EPTCS 379

Proceedings of the

Nineteenth conference on
Theoretical Aspects of Rationality and Knowledge

Oxford, United Kingdom, 28-30th June 2023

Edited by: Rineke Verbrugge

Published: 11th July 2023
DOI: 10.4204/EPTCS.379
ISSN: 2075-2180
Open Publishing Association

Preface

Rineke Verbrugge
Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence
University of Groningen
L.C.Verbrugge@rug.nl

The TARK conference (Theoretical Aspects of Rationality and Knowledge) is a conference that aims to bring together researchers from a wide variety of fields, including computer science, artificial intelligence, game theory, decision theory, philosophy, logic, linguistics, and cognitive science. Its goal is to further our understanding of interdisciplinary issues involving reasoning about rationality and knowledge.

Previous conferences have been held biennially around the world since 1986, on the initiative of Joe Halpern (Cornell University). Topics of interest include, but are not limited to, semantic models for knowledge, belief, awareness and uncertainty, bounded rationality and resource-bounded reasoning, commonsense epistemic reasoning, epistemic logic, epistemic game theory, knowledge and action, applications of reasoning about knowledge and other mental states, belief revision, computational social choice, algorithmic game theory, and foundations of multi-agent systems. Information about TARK, including conference proceedings, is available at the website http://www.tark.org/

These proceedings contain the papers that have been accepted for presentation at the Nineteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2023), held between June 28 and June 30, 2023, at the University of Oxford, United Kingdom. The conference website can be found at https://sites.google.com/view/tark-2023

The conference is enlivened by four invited talks, given by:

- Aviad Heifetz (Open University of Israel)
- Willemien Kets (Utrecht University)
- Jon Kleinberg (Cornell University)
- Anna Mahtani (London School of Economics)

The Programme Committee received 82 regular paper submissions. Of these, 40 were selected for this volume in a reviewing process during which every paper received three independent expert reviews. Decisions were often difficult and were based on lively discussions between PC members. Of the 40 accepted papers, 21 will be presented as an oral lecture and 19 as a poster presentation accompanied by a flash talk. This volume evidences the interdisciplinary nature of research on theoretical aspects of rationality and knowledge: Several papers are contributions to more than one of the fields listed above, and all of them have been written to be understood by readers across discipline boundaries.

The members of the Programme Committee for the conference were:

- Christian Bach, University of Liverpool

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. i–iii, doi:10.4204/EPTCS.379.0

- Adam Bjorndahl, Carnegie Mellon University
- Zoé Christoff, University of Groningen
- Hans van Ditmarsch, CNRS, IRIT, University of Toulouse
- Malvin Gattinger, University of Amsterdam
- Sujata Ghosh, ISI Chennai
- Nina Gierasimczuk, Technical University of Copenhagen
- Joe Halpern, Cornell University
- Willemien Kets, Utrecht University
- Louwe Kuijer, University of Liverpool
- Jérôme Lang, LAMSADE, Université Paris-Dauphine
- Silvia Milano, University of Exeter
- Larry Moss, University of Indiana
- Andrés Perea, University of Maastricht
- Gabriella Pigozzi, LAMSADE, Université Paris-Dauphine
- Olivier Roy, University of Bayreuth
- Burkhard Schipper, University of California at Davis
- Paolo Turrini, University of Warwick
- Rineke Verbrugge, University of Groningen (chair)

Many other people assisted with the reviewing process, including: Leyla Ade, Edoardo Baccini, Philippe Balbiani, Jacques Bara, Fausto Barbero, Gaia Belardinelli, Francesco Berto, Patrick Blackburn, Giacomo Bonanno, Richard Booth, Martin Caminada, Sourav Chakraborty, Michele Crescenzi, Ramit Das, Adam Dominiak, Soma Dutta, Peter van Emde Boas, Jie Fan, Peter Fritz, Asta Halkjær From, Satoshi Fukuda, Paolo Galeazzi, Rustam Galimullin, Avijeet Ghosh, Patrick Girard, Olga Gorelkina, Davide Grossi, Pierfrancesco Guarino, Shreyas Gupta, Jens Ulrik Hansen, Adrian Haret, Aviad Heifetz, Wesley Holliday, Prosenjit Howlader, Neil Hwang, Stephan Jagau, Dominik Klein, Barteld Kooi, Jan Lastovicka, Dazhu Li, Grzegorz Lisowski, Shuige Liu, Emiliano Lorini, Maaike Los, Munyque Mittelmann, Niels Mourmans, Eric Pacuit, Anantha Padmanabha, Timothy Parker, Mina Young Pedersen, Rafael Peñaloza, Elise Perrotin, Charlie Pilgrim, Robert Routledge, Ocan Sankur, Katsuhiko Sano, François Schwarzentruber, Ted Shear, Chenwei Shi, Sonja Smets, Tomasz Steifer, Katrine B. P. Thoft, Paolo Viappiani, Yanjing Wang, Yì Nicholas Wáng, Nic Wilson, Fabio Massimo Zennaro, Stanislav Zhydkov, Gabriel Ziegler, Aybüke Özgün.

I would like to thank the members of the Programme Committee and all other reviewers for the time, professional effort and the expertise that they invested in ensuring the high scientific standards of the conference and its proceedings and for providing a lot of useful suggestions for further improvements to the authors. It was an honor and pleasure for me to read your thoughtful reviews and share in the discussions about the papers. I also thank the authors for their excellent contributions. Moreover, I thank Rob Glabbeek of EPTCS for bringing this volume to publication and for his kind support to the authors and to me as editor.

I want to express my thanks to the organizers of the conference, Mike Wooldridge and Jenny Dollard, for their dedication in bringing TARK 2023 to life in the beautiful grounds of Worcester College at the University of Oxford. Special thanks go to the TARK General Chair Joe Halpern who started the conference series in 1986 and who supported us with his advice in all phases of the conference preparations.

Rineke Verbrugge
Programme chair, TARK 2023
Groningen, The Netherlands
June 20th, 2023

Table of Contents

Epistemic Conditions for Bayesian Equilibrium

Christian W. Bach
EPICENTER & University of Liverpool
cwbach@liverpool.ac.uk

Andrés Perea
EPICENTER & University of Maastricht
a.perea@maastrichtuniversity.nl

Bayesian equilibrium constitutes the prevailing solution concept for games with incomplete information. It is known that from an ex-ante perspective Harsanyi's seminal notion is related both to Nash equilibrium as well as to canonical correlated equilibrium. We provide an epistemic characterization of Bayesian equilibrium from an interim perspective by means of common belief in rationality and a common prior. Since these epistemic conditions also characterize correlated equilibrium in the special case of complete information, our result substantiates that Bayesian equilibrium forms the incomplete information analogue to correlated equilibrium – and not to Nash equilibrium – in terms of reasoning.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 1–1, doi:10.4204/EPTCS.379.1

Group Knowledge and Individual Introspection

Michele Crescenzi
Discipline of Economics
University of Helsinki
Helsinki, Finland
michele.crescenzi@helsinki.fi

The goal of the paper is to examine distributed knowledge in groups with differently introspective agents. Three categories of agents are considered: non-introspective, positively introspective, and fully introspective. When a non-introspective agent knows something, she may fail to know that she knows it. On the contrary, when a fully introspective agent knows something, she always knows that she knows it. A fully introspective agent is positively introspective and, when she does not know something, she also knows that she does not know it. We give two equivalent characterizations of distributed knowledge: one in terms of knowledge operators and the other in terms of possibility relations, i.e., binary relations. We show that two different cases emerge. In the first, distributed knowledge is fully determined by the group member who is sophisticated enough to replicate all the inferences that anyone else in the group can make. In the second case, no member is sophisticated enough to replicate what anyone else in the group can infer. As a result, distributed knowledge is determined by a two-person subgroup who can jointly replicate what others infer. The latter case depicts a wisdom-of-the-crowd effect, in which the group knows more than what any of its members could possibly know by having access to all the information available within the group. Finally, we show that distributed knowledge is not always represented by the intersection of the group members' possibility relations. Depending on how introspective agents are, distributed knowledge may be represented by strict refinements of the aforementioned intersection. A full version of the paper is available at https://arxiv.org/abs/2305.08729

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 2–2, doi:10.4204/EPTCS.379.2

Incomplete Preferences, Multi-Utility Representations, and the Axiom of Parity

Harvey Lederman
UT Austin
Department of Philosophy
harvey.lederman@austin.utexas.edu

This paper studies extensions of incomplete preferences over basic alternatives to preferences over lotteries on those alternatives. I begin by introducing the normatively and descriptively plausible *axiom of parity*, an analog of typical principles of dominance: for an outcome o and a lottery L, if, for every outcome o' in the support of L, both $o \not\succeq o'$ and $o' \not\succeq o$, then both $L \not\succ o$ and $o \not\succ L$. The main result of the paper shows that, in a natural setting, where incomplete preferences are sensitive to an underlying space of totally ordered 'dimensions', the axiom of parity is incompatible with all natural ways of extending preferences from basic outcomes to lotteries over them. In particular, the axiom of parity is inconsistent with the natural principle that, if the outcomes in the support of a lottery only vary along a single dimension, the decision-maker should be indifferent between the lottery and its expected value.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 3–3, doi:10.4204/EPTCS.379.3

Strengthening Consistency Results in Modal Logic

Samuel Allen Alexander
US Securities and Exchange Commission
New York, USA
samuelallenalexander@gmail.com

Arthur Paul Pedersen
City University of New York
New York, USA
apedersen@cs.ccny.cuny.edu

A fundamental question asked in modal logic is whether a given theory is consistent. But consistent with what? A typical way to address this question identifies a choice of background knowledge axioms (say, S4, D, etc.) and then shows the assumptions codified by the theory in question to be consistent with those background axioms. But determining the specific choice and division of background axioms is, at least sometimes, little more than tradition. This paper introduces *generic theories* for propositional modal logic to address consistency results in a more robust way. As building blocks for background knowledge, generic theories provide a standard for categorical determinations of consistency. We argue that the results and methods of this paper help to elucidate problems in epistemology and enjoy sufficient scope and power to have purchase on problems bearing on modalities in judgement, inference, and decision making.

1 Introduction

Many treatments of epistemological paradoxes in modal logic proceed along the following lines. Begin with some enumeration of assumptions that are individually plausible but when taken together fail to be jointly consistent (or at any rate fail to stand to reason in some way). Thereupon proceed to propose a resolution to the emerging paradox that identifies one or more assumptions that may be comfortably discarded or weakened and that in the presence of the remaining assumptions circumvents the troubling inconsistency defining the paradox [11] (cf. Chow [8] and de Vos et al. [16]). Typical among such assumptions are logical standards expressed in the form of inference rules and axioms pertaining to knowledge and belief, such as axiom scheme **K** — that is to say, the distributive axiom scheme of the form $\mathrm{K}(\varphi \rightarrow \psi) \rightarrow (\mathrm{K}\varphi \rightarrow \mathrm{K}\psi)$.

The choice of precisely *which* assumptions to temper can, at times, have an element of arbitrariness to it, especially when the choice is made from among several independent alternatives underpinning distinct resolutions in the absence of clear criteria or compelling grounds for distinguishing among them. In the present paper, we introduce a criterion for addressing this predicament based on the *genericity* of what a resolution assumes.

As a standard for knowledge, a theory is generic when its factivity cannot be overturned however the questions it leaves open are answered and what is known accordingly grows. Generic theories enjoy various desirable properties which are common in formal epistemology — arbitrary unions of generic theories, for example, are generic. We present both positive and negative results turning on genericness, which cast light on the structure of popular logics for belief and knowledge.

The concept of generic theories, as introduced in [4] and [5] for quantified modal logic, emerged in response to Carlson's proof [7] of a conjecture due to Reinhardt [13]. Carlson's proof, despite its significance, was limited by its dependency on a somewhat arbitrary choice of background knowledge axioms. Carlson proof, subject to but small changes, is likewise valid for various other sets of background axioms. The present paper examines generic theories for propositional modal logic. In our concluding

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 4–15, doi:10.4204/EPTCS.379.4

remarks we discuss the developments of this paper in connection with work done to generalize Carlson's consistency result.

The paper is organized as follows. In Section 2 we state preliminaries. In Section 3 we state a propositional version of the Knower Paradox: a certain theory, consisting of standard background knowledge axioms plus an axiom intended to be read as "This sentence is known to be false," is inconsistent. We discuss a possible resolution to the paradox: weaken the background knowledge axioms in order to render the theory consistent. In Section 4 we introduce generic and closed generic theories. In Section 5 we use generic and closed generic theories to state very generalized versions of the consistency result from Section 3. In Section 6 we state some negative results about genericness and closed genericness. Proofs of these negative results naturally lead to the construction of exotic models which satisfy certain standard knowledge axioms while failing certain other standard knowledge axioms. In Section 7 we conclude the paper with a high-level discussion. In Appendix A we give proofs of some of the claims made in the above sections.

2 Preliminaries

Throughout, we fix a nonempty set of symbols called *propositional atoms* and a symbol K which is not a propositional atom. The following logic is a propositional version of Carlson's so-called *base logic* [7] (cf. [2] and [1]).

Definition 1. The set of *formulas* is defined recursively as follows:

(i) Every propositional atom is a formula;

(ii) Whenever φ and ψ are formulas, so are $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, and $(\varphi \rightarrow \psi)$; and

(iii) Whenever φ is formula, so too is $\mathrm{K}(\varphi)$.

A formula is said to be *basic* if it is either a propositional atom or a formula of the form $\mathrm{K}(\varphi)$ for some formula φ. A set of formulas is called a *theory*. ◂

We adopt standard conventions for omitting parentheses. Parentheses omitted from conditional formulas are assumed to be right-nested; thus, for example, we write $\phi \rightarrow \psi \rightarrow \rho$ for $\phi \rightarrow (\psi \rightarrow \rho)$, and similarly for longer chains of implications.

Definition 2. A *model* is a function mapping each basic formula to a truth value in $\{\mathrm{True}, \mathrm{False}\}$. ◂

Thus, in contrast with classical treatments of semantics for modalities, a model assigns truth values not only to propositional atoms but also to formulas prefixed with K.

We may define a binary relation $\models$ from models to basic formulas in the usual way — that is, by stipulating that $\mathscr{M} \models \varphi$ just in case $\mathscr{M}$ assigns to φ the value True. The next definition extends this relation to all formulas. We adopt the standard convention to write $\mathscr{M} \not\models \varphi$ if it is not the case that $\mathscr{M} \models \varphi$.

Definition 3. Let $\mathscr{M}$ be a model. Define formula φ to be *true* in $\mathscr{M}$, $\mathscr{M} \models \varphi$, by recursion on φ:

(i) If φ is a basic formula, then $\mathscr{M} \models \varphi$ if and only if $\mathscr{M}$ assigns to φ the value True;

(ii) $\mathscr{M} \models \neg\varphi$ if and only if $\mathscr{M} \not\models \varphi$;

(iii) $\mathscr{M} \models \varphi \wedge \psi$ if and only if both $\mathscr{M} \models \varphi$ and $\mathscr{M} \models \psi$;

(iv) $\mathscr{M} \models \varphi \vee \psi$ if and only if either $\mathscr{M} \models \varphi$ or $\mathscr{M} \models \psi$; and

(v) $\mathscr{M} \models \varphi \rightarrow \psi$ if and only if either $\mathscr{M} \not\models \varphi$ or $\mathscr{M} \models \psi$.

Given a theory T, we write $\mathscr{M} \models T$ just in case $\mathscr{M} \models \varphi$ for every $\varphi \in T$. ◀

Entailment and validity are given standard treatment.

Definition 4. A theory T is said to *entail* a formula φ, written $T \models \varphi$, if for all models $\mathscr{M}$, $\mathscr{M} \models T$ implies $\mathscr{M} \models \varphi$. A formula φ is said to be *valid*, written $\models \varphi$, if $\emptyset \models \varphi$. ◀

Since modal formulas of the form $\mathrm{K}\varphi$ are treated like propositional atoms, it follows that if p is a propositional atom, then $\mathrm{K}p \vee \neg\mathrm{K}p$ is valid but $\mathrm{K}(p \vee \neg p)$ is not. Routine argument establishes compactness. A useful result is the following corollary of compactness.

Lemma 5. *Let T be a theory and φ be a formula. Then $T \models \varphi$ if and only if there is a finite sequence of formulas $\varphi_1, \ldots, \varphi_n \in T$ for which $\models \varphi_1 \to \cdots \to \varphi_n \to \varphi$.*

Lemma 5 provides a basis for adopting the following proof-theoretic terminology in what follows.

Definition 6. A theory T is said to be *consistent* if there is a model $\mathscr{M}$ for which $\mathscr{M} \models T$. ◀

The following definition captures the familiar notion of closedness under the K operator.

Definition 7. A theory T is *closed* if $\left\{ \mathrm{K}\varphi \;:\; \varphi \in T \right\} \subseteq T$. ◀

Thus a theory T is closed just in case for every formula φ, if $\varphi \in T$, then $\mathrm{K}\varphi \in T$.

Definition 8. We adopt the following conventions for naming standard schemas:

V is the theory consisting of all formulas of the form $\mathrm{K}\varphi$ such that φ is valid (Definition 4).

K is the theory consisting of all formulas of the form $\mathrm{K}(\varphi \to \psi) \to (\mathrm{K}\varphi \to \mathrm{K}\psi)$.

T is the theory consisting of all formulas of the form $\mathrm{K}\varphi \to \varphi$.

KK (sometimes also called **4**) is the theory consisting of all formulas of the form $\mathrm{K}\varphi \to \mathrm{KK}\varphi$. ◀

We conclude this section with an observation about necessitation (proved in Appendix A).

Lemma 9. (Simulated Necessitation) *Let T be a closed theory. If T includes both* **V** *and* **K**, *then for every formula φ : if $T \models \varphi$, then $T \models \mathrm{K}\varphi$.*

3 A Formalization of the Knower Paradox

We will use a propositional version of the well-known Knower Paradox [12] to illustrate the ideas of this paper. The paradox is usually formalized in first-order modal logic, where appeal to Gödel's Diagonal Lemma admits construction of the problematic sentence without having to assume it as an axiom. In our propositional version, we instead assume the problematic sentence axiomatically, allowing us to focus on the epistemological contents of the paradox without arithmetical distractions.

Theorem 10 (The Knower Paradox)**.** *Let p be some propositional atom. Let T_{KP} be the smallest closed theory which contains*:

(i) **V**, **K**, *and* **T**

(ii) $p \leftrightarrow \mathrm{K}\neg p$ *"This sentence is known to be false"*

Then the theory T_{KP} is inconsistent. ■

Proof. From schema **T** and axiom (ii), it follows that $T_{KP} \models \neg p$ and therefore $T_{KP} \models \mathrm{K}\neg p$ by Lemma 9, whence $T_{KP} \models p$ by axiom (ii). Hence, T_{KP} is inconsistent. □

The next theorem provides one way the theory in Theorem 10 may be weakened in order to restore consistency (and so constitutes a candidate for resolving the paradox, in the sense of Haack [11] or Chow [8]).

Theorem 11. *Let p be some propositional atom. Inductively, let $(T^-_{KP})_0$ be the smallest closed theory which contains*:

(i) **V** *and* **K**

(ii) $p \leftrightarrow \mathrm{K}\neg p$ *"This sentence is known to be false"*

In addition, let T^-_{KP} be the theory which contains:

(a) $(T^-_{KP})_0$.

(b) **T**.

Then theory T^-_{KP} is consistent. ■

Observe that the Knower Paradox (Theorem 10), so formalized, rests on the assumption that the knower know its own truthfulness. The key difference between T_{KP} and T^-_{KP} is that, while the schema $\mathrm{K}\varphi \rightarrow \varphi$ is included in both theories, only T_{KP} includes the schema $\mathrm{K}(\mathrm{K}\varphi \rightarrow \varphi)$. Some treatments[1] of the Knower Paradox do not explicitly include $\mathrm{K}(\mathrm{K}\varphi \rightarrow \varphi)$ as an assumption at all, instead including $\mathrm{K}\varphi \rightarrow \varphi$ and using a logic where the *rule of necessitation* holds—the rule permitting one to conclude $T \models \mathrm{K}\varphi$ from $T \models \varphi$. In such logics, if T contains the schema $\mathrm{K}\varphi \rightarrow \varphi$, then trivially $T \models \mathrm{K}\varphi \rightarrow \varphi$, so by necessitation, $T \models \mathrm{K}(\mathrm{K}\varphi \rightarrow \varphi)$. Thus, $\mathrm{K}(\mathrm{K}\varphi \rightarrow \varphi)$ sneaks in implicitly, in such logics.

The logic (Definition 1) studied in this paper does not presume the rule of necessitation. The rule of necessitation can be simulated in our logic by using Lemma 9, but only if the Lemma's conditions are met—which, in the case of T^-_{KP}, they are not, as T^-_{KP} is not closed. Thus, it becomes possible to weaken knowledge-of-factivity without weakening factivity itself. Theorem 11 shows that doing so is one possible resolution, in the sense of Haack [11] or Chow [8], to the paradox.[2] See [1, 15] for discussion about the weakening of knowledge-of-factivity. Note that this requires departing from Kripke semantics, as the rule of necessitation always holds in Kripke semantics.

Rather than prove Theorem 11 directly, we will (in Section 5) prove a pair of more general theorems, and Theorem 11 is a special case of either one of them. In order to state the more general theorems, we need to first introduce certain notions of genericity.

4 Generic and Closed Generic Theories

The following definition is a variant of Carlson's concept of a *knowing entity* [7].

Definition 12. Let T be a theory, and let S be a set of propositional atoms. Let $\mathscr{M}_{T,S}$ be the model defined by stipulating:

(i) For any propositional atom p: $\mathscr{M}_{T,S} \models p$ if and only if $p \in S$; and

(ii) For any formula of the form $\mathrm{K}\varphi$: $\mathscr{M}_{T,S} \models \mathrm{K}\varphi$ if and only if $T \models \varphi$. ◂

[1] See [9, 10] for an exception.

[2] The same technique has been used to resolve (in Haack's or Chow's sense) a version of the surprise exam paradox [1]; to resolve a version of Fitch's paradox [2]; and to construct a machine that knows its own code [3]. Aldini et al suggest [1] it might be possible to *simultaneously* resolve multiple paradoxes at once by dropping $\mathrm{K}(\mathrm{K}\varphi \rightarrow \varphi)$, i.e., the *union* of *multiple* paradoxically inconsistent theories might be consistent when so weakened.

The model $\mathscr{M}_{T,S}$ may be loosely interpreted to be that of an agent who knows exactly the consequences of theory T in a world in which all propositions from S are true. We will see that these models are useful for establishing consistency results.

The following definition strengthens the notion of consistency.

Definition 13. A theory T is said to be *generic* (resp. *closed generic*) if for each set S of propositional atoms and each theory (resp. *closed theory*) T': if $T' \supseteq T$, then $\mathscr{M}_{T',S} \models T$. ◂

A theory T is generic when T is known regardless of contingent facts S and however theoretical knowledge might grow in conjunction with them. Generic theories are theories that cannot be made false by the addition of more information.

We catalogue basic properties of genericity.

Proposition 14. *Genericity enjoys the following properties*:

(1) *Unions of generic theories are generic*;

(2) *Unions of closed generic theories are closed generic*;

(3) *Every generic theory is closed generic*;

(4) $\mathbf{V}$ *is generic*; and

(5) $\mathbf{K}$ *is generic.* ■

Proof. Properties (1)–(3) are readily verified.

(4) Let S be a set of propositional atoms and let $T' \supseteq \mathbf{V}$. Let $\varphi \in \mathbf{V}$, we must show $\mathscr{M}_{T',S} \models \varphi$. By definition of $\mathbf{V}$, φ is $\mathrm{K}\psi$ for some valid ψ. Since ψ is valid, $T' \models \psi$. Thus $\mathscr{M}_{T',S} \models \mathrm{K}\psi$, as desired.

(5) Let S be a set of propositional atoms and let $T' \supseteq \mathbf{K}$. Let $\varphi \in \mathbf{K}$, we must show $\mathscr{M}_{T',S} \models \varphi$. By definition of $\mathbf{K}$, φ is $\mathrm{K}(\psi \rightarrow \rho) \rightarrow (\mathrm{K}\psi \rightarrow \mathrm{K}\rho)$ for some ψ and ρ. Assume $\mathscr{M}_{T',S} \models \mathrm{K}(\psi \rightarrow \rho)$ and $\mathscr{M}_{T',S} \models \mathrm{K}\psi$. This means $T' \models \psi \rightarrow \rho$ and $T' \models \psi$. By modus ponens, $T' \models \rho$. So $\mathscr{M}_{T',S} \models \mathrm{K}\rho$, as desired. □

Lemma 15. *The theory* $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ *is closed generic.* ■

Proof. Let $T = \mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$. Let S be a set of propositional atoms and let $T' \supseteq T$ be closed. Let $\varphi \in T$, we must show $\mathscr{M}_{T',S} \models \varphi$. Consider two cases:

Case 1 $\varphi \in \mathbf{V} \cup \mathbf{K}$. Then $\mathscr{M}_{T',S} \models \varphi$ because $\mathbf{V} \cup \mathbf{K}$ is generic by Proposition 14, parts (1), (4), and (5).

Case 2 $\varphi \in \mathbf{KK}$. Then φ is $\mathrm{K}\psi \rightarrow \mathrm{KK}\psi$ for some ψ. Assume $\mathscr{M}_{T',S} \models \mathrm{K}\psi$. This means $T' \models \psi$. Since T' contains $\mathbf{V}$ and $\mathbf{K}$ and is closed, we may simulate necessitation: Lemma 9 implies $T' \models \mathrm{K}\psi$. Thus $\mathscr{M}_{T',S} \models \mathrm{KK}\psi$, as desired. □

Lemma 16. *Let T_0 be a theory, and let T be the smallest closed theory including theory T_0. Suppose theory T_0 is (closed) generic. Then T is (closed) generic.* ■

Proof. Let S be a set of propositional atoms and let T' be a theory (resp. closed theory) such that $T' \supseteq T$. Let $\varphi \in T$, we must show $\mathscr{M}_{T',S} \models \varphi$. Consider two cases:

Case 1 $\varphi \in T_0$. Then $\mathscr{M}_{T',S} \models \varphi$ because T_0 is generic (resp. closed generic).

Case 2 $\varphi \notin T_0$. The only other way for φ to be in T (besides being in T_0) is by way of the closure of T. So φ is $K\psi$ for some $\psi \in T$. Since $T' \supseteq T$ and $\psi \in T$, we have $T' \models \psi$, which means $\mathscr{M}_{T',S} \models \mathrm{K}\psi$, as desired. □

Lemma 17. *Suppose T_0 is a generic (resp. closed generic) theory. Let $T = \{\varphi : T_0 \models \varphi\}$. Then T is generic (resp. closed generic).* ■

Proof. Let S be an arbitrary set of propositional atoms, and let T' be a theory (resp. closed theory) such that $T' \supseteq T$. We establish that $\mathscr{M}_{T',S} \models T$.

For each formula φ such that $T \models \varphi$, let $N(\varphi)$ be the smallest positive integer n for which there is a sequence $\varphi_1, \dots, \varphi_n$, with $\varphi_n = \varphi$, such that for each $i = 1, \dots, n$, either $\varphi_i \in T_0$ or there exist $j, k < i$ such that φ_k is $\varphi_j \rightarrow \varphi_i$. Such an $N(\varphi)$ exists by the deduction theorem.

We prove by induction on $N(\varphi)$ that for every φ such that $T_0 \models \varphi$, $\mathscr{M}_{T',S} \models \varphi$.

Basis Step $N(\varphi) = 1$ can clearly only hold if $\varphi \in T_0$. In that case, $\mathscr{M}_{T',S} \models \varphi$ because T_0 is generic (resp. closed generic).

Inductive Step $N(\varphi) > 1$. If $\varphi \in T_0$, we are done as in the Base Case, but assume not. Let $\varphi_1, \dots, \varphi_n$ be a sequence of length $n = N(\varphi)$ with the above properties.

For each $i < n$, the subsequence $\varphi_1, \dots, \varphi_i$ is a shorter sequence (with the above properties) for φ_i, showing $N(\varphi_i) < N(\varphi)$. Thus by induction, $(*)$ for each $i < n$, $\mathscr{M}_{T',S} \models \varphi_i$. Since $\varphi \notin T_0$, there must be $j, k < n$ such that φ_k is $\varphi_j \rightarrow \varphi_n$. By $*$, $\mathscr{M}_{T',S} \models \varphi_j$ and $\mathscr{M}_{T',S} \models \varphi_k$. So $\mathscr{M}_{T',S} \models \varphi_j \rightarrow \varphi_n$. By modus ponens, $\mathscr{M}_{T',S} \models \varphi_n$, as desired.

□

We conclude this section with a result throwing light on the relationship between generic theories and normal modal logics. The proof is immediate by combining Lemmas 14, 16, and 17.

Theorem 18. *Suppose T_0 is a (closed) generic theory. Let T be the* normal Kripke closure *of T_0, i.e., the smallest closed theory containing T_0,* **V**, **K**, *and with the property that T contains ϕ whenever $T \models \phi$. Then T is (closed) generic.* ■

5 Two Generalized Consistency Statements

In what follows, we state two theorems, each generalizing Theorem 11. One might be curious whether adding **KK** to the statement of Theorem 11 would make the paradox reappear. Certainly the paradox as formulated in Theorem 10 does not use **KK** in its proof. But what if there is some other form of the Knower's Paradox that makes use of **KK**, and what if in fact we only managed to achieve consistency because we neglected to include **KK** among the background axioms? We could state a separate version of Theorem 11 which includes **KK** and then prove that separate version, with a proof that is extremely similar to a proof of Theorem 11 itself, but then maybe there's still some further background axiom that we are still neglecting, and we would then have to state and prove yet a third version of the theorem. This process might go on forever, we might never exhaustively think of all the different background axioms that critics might insist upon.

Theorem 19. *Let p be a propositional atom, and let H be a generic theory. Let $(T_{KP})_0$ be the smallest closed theory containing:*

(i) H

(ii) $p \leftrightarrow \mathrm{K}\neg p$ *"This sentence is known to be false"*

In addition, let T_{KP} be the theory containing:

(a) $(T_{KP})_0$; and

(b) $\mathbf{T}$.

For any set S of propositional atoms, if $p \notin S$ then $\mathscr{M}_{(T_{KP})_0,S} \models T_{KP}$. In particular, T_{KP} is consistent. ■

We prove Theorem 19 in Appendix A. Observe that since theory $\mathbf{V} \cup \mathbf{K}$ is generic by Proposition 14, Theorem 11 is a special case of Theorem 19.

Now modify Theorem 11 by replacing $\mathbf{V} \cup \mathbf{K}$ with $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$. We could not do that using Theorem 19 unless we first established that $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ was generic (in fact, in the next section, we will show that $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ is *not* generic). We do know that $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ is closed generic (Lemma 15), so we would be done if we had a version of Theorem 19 involving closed generic theories.

Theorem 20. *Same as Theorem 19 but with "generic" replaced by "closed generic."* ■

A proof similar to the one for Theorem 19 establishes Theorem 20.

6 Negative Results about Genericness

We have established theory $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ to be closed generic. Are these results preserved if one or more of the arguments to the union is dropped? For example, is theory $\mathbf{V} \cup \mathbf{KK}$ closed generic? Or the theory $\mathbf{K} \cup \mathbf{KK}$? What about the theory $\mathbf{KK}$ alone? Similarly, can we strengthen closed genericity of $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ to full genericity? We show each of these questions has one and the same answer: No.

Theorem 21. *The theory $\mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$ fails to be generic.* ■

Proof. Let $T = \mathbf{V} \cup \mathbf{K} \cup \mathbf{KK}$. Let p be some propositional atom and let $T' = T \cup \{p\}$. We show that $\mathscr{M}_{T',\emptyset} \not\models \mathrm{K}p \rightarrow \mathrm{KK}p$, whereby $\mathscr{M}_{T',\emptyset} \not\models \mathbf{KK}$ and so $\mathscr{M}_{T',\emptyset} \not\models T$, showing T is not generic. Clearly $T' \models p$, so $\mathscr{M}_{T',\emptyset} \models \mathrm{K}p$. What remains to show is that $\mathscr{M}_{T',\emptyset} \not\models \mathrm{KK}p$ — that is, $T' \not\models \mathrm{K}p$.

To this end, inductively define models $\mathscr{N}_1$ and $\mathscr{N}_2$ simultaneously by stipulating $\mathscr{N}_1 \models q$ and $\mathscr{N}_2 \not\models q$ for each propositional atom q and requiring that $\mathscr{N}_1$ and $\mathscr{N}_2$ interpret formulas $\mathrm{K}\varphi$ in the following way:

$\mathscr{N}_2 \models \mathrm{K}\varphi$ if and only if $\mathscr{N}_2 \models \varphi$; and

$\mathscr{N}_1 \models \mathrm{K}\varphi$ if and only if $\mathscr{N}_2 \models \varphi$.

Since $\mathscr{N}_2 \not\models p$, $\mathscr{N}_1 \not\models \mathrm{K}p$. Thus, to show that $T' \not\models \mathrm{K}p$, and so conclude the proof, it suffices to show $\mathscr{N}_1 \models T'$.

Let $\varphi \in T'$. Consider four cases:

Case 1 $\varphi \in \mathbf{V}$. Then φ is $\mathrm{K}\varphi_0$ for some valid φ_0. Since φ_0 is valid, $\mathscr{N}_2 \models \varphi_0$, so $\mathscr{N}_1 \models \mathrm{K}\varphi_0$.

Case 2 $\varphi \in \mathbf{K}$. Then φ has the form $\mathrm{K}(\psi \rightarrow \rho) \rightarrow (\mathrm{K}\psi \rightarrow \mathrm{K}\rho)$. Assume $\mathscr{N}_1 \models \mathrm{K}(\psi \rightarrow \rho)$ and $\mathscr{N}_1 \models \mathrm{K}\psi$. Then $\mathscr{N}_2 \models \psi \rightarrow \rho$ and $\mathscr{N}_2 \models \psi$. By modus ponens, $\mathscr{N}_2 \models \rho$. Thus $\mathscr{N}_1 \models \mathrm{K}\rho$, as desired.

Case 3 $\varphi \in \mathbf{KK}$. Then φ has the form $\mathrm{K}\psi \rightarrow \mathrm{KK}\psi$. Assume $\mathscr{N}_1 \models \mathrm{K}\psi$. Then $\mathscr{N}_2 \models \psi$, so $\mathscr{N}_2 \models \mathrm{K}\psi$, whence $\mathscr{N}_1 \models \mathrm{KK}\psi$, as desired.

Case 4 φ is p. Then $\mathscr{N}_1 \models \varphi$ by construction. □

Proposition 14 can be used to establish an immediate corollary of Theorem 21.

Corollary 22. *The theories* $\mathbf{V} \cup \mathbf{KK}$ *and* $\mathbf{K} \cup \mathbf{KK}$ *fail to be generic.* ■

The following corollary follows from Theorem 21 and Lemma 15.

Corollary 23. *Not every closed generic theory is generic.* ■

Theorem 24. *If* $\mathbf{V} \cup \mathbf{KK}$ *is closed generic, then there is at most one propositional atom.*

The preceding theorem, like the one stated next, is proven in Appendix A.

Theorem 25. *The theory* $\mathbf{K} \cup \mathbf{KK}$ *is not closed generic.* ■

The following corollary follows by Proposition 14.

Corollary 26. *The theory* $\mathbf{KK}$ *is not closed generic.* ■

The proof of Theorem 21 illustrates a technique common to all proofs appearing in Appendix A for the results stated in this section — each argument proceeds by constructing pathological models. Investigating negative results about genericness and closed genericness using this technique locates sharp edges at the boundaries of modal logic: we are led to consider models where common assumptions no longer hold, such as models where **K** fails or where **V** fails.

We have applied the theory of genericity to the Knower Paradox. In the proofs of the following theorems, we will reverse the direction of application, applying the Knower Paradox to the theory of genericity, rather than vice versa.

Theorem 27. *The theory* $\mathbf{T}$ *is not closed generic. In fact, no superset of* $\mathbf{T}$ *is closed generic.* ■

Proof. Assume $\mathbf{T}^+ \supseteq \mathbf{T}$ is closed generic. By Lemma 14, $H = \mathbf{V} \cup \mathbf{K} \cup \mathbf{T}^+$ is closed generic. Let T_{KP} be as in Theorem 20. By Theorem 20, T_{KP} is consistent. But it is easy to see that T_{KP} is at least as strong as the theory of the same name from Theorem 10 (the Knower Paradox), which is *inconsistent*. Absurd. □

In particular, $S4$ is not closed generic (and thus not generic), and the same goes for $S5$. The following theorem implies that the same also goes for $KD45$.

Theorem 28. *Let* **5** *be the schema consisting of all formulas of the form* $\neg \mathrm{K}\phi \rightarrow \mathrm{K}\neg\mathrm{K}\phi$. *No superset of* **5** *is closed generic.* ■

Proof. Similar to Theorem 27 by reformulating the Knower's Paradox using **5** instead of **T**. □

7 Discussion

There are different forms of genericity, two of which we have examined above: generic theories and closed generic theories. These forms are particularly nice because of closure under union (Proposition 14 parts 1–2) and because they are simple enough that we can prove some results about them.

In future work, we intend to use closed generic theories to generalize Carlson's consistency result [7] (this is almost already done in [5], but not quite, because the latter paper relies on an axiom called *assigned validity* to avoid some tricky nuances, whereas Carlson does not).

Acknowledgements

The authors thank three anonymous reviewers as well as Rineke Verbrugge for valuable comments and suggestions to help improve this manuscript. The authors also extend their gratitude to Alessandro Aldini, Michael Grossberg, Ali Kahn, Rohit Parikh, and Max Stinchcombe, for their generous feedback on prior drafts of this manuscript.

A Proofs

Lemma 9. (Simulated Necessitation) *Let T be a closed theory. If T includes both* **V** *and* **K**, *then for every formula φ : if $T \models \varphi$, then $T \models \mathrm{K}\varphi$.*

Proof of Lemma 9. By Lemma 5, there are $\varphi_1, \ldots, \varphi_n \in T$ such that $\varphi_1 \rightarrow \cdots \rightarrow \varphi_n \rightarrow \varphi$ is valid. By **V**,

$$T \models \mathrm{K}(\varphi_1 \rightarrow \cdots \rightarrow \varphi_n \rightarrow \varphi).$$

By repeated applications of **K**,

$$T \models \mathrm{K}(\varphi_1 \rightarrow \cdots \rightarrow \varphi_n \rightarrow \varphi) \rightarrow \mathrm{K}\varphi_1 \rightarrow \cdots \rightarrow \mathrm{K}\varphi_n \rightarrow \mathrm{K}\varphi.$$

Since T contains each φ_i, the closure of T ensures T contains each $\mathrm{K}\varphi_i$. Thus $T \models \mathrm{K}\varphi$. □

Theorem 19. *Let p be a propositional atom, and let H be a generic theory. Let $(T_{KP})_0$ be the smallest closed theory containing:*

(i) H

(ii) $p \leftrightarrow \mathrm{K}\neg p$ *"This sentence is known to be false"*

In addition, let T_{KP} be the theory containing:

(a) $(T_{KP})_0$; and

(b) **T**.

For any set S of propositional atoms, if $p \notin S$ then $\mathscr{M}_{(T_{KP})_0,S} \models T_{KP}$. In particular, T_{KP} is consistent. ■

Proof of Theorem 19. Let $\varphi \in T_{KP}$, we must show $\mathscr{M}_{(T_{KP})_0,S} \models \varphi$. Consider four cases:

Case 1 $\varphi \in H$. Then $\mathscr{M}_{(T_{KP})_0,S} \models \varphi$ because $(T_{KP})_0 \supseteq H$ and H is generic.

Case 2 φ is $p \leftrightarrow \mathrm{K}\neg p$. Since $p \notin S$, $\mathscr{M}_{(T_{KP})_0,S} \not\models p$, thus it suffices to show $\mathscr{M}_{(T_{KP})_0,S} \not\models \mathrm{K}(\neg p)$. Let S' be a set of propositional atoms with $p \in S'$, and let T_∞ be the set of all formulas.

We claim $\mathscr{M}_{T_\infty,S'} \models (T_{KP})_0$. To see this, let $\psi \in (T_{KP})_0$, we must show $\mathscr{M}_{T_\infty,S'} \models \psi$. Three subcases are to be considered:

Subcase 1 $\psi \in H$. Then $\mathscr{M}_{T_\infty,S'} \models \psi$ because H is generic and $T_\infty \supseteq H$.

Subcase 2 ψ is $p \leftrightarrow \mathrm{K}\neg p$. Since $p \in S'$, $\mathscr{M}_{T_\infty,S'} \models p$. And since T_∞ contains all formulas, $T_\infty \models \neg p$, thus $\mathscr{M}_{T_\infty,S'} \models \mathrm{K}\neg p$. So $\mathscr{M}_{T_\infty,S'} \models \psi$.

Subcase 3 ψ is $\mathrm{K}\rho$ for some ρ such that $\rho \in (T_{KP})_0$. Since T_∞ contains all formulas, $T_\infty \models \rho$, so $\mathscr{M}_{T_\infty,S'} \models \mathrm{K}\rho$.

This shows $\mathscr{M}_{T_\infty,S'} \models (T_{KP})_0$. Now since $\mathscr{M}_{T_\infty,S'} \models (T_{KP})_0$ and $\mathscr{M}_{T_\infty,S'} \models p$, this shows $(T_{KP})_0 \not\models \neg p$. Thus $\mathscr{M}_{(T_{KP})_0,S} \not\models \mathrm{K}(\neg p)$, as desired.

Case 3 φ is $\mathrm{K}\psi$ for some ψ such that $\psi \in (T_{KP})_0$. Since $(T_{KP})_0 \models \psi$, by definition $\mathscr{M}_{(T_{KP})_0,S} \models \mathrm{K}\psi$.

Case 4 $\varphi \in T_{KP} \backslash (T_{KP})_0$. Then φ is an instance of $\mathbf{T}$, i.e., φ is $\mathrm{K}\psi \rightarrow \psi$ for some ψ. Assume $\mathscr{M}_{(T_{KP})_0,S} \models \mathrm{K}\psi$. Then $(T_{KP})_0 \models \psi$. By Cases 1–3, $\mathscr{M}_{(T_{KP})_0,S} \models (T_{KP})_0$. Thus $\mathscr{M}_{(T_{KP})_0,S} \models \psi$.

□

Definition 29. Given a formula φ, define $\mathrm{K}^n\varphi$ by recursion on $n \in \mathbb{N}$ by $\mathrm{K}^0\varphi = \varphi$ and $\mathrm{K}^{n+1}\varphi = \mathrm{K}\mathrm{K}^n\varphi$. ◂

Theorem 24. *If* $\mathbf{V} \cup \mathbf{KK}$ *is closed generic, then there is at most one propositional atom.*

Proof of Theorem 24. Let $T = \mathbf{V} \cup \mathbf{KK}$. Assume there exist distinct propositional atoms p and q. Let T' be the theory which contains:

- $K^n\varphi$ for all $n \in \mathbb{N}$ and all $\varphi \in \mathbf{V}$.
- $K^n\varphi$ for all $n \in \mathbb{N}$ and all $\varphi \in \mathbf{KK}$.
- $K^n(p \rightarrow q)$ for all $n \in \mathbb{N}$.
- $K^n p$ for all $n \in \mathbb{N}$.

Clearly T' is closed and $T' \supseteq T$. We will show $\mathscr{M}_{T',\emptyset} \not\models Kq \rightarrow KKq$, so $\mathscr{M}_{T',\emptyset} \not\models T$, so T is not closed generic. Since T' contains p and $p \rightarrow q$, by modus ponens $T' \models q$, so $\mathscr{M}_{T',\emptyset} \models Kq$. It remains only to show $\mathscr{M}_{T',\emptyset} \not\models KKq$, i.e., that $T' \not\models Kq$.

Define models $\mathscr{N}_1$ and $\mathscr{N}_2$ inductively so that:

- For every propositional atom a, $\mathscr{N}_1 \models a$.
- For every propositional atom a, $\mathscr{N}_2 \not\models a$.
- For every formula φ, $\mathscr{N}_2 \models K\varphi$ iff $\mathscr{N}_2 \models \varphi$.
- For every formula φ, $\mathscr{N}_1 \models K\varphi$ iff $\mathscr{N}_2 \models \varphi$ or φ is $K^n p$ for some $n \in \mathbb{N}$.

Since q is distinct from p and $\mathscr{N}_2 \not\models q$, we have $\mathscr{N}_1 \not\models Kq$. So to show $T' \not\models Kq$ (and thus finish the proof), it suffices to show $\mathscr{N}_1 \models T'$. Let $\varphi \in T'$.

Case 1: φ is $K^n\psi$ for some $n \in \mathbb{N}$ and some $\psi \in \mathbf{V}$. Then φ is $K^{n+1}\psi_0$ for some valid ψ_0. Since ψ_0 is valid, $\mathscr{N}_2 \models \psi_0$, and it follows that $\mathscr{N}_1 \models K^{n+1}\psi_0$.

Case 2: φ is $K^n\psi$ for some $n \in \mathbb{N}$ and some $\psi \in \mathbf{KK}$. Then φ is $K^n(K\rho \rightarrow KK\rho)$ for some ρ. To show $\mathscr{N}_1 \models \varphi$, it suffices to show $\mathscr{N}_2 \models K\rho \rightarrow KK\rho$. Assume $\mathscr{N}_2 \models K\rho$, then by definition $\mathscr{N}_2 \models KK\rho$, as desired.

Case 3: φ is $K^n(p \rightarrow q)$ for some $n \in \mathbb{N}$. Since $\mathscr{N}_2 \not\models p$, we have $\mathscr{N}_2 \models p \rightarrow q$, thus $\mathscr{N}_1 \models K^n(p \rightarrow q)$.

Case 4: φ is p. Then $\mathscr{N}_1 \models \varphi$ by definition.

Case 5: φ is $K^n p$ for some $n > 0$. Then $\mathscr{N}_1 \models \varphi$ by definition. □

Proof of Theorem 25. Let $T = \mathbf{K} \cup \mathbf{KK}$. Let T' be the theory consisting of:

- $K^n\varphi$ for all $n \in \mathbb{N}$ and all $\varphi \in \mathbf{K}$.
- $K^n\varphi$ for all $n \in \mathbb{N}$ and all $\varphi \in \mathbf{KK}$.

Clearly T' is closed and $T' \supseteq T$. Let p be a propositional atom. We will show $\mathscr{M}_{T',\emptyset} \not\models K(p \vee \neg p) \rightarrow KK(p \vee \neg p)$, showing $\mathscr{M}_{T',\emptyset} \not\models T$ and thus proving T is not closed generic. Clearly $T' \models p \vee \neg p$, thus $\mathscr{M}_{T',\emptyset} \models K(p \vee \neg p)$. It remains to show $\mathscr{M}_{T',\emptyset} \not\models KK(p \vee \neg p)$, i.e., that $T' \not\models K(p \vee \neg p)$.

Call a formula *bad* if it is either $p \vee \neg p$ or is of the form $\varphi_1 \rightarrow \cdots \rightarrow \varphi_n \rightarrow (p \vee \neg p)$. Let $\mathscr{N}$ be the model such that:

- For every propositional atom a, $\mathscr{N} \models a$.
- For every formula φ, $\mathscr{N} \models K\varphi$ iff φ is not bad.

Since $p \vee \neg p$ is bad, we have $\mathscr{N} \not\models K(p \vee \neg p)$. Thus to show $T' \not\models K(p \vee \neg p)$ (and thus finish the proof), it suffices to show $\mathscr{N} \models T'$. Let $\varphi \in T'$.

Case 1: $\varphi \in \mathbf{K}$. Then φ has the form $K(\psi \rightarrow \rho) \rightarrow K\psi \rightarrow K\rho$. Assume $\mathscr{N} \models K(\psi \rightarrow \rho)$ and $\mathscr{N} \models K\psi$. Then $\psi \rightarrow \rho$ is not bad. This implies ρ is not bad, thus $\mathscr{N} \models K\rho$, as desired.

Case 2: $\varphi \in \mathbf{KK}$. Then φ has the form $K\psi \rightarrow KK\psi$. Clearly $K\psi$ is not bad, thus $\mathscr{N} \models KK\psi$, thus $\mathscr{N} \models \varphi$.

Case 3: φ is of the form $K(K(\psi \rightarrow \rho) \rightarrow K\psi \rightarrow K\rho)$. Clearly $K(\psi \rightarrow \rho) \rightarrow K\psi \rightarrow K\rho$ is not bad, so $\mathscr{N} \models \varphi$.

Case 4: φ is of the form $K(K\psi \rightarrow KK\psi)$. Clearly $K\psi \rightarrow KK\psi$ is not bad, so $\mathscr{N} \models \varphi$.

Case 5: φ is $K^n\psi$ for some $\psi \in \mathbf{K} \cup \mathbf{KK}$ and some $n \geq 2$. Then φ has the form $KK\rho$ for some ρ. Clearly $K\rho$ is not bad, thus $\mathscr{N} \models KK\rho$. □

References

[1] Alessandro Aldini, Samuel A. Alexander & Pierluigi Graziani (2022): *Knowledge-of-own-Factivity, the Definition of Surprise, and a Solution to the Surprise Examination Paradox*. In: *CIFMA*, Springer, pp. 383–399, doi:10.1007/978303126236430.

[2] Samuel A. Alexander (2013): *An Axiomatic Version of Fitch's Paradox*. *Synthese* 190(12), pp. 2015–2020, doi:10.1007/s11229-011-9954-0.

[3] Samuel A Alexander (2014): *A Machine that Knows its own Code*. *Studia Logica*, pp. 567–576, doi:10.1007/s11225-013-9491-6.

[4] Samuel A Alexander (2015): *Fast-Collapsing Theories*. *Studia Logica* 103(1), pp. 53–73, doi:10.1007/s11225-013-9537-9.

[5] Samuel A. Alexander (2020): *Self-Referential Theories*. *The Journal of Symbolic Logic* 85(4), pp. 1687–1716, doi:10.1017/jsl.2020.54.

[6] C. Anthony Anderson (1983): *The Paradox of the Knower*. *The Journal of Philosophy* 80(6), pp. 338–355, doi:10.2307/2026335.

[7] Timothy J Carlson (2000): *Knowledge, Machines, and the Consistency of Reinhardt's Strong Mechanistic Thesis*. *Annals of Pure and Applied Logic* 105(1-3), pp. 51–82, doi:10.1016/S0168-0072(99)00048-2.

[8] Timothy Y Chow (1998): *The Surprise Examination or Unexpected Hanging Paradox*. *The American Mathematical Monthly* 105(1), pp. 41–51, doi:10.2307/2589525.

[9] Charles B Cross (2001): *The Paradox of the Knower without Epistemic Closure*. *Mind* 110(438), pp. 319–333, doi:10.1093/mind/110.438.319.

[10] Charles B Cross (2012): *The Paradox of the Knower without Epistemic Closure — Corrected*. *Mind* 121(482), pp. 457–466, doi:10.1093/mind/fzs067.

[11] Susan Haack (1978): *Philosophy of Logics*. Cambridge University Press, doi:10.1017/CBO9780511812866.

[12] David Kaplan & Richard Montague (1960): *A Paradox Regained*. *Notre Dame Journal of Formal Logic* 1(3), pp. 79–90, doi:10.1305/ndjfl/1093956549.

[13] William N Reinhardt (1985): *Absolute Versions of Incompleteness Theorems*. *Noûs*, pp. 317–346, doi:10.2307/2214945.

[14] Jan-Willem Romeijn, E Pacuit & AP Pedersen (2013): *When is an Example a Counterexample?* In: *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge: TARK 2013*, ACM Press, pp. 156–165, doi:10.48550/arXiv.1310.6432.

[15] Fredrik Stjernberg (2009): *Restricting Factiveness*. *Philosophical Studies* 146, pp. 29–48, doi:10.1007/s11098-008-9243-z.

[16] Mirjam de Vos, Rineke Verbrugge & Barteld Kooi (2023): *Solutions to the Knower Paradox in the Light of Haack's Criteria*. *Journal of Philosophical Logic*, pp. 1–32, doi:10.1007/s10992-023-09699-3.

Iterated Elimination of Weakly Dominated Strategies in Well-Founded Games

Krzysztof R. Apt
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
University of Warsaw
Warsaw, Poland
k.r.apt@cwi.nl

Sunil Simon
Department of CSE,
IIT Kanpur, Kanpur, India
simon@cse.iitk.ac.in

Recently, in [3], we studied well-founded games, a natural extension of finite extensive games with perfect information in which all plays are finite. We extend here, to this class of games, two results concerned with iterated elimination of weakly dominated strategies, originally established for finite extensive games.

The first one states that every finite extensive game with perfect information and injective payoff functions can be reduced by a specific iterated elimination of weakly dominated strategies to a trivial game containing the unique subgame perfect equilibrium. Our extension of this result to well-founded games admits transfinite iterated elimination of strategies. It applies to an infinite version of the centipede game. It also generalizes the original result to a class of finite games that may have several subgame perfect equilibria.

The second one states that finite zero-sum games with n outcomes can be solved by the maximal iterated elimination of weakly dominated strategies in $n-1$ steps. We generalize this result to a natural class of well-founded strictly competitive games.

1 Introduction

This paper is concerned with the iterated elimination of weakly dominated strategies (IEWDS) in the context of natural class of infinite extensive games with perfect information. While simple examples show that the deletion of weakly dominated strategies may result in removal of a unique Nash equilibrium, IEWDS has some merit if it results in solving a game. It is for instance used to show that the so-called "beauty contest" game has exactly one Nash equilibrium (see, e.g., [7, Chapter 5]). Other games can be solved this way, see, e.g., [11, pages 63, 110-114].

This procedure was also studied in the realm of finite extensive games with perfect information. In [8] the correspondence between the outcomes given by the iterated elimination of weakly dominated strategies and backward induction was investigated in the context of binary voting agendas with sequential voting. More recently, this procedure was studied in [16] in the context of supermodular games.

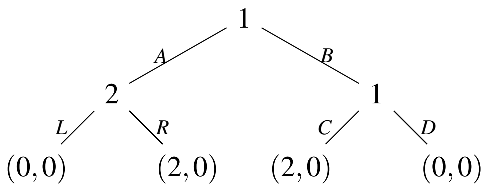
For arbitrary games two important results were established. The first one states, see [11], that in such games with injective payoff functions (such games are sometimes called *generic*) a specific iterated elimination of weakly dominated strategies (that mimics the backward induction) yields a trivial game which contains the unique subgame perfect equilibrium. It was noticed in [4] that this result holds for a slightly more general class of games *without relevant ties*.[1]

[1] All mentioned concepts are explained in Sections 2, 4, and 5. We did not find any precise proofs in the literature. The proof is briefly sketched in [11, pages 108-109] and summarized in [4, pages 48-49] as follows: "if backward induction deletes action a at node x, delete all the strategies reaching x and choosing a". We provided in [2] a detailed proof of the stronger result of [4] in which we clarified how the backward induction algorithm needs to be modified to achieve the desired outcome.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 16–30, doi:10.4204/EPTCS.379.5

The second result, due to [6], is concerned with finite extensive zero-sum games. It states that such games can be reduced to a trivial game by the 'maximal' iterated elimination of weakly dominated strategies in $n-1$ steps, where n is the number of outcomes.[2]

In [3] we studied a natural extension of finite extensive games with perfect information in which one assumes that all plays are finite. We called these games well-founded games.[3] The subject of this paper is to extend the above two results to well-founded games. In both cases some non-trivial difficulties arise.



	L	R
AC	(0,0)	(2,0)
AD	(0,0)	(2,0)
BC	(2,0)	(2,0)
BD	(0,0)	(0,0)

Figure 1: An extensive game G and the corresponding strategic game $\Gamma(G)$

Example 1 Consider the extensive game G and the corresponding strategic game $\Gamma(G)$ given in Figures 1. G has three subgame perfect equilibria which are all payoff equivalent: $\{(AC,R),(BC,L),(BC,R)\}$. We can observe that in $\Gamma(G)$ no sequence of iterated elimination of weakly dominated strategies results in a trivial game that contains all the subgame perfect equilibria in G. To see this, first note that the strategies L and R of player 2 are never weakly dominated irrespective of the elimination done with respect to the strategies of player 1. Also, note that the strategy BD of player 1 is strictly dominated by BC in $\Gamma(G)$. Thus the only possibility of reducing $\Gamma(G)$ to a trivial game is to eliminate all strategies of player 1 except BC. But this results in the elimination of (AC,R) which is a subgame perfect equilibrium in G. □

This might suggest that one should limit oneself to extensive games with a unique subgame perfect equilibrium. Unfortunately, this restriction does not work either as shown in Example 2. Additional complication arises when the game has no subgame perfect equilibrium as shown in 3.

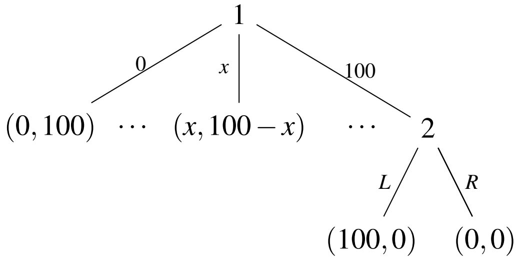


Figure 2: A game G with a unique SPE

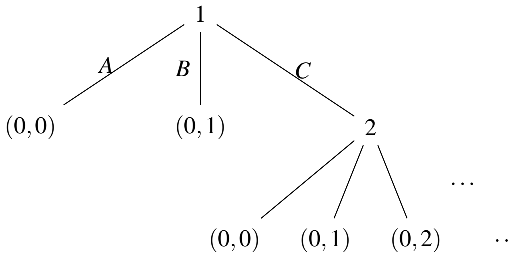


Figure 3: A game G with no SPE

Example 2 Consider a 'trimmed version' of the ultimatum game from [3] given in Figure 2, in which for each $x \in [0,100]$ the root has a direct descendant x. This game has a unique subgame perfect equilibrium, namely $(100,L)$. Consider an iterated elimination of weakly dominated strategies. For each strategy of player 1 the strategies L and R of player 2 yield the same payoff. So these two strategies are never

[2] An alternative proof given in [17] shows that the result holds for the larger class of strictly competitive games. In [2] we clarified that the original proof also holds for this class of games.

[3] In the economic literature such games are sometimes called 'games with finite horizon'.

eliminated. Further, strategy 100 of player 1 is never eliminated either, since for any strategy $x < 100$ we have $p_1(x,L) = x < 100 = p_1(100,L)$ and $p_1(x,R) = x > 0 = p_1(100,R)$. So the joint strategies $(100,L)$ and $(100,R)$ are never eliminated and they are not payoff equivalent. (In fact, each iterated elimination of weakly dominated strategies yields the game with the sets of strategies $\{100\}$ and $\{L,R\}$.) □

Example 3 Consider the well-founded game G given in Figure 3. Clearly G has no subgame perfect equilibrium. Further, strategies A and B of player 1 yield the same outcome for him, so cannot be eliminated by any iterated elimination of weakly dominated strategies. Thus any result of such an elimination contains at least two outcomes, $(0,0)$ and $(0,1)$. So G cannot be reduced to a trivial game. □

To address these issues, we introduce the concept of an *SPE-invariant* well-founded game. These are games in which subgame perfect equilibria exist and moreover in each subgame such equilibria are payoff equivalent. Then we show that the first result can be extended to such games. In view of the above examples it looks like the strongest possible generalization of the original result. In particular, it applies to an infinite version of the well-known centipede game of [15].

This result calls for a careful extension of the iterated elimination of weakly dominated strategies to infinite games: its stages have to be indexed by ordinals and one has to take into account that the outcome can be the empty game.

When limited to finite games, our theorem extends the original result. In particular it applies to the class of extensive games that satisfy the ***transference of decisionmaker indifference (TDI)*** condition due to [10], a class that includes strictly competitive games. We also show that the well-founded games with finitely many outcomes that satisfy the TDI condition are SPE-invariant. Also when extending the second result, about strictly competitive games, to well-founded games one has to be careful. The original proof crucially relies on the fact that finite extensive zero-sum games have a value. Fortunately, as we showed in [3], well-founded games with finitely many outcomes have a subgame perfect equilibrium, so a fortiori a Nash equilibrium, which suffices to justify the relevant argument (Lemma 21 in Section 5).

By carefully checking of the crucial steps of the original proof we extend the original result to a class of well-founded strictly competitive games that includes *almost constant* games, in which for all but finitely many leaves the outcome is the same. It remains an open problem whether this result holds for all strictly competitive games with finitely many outcomes.

IEWDS is one of the early approaches applied to analyze strategies and extensive games. It does not take into account epistemic reasoning of players in the presence of assumptions such as common knowledge of rationality. The vast literature on this subject, starting with [5] and [12], led to identification of several more informative ways of analyzing finite extensive games with imperfect information. We just mention here two representative references. In [4] Pearce's notion of *extensive form rationalizability* (EFR) was studied and it was shown that for extensive games without relevant ties it coincides with the IEWDS. A more general notion of common belief in future rationality was studied in [13] that led to identification of a new iterative elimination procedure called *backward dominance*.

In our paper IEWDS is defined as a transfinite elimination procedure. A number of papers, starting with [9], analyzed when such a transfinite elimination of strategies cannot be reduced to an iteration over ω steps. In our framework it is a simple consequence of the fact that the ranks of the admitted game trees can be arbitrary ordinals. In particular, an infinite version of the centipede game considered in Example 12 requires more than ω elimination rounds.

2 Preliminaries

2.1 Strategic games

A ***strategic game*** $H = (H_1,\ldots,H_n,p_1,\ldots,p_n)$ consists of a set of players $\{1,\ldots,n\}$, where $n \geq 1$, and for each player i, a set H_i of ***strategies*** along with a ***payoff function*** $p_i : H_1 \times \cdots \times H_n \rightarrow \mathbb{R}$.

We call each element of $H_1 \times \cdots \times H_n$ a ***joint strategy*** of players $1,\ldots,n$, denote the ith element of $s \in H_1 \times \cdots \times H_n$ by s_i, and abbreviate the sequence $(s_j)_{j \neq i}$ to s_{-i}. We write (s'_i, s_{-i}) to denote the joint strategy in which player's i strategy is s'_i and each other player's j strategy is s_j. Occasionally we write (s_i, s_{-i}) instead of s. Finally, we abbreviate the Cartesian product $\times_{j \neq i} H_j$ to H_{-i}.

Given a joint strategy s, we denote the sequence $(p_1(s),\ldots,p_n(s))$ by $p(s)$ and call it an ***outcome*** of the game. We say that H ***has*** k ***outcomes*** if $|\{p(s) \mid s \in H_1 \times \cdots \times H_n\}| = k$ and call a game ***trivial*** if it has one outcome. If one of the sets H_i is empty, we call the game ***empty*** and ***non-empty*** otherwise. Unless explicitly stated, all used strategic games are assumed to be non-empty. We say that two joint strategies s and t are ***payoff equivalent*** if $p(s) = p(t)$.

We call a joint strategy s a ***Nash equilibrium*** if $\forall i \in \{1,\ldots,n\} \forall s'_i \in H_i : p_i(s_i, s_{-i}) \geq p_i(s'_i, s_{-i})$. When the number of players and their payoff functions are known we can identify the game H with the set of strategies in it.

By a ***subgame*** of a strategic game H we mean a game obtained from H by removing some strategies. Given a set $\mathscr{J}$ of subgames of a strategic game H we define $\bigcap \mathscr{J}$ as the subgame of H in which for each player i his set of strategies is $\bigcap_{J \in \mathscr{J}} J_i$. Also, given two subgames H' and H'' of a strategic game H we write $H' \subseteq H''$ if for each player i, $H'_i \subseteq H''_i$.

Consider two strategies s_i and s'_i of player i in a strategic game H. We say that s_i ***weakly dominates*** s'_i (or equivalently, that s'_i is ***weakly dominated by*** s_i) in H if $\forall s_{-i} \in H_{-i} : p_i(s_i, s_{-i}) \geq p_i(s'_i, s_{-i})$ and $\exists s_{-i} \in H_{-i} : p_i(s_i, s_{-i}) > p_i(s'_i, s_{-i})$.

In what follows, given a strategic game we consider, possibly transfinite, sequences of sets of strategies. They are written as $(\rho_\alpha, \alpha < \gamma)$, where α ranges over all ordinals smaller than some ordinal γ. Given two such sequences $\rho := (\rho_\alpha, \alpha < \gamma)$ and $\rho' := (\rho'_{\alpha'}, \alpha' < \gamma')$, we denote by (ρ, ρ') their concatenation (which is indexed by $\gamma + \gamma'$), by ρ^β the subsequence $(\rho_\alpha, \alpha < \beta)$ of ρ, and for $\alpha < \beta$ by $\rho^{\beta-\alpha}$ the subsequence such that $(\rho^\alpha, \rho^{\beta-\alpha}) = \rho^\beta$. Further, we write $H \rightarrow^\rho H'$ to denote the fact that the game H' is the outcome of the iterated elimination from the non-empty game H of the sets of strategies that form ρ. In each step all eliminated strategies are weakly dominated in the current game. As a result H' may be empty. The relation $\rightarrow^\rho$ is defined as follows.

If $\rho = (\rho_0)$, that is, if $\gamma = 1$, then $H \rightarrow^\rho H'$ holds if each strategy in the set ρ_0 is weakly dominated in H and H' is the outcome of removing from H all strategies from ρ_0. If γ is a successor ordinal > 1, say $\gamma = \delta + 1$, and $H \rightarrow^{\rho'} H'$, $H' \rightarrow^{(\rho_\delta)} H''$, where H' is non-empty, and $\rho' := (\rho_\alpha, \alpha < \delta)$, then $H \rightarrow^\rho H''$. Finally, if γ is a limit ordinal and for all $\beta < \gamma$, $H \rightarrow^{\rho^\beta} H^\beta$, then $H \rightarrow^\rho \bigcap_{\beta < \gamma} H^\beta$. In general, the strategic game H from which we eliminate strategies will be a subgame of a game $\Gamma(G)$, where G is an extensive game (to be defined shortly). It will be then convenient to allow in ρ strategies from $\Gamma(G)$. In the definition of $H \rightarrow^\rho H'$ we then disregard the strategies from ρ that are not from H. In the proofs below we rely on the following observations about the $\rightarrow^\rho$ relation, the proofs of which we omit.

Note 4

(i) Suppose $H \rightarrow^\rho H'$ and $H' \rightarrow^{\rho'} H''$, where H' is non-empty. Then $H \rightarrow^{(\rho,\rho')} H''$.

(ii) Suppose $H \rightarrow^\rho H'$, where $\rho = (\rho_\alpha, \alpha < \gamma)$ and γ is a limit ordinal. Suppose further that for a sequence of ordinals $(\alpha_\delta)_{\delta < \varepsilon}$ converging to γ we have $H \rightarrow^{\rho^{\alpha_\delta}} H^{\alpha_\delta}$ for all $\delta < \varepsilon$. Then $H' = \bigcap_{\delta < \varepsilon} H^{\alpha_\delta}$.

2.2 Well-founded games

We recall from [3] the definition of a well-founded game. A ***tree*** is an acyclic directed connected graph, written as (V,E), where V is a non-empty set of nodes and E is a possibly empty set of edges. An ***extensive game with perfect information*** $(T, turn, p_1,\ldots,p_n)$ consists of a set of players $\{1,\ldots,n\}$, where $n \geq 1$ along with the following. A ***game tree***, which is a tree $T := (V,E)$ with a ***turn function*** $turn : V \setminus Z \to \{1,\ldots,n\}$, where Z is the set of leaves of T. For each player i a ***payoff function*** $p_i : Z \to \mathbb{R}$, for each player i. The function *turn* determines at each non-leaf node which player should move. The edges of T represent possible ***moves*** in the considered game, while for a node $v \in V \setminus Z$ the set of its children $C(v) := \{w \mid (v,w) \in E\}$ represents possible ***actions*** of player $turn(v)$ at v.

We say that an extensive game with perfect information is ***finite***, ***infinite***, or ***well-founded*** if, respectively, its game tree is finite, infinite, or well-founded. Recall that a tree is called ***well-founded*** if it has no infinite paths. *From now on by an* **extensive game** *we mean a well-founded extensive game with perfect information.*

For a node u in T we denote the subtree of T rooted at u by T^u. In the proofs we shall often rely on the concept of a *rank* of a well-founded tree T, defined inductively as follows, where v is the root of T:

$$rank(T) := \begin{cases} 0 & \text{if } T \text{ has one node} \\ sup\{rank(T^u) + 1 \mid u \in C(v)\} & \text{otherwise,} \end{cases}$$

where $sup(X)$ denotes the least ordinal larger than all ordinals in the set X.

For an extensive game $G := (T, turn, p_1,\ldots,p_n)$ let $V_i := \{v \in V \setminus Z \mid turn(v) = i\}$. So V_i is the set of nodes at which player i moves. A ***strategy*** for player i is a function $s_i : V_i \to V$, such that $(v, s_i(v)) \in E$ for all $v \in V_i$. We denote the set of strategies of player i by S_i. Let $S = S_1 \times \cdots \times S_n$. As in the case of the strategic games we use the '$-i$' notation, when referring to sequences of strategies or sets of strategies.

Each joint strategy $s = (s_1,\ldots,s_n)$ determines a rooted path $play(s) := (v_1,\ldots,v_m)$ in T defined inductively as follows. v_1 is the root of T and if $v_k \notin Z$, then $v_{k+1} := s_i(v_k)$, where $turn(v_k) = i$. So when the game tree consists of just one node, v, we have $play(s) = v$. Informally, given a joint strategy s, we can view $play(s)$ as the resulting *play* of the game. For each joint strategy s the rooted path $play(s)$ is finite since the game tree is assumed to be well-founded. Denote by $leaf(s)$ the last element of $play(s)$. To simplify the notation we just write everywhere $p_i(s)$ instead of $p_i(leaf(s))$.

With each extensive game $G := (T, turn, p_1,\ldots,p_n)$ we associate a strategic game $\Gamma(G)$ defined as follows. $\Gamma(G) := (S_1,\ldots,S_n, p_1,\ldots,p_n)$, where each S_i is the set of strategies of player i in G. In the degenerate situation when the game tree consists of just one node, each strategy is the empty function, denoted by $\emptyset$, and there is only one joint strategy, namely the n-tuple $(\emptyset,\ldots,\emptyset)$ of these functions. In that case we just stipulate that $p_i(\emptyset,\ldots,\emptyset) = 0$ for all players i. All notions introduced in the context of strategic games can now be reused in the context of an extensive game G simply by referring to the corresponding strategic form $\Gamma(G)$. In particular, the notion of a Nash equilibrium is well-defined.

The ***subgame*** of an extensive game $G := (T, turn, p_1,\ldots,p_n)$, rooted at the node w and denoted by G^w, is defined as follows. The set of players is $\{1,\ldots,n\}$, the game tree is T^w. The *turn* and payoff functions are the restrictions of the corresponding functions of G to the nodes of T^w. We call G^w a ***direct subgame*** of G if w is a child of the root v.

Note that some players may 'drop out' in G^w, in the sense that at no node of T^w it is their turn to move. Still, to keep the notation simple, it is convenient to admit in G^w all original players in G.

Each strategy s_i of player i in G uniquely determines his strategy s_i^w in G^w. Given a joint strategy $s = (s_1,\ldots,s_n)$ of G we denote by s^w the joint strategy $(s_1^w,\ldots,s_n^w)$ in G^w. Further, we denote by S_i^w the set of strategies of player i in the subgame G^w and by S^w the set of joint strategies in this subgame.

Finally, a joint strategy s of G is called a ***subgame perfect equilibrium*** in G if for each node w of T, the joint strategy s^w of G^w is a Nash equilibrium in the subgame G^w. We denote by $SPE(G)$ the set of subgame perfect equilibria in G. Finally, we say that a game is ***SPE-invariant*** if it has a subgame perfect equilibrium and in each subgame of it all subgame perfect equilibria are payoff equivalent.

We shall often use the following result.

Theorem 5 ([3]) *Every extensive game with finitely many outcomes has a subgame perfect equilibrium.*

3 Preliminary lemmas

In this section we present a sequence of lemmas needed to prove our first main result. In the proofs we often switch between a game and its direct subgames.

Consider an extensive game $G := (T, turn, p_1, \ldots, p_n)$ with the root v and a child w of v. For each player j to each of his strategy t_j in a direct subgame G^w there corresponds a natural set $[t_j]$ of his strategies in the game G defined by $[t_j] := \{s_j \mid t_j = s^w_j \text{ and } s_j(v) = w \text{ if } j = turn(v)\}$. So for a player j, $[t_j]$ is the set of his strategies in G the restriction of which to G^w is t_j, with the additional proviso that if $j = turn(v)$, then each strategy in $[t_j]$ selects w at the root v. We call $[t_j]$ the *lifting* of t_j to the game G. The following lemma clarifies the relevance of lifting.

Lemma 6 *Consider a direct subgame G^w of G. Suppose that the strategy t_j is weakly dominated in G^w. Then each strategy in $[t_j]$ is weakly dominated in G.*

Proof. Suppose that t_j is weakly dominated in G^w by some strategy u_j. Take a strategy v_j in $[t_j]$. We show that v_j is weakly dominated in G by the strategy w_j in $[u_j]$ that coincides with v_j on all the nodes that do not belong to G^w. So w_j is obtained from v_j by replacing in it v^w_j, i.e., t_j, by u_j. Below s_{-j} denotes a sequence of strategies in G of the opponents of player j.

Case 1. $j = turn(v)$.

By the choice of u_j for all s_{-j} $p_j(t_j, s^w_{-j}) \leq p_j(u_j, s^w_{-j})$ and for some s_{-j} $p_j(t_j, s^w_{-j}) < p_j(u_j, s^w_{-j})$. Further, by the definition of $[\cdot]$ we have $v_j(v) = w$, so for all s_{-j} we have $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(u_j, s^w_{-j}) = p_j(w_j, s_{-j})$, so the claim follows.

Case 2. $j \neq turn(v)$.

Let $i = turn(v)$. Take some s_{-j}. If $s_i(v) = w$, then $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(w_j, s_{-j}) = p_j(u_j, s^w_{-j})$. Thus $p_j(v_j, s_{-j}) \leq p_j(w_j, s_{-j})$ by the choice of u_j and w_j. Further, if $s_i(v) \neq w$, then $p_j(v_j, s_{-j}) = p_j(w_j, s_{-j})$ by the choice of w_j.

Choose an arbitrary s_{-j} such that $s_i(v) = w$ and $p_j(t_j, s^w_{-j}) < p_j(u_j, s^w_{-j})$. By the choice of s_i we have $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(w_j, s_{-j}) = p_j(u_j, s^w_{-j})$, so $p_j(v_j, s_{-j}) < p_j(w_j, s_{-j})$. Thus the claim follows. □

We now extend the notation $[\cdot]$ to sets of strategies and sequences of sets strategies. First, given a set of strategies A in a direct subgame G^w of G we define $[A] := \bigcup_{s_j \in A} [s_j]$. Next, given a sequence ρ of sets of strategies of players, each set taken from a direct subgame of G, we denote by $[\rho]$ the corresponding sequence of sets of strategies of players in G obtained by replacing each element A in ρ by $[A]$.

Given a set A of strategies of players in a direct subgame G^w we define the corresponding set of strategies in the game G by putting $\langle A \rangle = \{s_j \mid s^w_j \in A\}$. Thus for a set A of strategies in a direct subgame G^w, the set $\langle A \rangle$ differs from $[A]$ in that we do include in the former set strategies s_j for which $s_j(v) \neq w$. Given a set A of strategies of player j in the subgame G^w, we call $\langle A \rangle$ an *extension* of A to the game G.

Further, given a subgame H of $\Gamma(G^w)$, we define $\langle H\rangle$ as the subgame of $\Gamma(G)$ in which for each player j we have $\langle H\rangle_j = \langle H_j\rangle$.

In what follows we need a substantially strengthened version of Lemma 6 that relies on the following concept. Given an extensive game G with a root v, we say that a non-empty subgame J of $\Gamma(G)$ ***does not depend on*** a direct subgame G^w if for any strategy s_j from J any modification of it on the non-leaf nodes of G^w or on v if $turn(v) = j$ is also in J. Note that in particular $\Gamma(G)$ does not depend on any of its direct subgame and that for any non-empty subgame H of a direct subgame G^w of G the subgame $\langle H\rangle$ does not depend on any other direct subgame of G.

Lemma 7 *Consider a direct subgame G^w of G, subgames H and H' of $\Gamma(G^w)$ and a set A of strategies in H. Suppose that $H \rightarrow^A H'$ and that the subgame J of $\Gamma(G)$ does not depend on G^w. Then $J\cap\langle H\rangle \rightarrow^{[A]} J\cap\langle H'\rangle$.*

Proof. Take a strategy v_j in $[A]$. For some strategy t_j from A that is weakly dominated in H by some strategy u_j we have $v_j \in [t_j]\cap J_j$. Select a strategy w_j in $[u_j]$ that coincides with v_j on the nodes that do not belong to G^w. So w_j is a modification of v_j on the non-leaf nodes of G^w and consequently, by the assumption about J, it is in J_j. Further, w_j is in $\langle H\rangle$, since u_j is from H.

We claim that v_j is weakly dominated in $J\cap\langle H\rangle$ by w_j. Below s_{-j} denotes a sequence of strategies of the opponents of player j in the original game G.

Case 1. $j = turn(v)$.

By the choice of u_j for all s_{-j} such that $s^w_{-j} \in H_{-j}$ $p_j(t_j, s^w_{-j}) \le p_j(u_j, s^w_{-j})$ and for some s_{-j} such that $s^w_{-j} \in H_{-j}$ $p_j(t_j, s^w_{-j}) < p_j(u_j, s^w_{-j})$. By the definition of 'does not depend on' and the fact that $j = turn(v)$ we can also assume that the latter s_{-j} is from J_{-j} by stipulating that $s_{-j} = t_{-j}$ for an arbitrary joint strategy t from J.

Further, by the definition of $[\cdot]$ we have $v_j(v) = w$, so for all s_{-j} such that $s^w_{-j} \in H_{-j}$ we have $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(u_j, s^w_{-j}) = p_j(w_j, s_{-j})$. Hence for all s_{-j} $p_j(v_j, s_{-j}) \le p_j(w_j, s_{-j})$ and for some s_{-j} such that $s_{-j} \in J_{-j}$ and $s^w_{-j} \in H_{-j}$ (i.e., for some $s_{-j} \in (J\cap\langle H\rangle)_{-j}$) $p_j(v_j, s_{-j}) < p_j(w_j, s_{-j})$. This establishes the claim.

Case 2. $j \ne turn(v)$.

Let $i = turn(v)$. Take some s_{-j}. If $s_i(v) = w$, then $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(w_j, s_{-j}) = p_j(u_j, s^w_{-j})$. Thus $p_j(v_j, s_{-j}) \le p_j(w_j, s_{-j})$ by the choice of u_j and w_j. Further, if $s_i(v) \ne w$, then $p_j(v_j, s_{-j}) = p_j(w_j, s_{-j})$ by the choice of w_j. So for all s_{-j} we have $p_j(v_j, s_{-j}) \le p_j(w_j, s_{-j})$.

Choose an arbitrary s_{-j} such that $s_i(v) = w$, $s^w_{-j} \in H_{-j}$, and $p_j(t_j, s^w_{-j}) < p_j(u_j, s^w_{-j})$. Additionally, we can claim that $s_{-j} \in J_{-j}$ by stipulating that $s_{-j} = t_{-j}$ for an arbitrary joint strategy t from J. Then $s_{-j} \in (J\cap\langle H\rangle)_{-j}$.

By the choice of s_i we have $p_j(v_j, s_{-j}) = p_j(t_j, s^w_{-j})$ and $p_j(w_j, s_{-j}) = p_j(u_j, s^w_{-j})$, so $p_j(v_j, s_{-j}) < p_j(w_j, s_{-j})$. This establishes the claim for this case. □

We continue with some lemmas concerned with the relation $\rightarrow^\rho$.

Lemma 8 *Consider a direct subgame G^w of G. Suppose that for some sequence ρ of sets of strategies of players in G^w and a subgame H of $\Gamma(G^w)$, $\Gamma(G^w) \rightarrow^\rho H$. Suppose further that the subgame J of $\Gamma(G)$ does not depend on G^w. Then $J \rightarrow^{[\rho]} J\cap\langle H\rangle$.*

Proof. We proceed by transfinite induction on the length γ of $\rho = (\rho_\alpha, \alpha < \gamma)$.

Case 1. $\gamma = 1$.

By Lemma 7 $J\cap\langle\Gamma(G^w)\rangle \rightarrow^{[\rho_0]} J\cap\langle H\rangle$, so the claim holds since $\langle\Gamma(G^w)\rangle = \Gamma(G)$ and $J\cap\Gamma(G) = J$.

Case 2. γ is a successor ordinal > 1.

Suppose $\gamma = \delta + 1$. Then $\rho = (\rho', \rho_\delta)$, where $\rho' := (\rho_\alpha, \alpha < \delta)$. By definition for some H' we have $\Gamma(G^w) \to^{\rho'} H'$ and $H' \to^{\rho_\delta} H$. By the induction hypothesis $J \to^{[\rho']} J \cap \langle H' \rangle$ and by Lemma 7 $J \cap \langle H' \rangle \to^{[\rho_\delta]} J \cap \langle H \rangle$, so the claim follows by Note 4(i), since $[\rho] = ([\rho'], [\rho_\delta])$.

Case 3. γ is a limit ordinal.

By definition for some games H^β, where $\beta < \gamma$, we have $\Gamma(G^w) \to^{\rho^\beta} H^\beta$ and $H = \bigcap_{\beta<\gamma} H^\beta$, where—recall—$\rho^\beta = (\rho_\alpha, \alpha < \beta)$. By the induction hypothesis for all $\beta < \gamma$, we have $J \to^{[\rho^\beta]} J \cap \langle H^\beta \rangle$. So by definition $J \to^{[\rho]} J \cap \langle H \rangle$, since $J \cap \langle H \rangle = \bigcap_{\beta<\gamma} \langle J \cap H^\beta \rangle$ as $\langle H \rangle = \bigcap_{\beta<\gamma} \langle H^\beta \rangle$. □

Lemma 9 *Consider an extensive game G with the root v. Suppose that $(w_\alpha, \alpha < \gamma)$ is a sequence of children of v and that for all $\alpha < \gamma$, ρ_α is a sequence of sets of strategies in the direct subgame G^{w_α}. Suppose further that for each $\alpha < \gamma$ $\Gamma(G^{w_\alpha}) \to^{\rho_\alpha} H^{w_\alpha}$, where each game H^{w_α} is non-empty. Let ρ be the concatenation of the sequences $(\rho_\alpha, \alpha < \gamma)$. Then $\Gamma(G) \to^{[\rho]} \bigcap_{\alpha<\gamma} \langle H^{w_\alpha} \rangle$.*

By assumption each H^{w_α} is a non-empty subgame of $\Gamma(G^{w_\alpha})$, so each $\langle H^{w_\alpha} \rangle$ is a non-empty subgame of $\Gamma(G)$, and consequently $\bigcap_{\alpha<\gamma} \langle H^{w_\alpha} \rangle$ is also a non-empty subgame of $\Gamma(G)$.

Informally, suppose that for each direct subgame G^{w_α} of G we can reduce the corresponding strategic game $\Gamma(G^{w_\alpha})$ to a non-empty game H^{w_α}. Then the strategic game $\Gamma(G)$ can be reduced to a strategic game the strategies of which are obtained by intersecting for each player the extensions of his strategy sets in all games H^{w_α}. To establish this lemma we do not assume that $(w_\alpha, \alpha < \gamma)$ contains all children of v, which makes it possible to proceed by induction.

Proof. We proceed by transfinite induction on the length γ of ρ.

Case 1. $\gamma = 1$. Follows from Lemma 8 with $J = \Gamma(G)$.

Case 2. γ is a successor ordinal > 1.

Suppose $\gamma = \delta + 1$. By the induction hypothesis $\Gamma(G) \to^{[\rho^\delta]} \bigcap_{\alpha<\delta} \langle H^{w_\alpha} \rangle$, where ρ^δ is the concatenation of the sequences $(\rho_\alpha, \alpha < \delta)$. We also have by assumption $\Gamma(G^{w_\delta}) \to^{\rho_\delta} H^{w_\delta}$.

Note that the subgame $\bigcap_{\alpha<\delta} \langle H^{w_\alpha} \rangle$ of $\Gamma(G)$ does not depend on G^{w_δ}, so by Lemma 8 we have that $\bigcap_{\alpha<\delta} \langle H^{w_\alpha} \rangle \to^{[\rho_\delta]} \bigcap_{\alpha<\delta} \langle H^{w_\alpha} \rangle \cap \langle H^{w_\delta} \rangle$. By Note 4($i$) the claim follows.

Case 3. γ is a limit ordinal.

By the induction hypothesis for all $\beta < \gamma$ $\Gamma(G) \to^{[\rho^\beta]} \bigcap_{\alpha<\beta} \langle H^{w_\alpha} \rangle$, where ρ^β is the concatenation of the sequences $(\rho_\alpha, \alpha < \beta)$. Then by Note 4($ii$) and by definition $\Gamma(G) \to^{[\rho]} \bigcap_{\beta<\gamma} \bigcap_{\alpha<\beta} \langle H^{w_\alpha} \rangle$. But $\bigcap_{\beta<\gamma} \bigcap_{\alpha<\beta} \langle H^{w_\alpha} \rangle = \bigcap_{\alpha<\gamma} \langle H^{w_\alpha} \rangle$, so the claim follows. □

The next lemma shows that when each subgame H^{w_α} of $\Gamma(G^{w_\alpha})$ is trivial, under some natural assumptions the subgame $\bigcap_{\alpha<\gamma} \langle H^{w_\alpha} \rangle$ of $\Gamma(G)$ can then be reduced in one step to a trivial game.

Lemma 10 *Consider an extensive game G with the root v. Suppose that*

(a) G has a subgame perfect equilibrium and all subgame perfect equilibria of G are payoff equivalent,

(b) for all $w \in C(v)$, $SPE(G^w) \subseteq H^w$, where H^w is a trivial subgame of $\Gamma(G^w)$.

Then for some set of strategies A we have $\bigcap_{w\in C(v)} \langle H^w \rangle \to^A H'$, where H' a trivial game and $SPE(G) \subseteq H'$.

Proof. Let $H := \bigcap_{w \in C(v)} \langle H^w \rangle$. Note that H is a non-empty subgame of $\Gamma(G)$.

Denote the unique outcome in the game H^w by val^w, i.e., for all joint strategies s in H^w we have $p(s) = val^w$. Then the possible outcomes in H are val^w, where $w \in C(v)$. More precisely, suppose that $i = turn(v)$. Then if s is a joint strategy in H, then $p(s) = val^w$, where $s_i(v) = w$.

Take two strategies t'_i and t''_i of player i in H with $t'_i(v) = w_1$ and $t''_i(v) = w_2$ such that $val_i^{w_1} < val_i^{w_2}$. This means that for any joint strategies s_{-i} from H_{-i} we have $p_i(t'_i, s_{-i} < p_i(t''_i, s_{-i}$, so t'_i is weakly dominated in H by t''_i (actually, even strictly dominated).

By assumption *(a)* G has a subgame perfect equilibrium, so by Corollary 7 of [3] $\max\{val_i^w \mid w \in C(v)\}$ exists. Denote it by val_i and let $W := \{w \in C(v) \mid val_i^w = val_i\}$. So W is the set of children w of v for which the corresponding value val_i^w is maximal. Finally, let A be the set of strategies t_i of player i in H such that $t_i(v) \notin W$.

By the above observation about t'_i and t''_i all strategies in A are weakly dominated in H. By removing them from H we get a game H' with the unique payoff val_i for player i. To prove that H' is trivial consider two joint strategies s and t in H'. Suppose that $s_i(v) = w_1$ and $t_i(v) = w_2$. Then $w_1, w_2 \in W$, $s^{w_1} \in H^{w_1}$, $t^{w_2} \in H^{w_2}$, $p(s) = p(s^{w_1})$, and $p(t) = p(t^{w_2})$.

By Theorem 8 of [3] subgame perfect equilibria u' and u'' in G exist such that $u'_i(v) = w_1$, $(u')^{w_1}$ is a subgame perfect equilibrium in G^{w_1}, $u''_i(v) = w_2$, and $(u'')^{w_2}$ is a subgame perfect equilibrium in G^{w_2}. Then $p(u') = p((u')^{w_1})$ and $p(u'') = p((u'')^{w_2})$, so $p((u')^{w_1}) = p((u'')^{w_2})$ by assumption *(a)*. Further, by assumption *(b)* both $(u')^{w_1} \in H^{w_1}$ and $(u'')^{w_2} \in H^{w_2}$, so since both subgames are trivial, $p(s^{w_1}) = p((u')^{w_1})$ and $p(t^{w_2}) = p((u')^{w_2})$. Consequently $p(s) = p(t)$, which proves that H' is trivial.

To prove that $SPE(G) \subseteq H'$ consider a subgame perfect equilibrium s in G. Take some $u \in C(v)$. By assumption *(b)*, $s^u \in H^u$, so $p_i(s^u) = val_i^u$ and, by the definition of $\langle \cdot \rangle$, $s \in H$. Suppose that $s_i(v) = w$. By Corollary 7 of [3] $val_i^w = val_i$, i.e., $s_i(v) \in W$. This means that $s_i \notin A$ and thus $s \in H'$. □

4 SPE-invariant games

We can now prove the desired result.

Theorem 11 *Consider an SPE-invariant extensive game G. There exists a sequence ρ of strategies of players in G and a subgame H of $\Gamma(G)$ such that $\Gamma(G) \rightarrow^{\rho} H$, H is trivial and $SPE(G) \subseteq H$.*

Proof. We proceed by induction on the rank of the game tree of G. For game trees of rank 0 all strategies are empty functions, so $\Gamma(G)$ is a trivial game with the unique joint strategy $(\emptyset, \ldots, \emptyset)$ and $SPE(G) = \{(\emptyset, \ldots, \emptyset)\}$, so the claim holds. Suppose that the rank of the game tree of G is $\alpha > 0$ and assume that claim holds for all extensive games with the game trees of rank smaller than α.

Let v be the root of G. Each direct subgame of G is SPE-invariant, so by the induction hypothesis for all $w \in C(v)$ there exists a sequence ρ^w of strategies of players in G^w and a subgame H^w of $\Gamma(G^w)$ such that $\Gamma(G^w) \rightarrow^{\rho^w} H^w$, H^w is trivial and $SPE(G^w) \subseteq H^w$. The claim now follows by Lemmas 9 and 10. □

The following example illustrates the use of this theorem. An extensive game is called ***generic*** if each payoff function is an injective.

Example 12 Recall that the centipede game, introduced in [15] (see also [11, pages 106-108]), is a two-players extensive game played for an even number of periods. We define it inductively as follows. The game with 2 periods is depicted in Figure 4. Here and below the argument of each non-leaf is the player whose turn is to move, and the leaves are followed by players' payoffs. The moves are denoted by

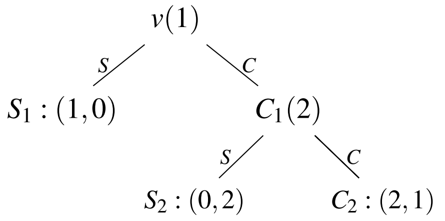


Figure 4: Centipede game with 2 periods

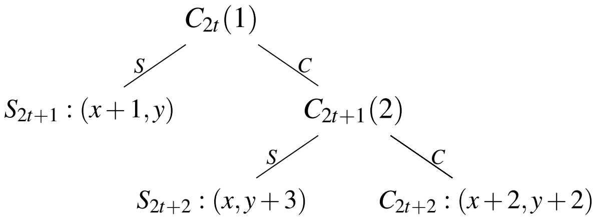


Figure 5: From t to $t+2$ periods

the letters C and S. The game with $2t+2$ periods is obtained from the game with $2t$ periods by replacing the leaf C_{2t} by the tree depicted in Figure 5.

By the the result of [11, pages 108-109]) each centipede game can be reduced by an iterated elimination of weakly dominated strategies to a trivial game which contains the unique subgame perfect equilibrium, with the outcome $(1,0)$. We now show that the same holds for an infinite version of the centipede game G in which player 2 begins the game by selecting an even number $2t > 0$. Subsequently, the centipede version with $2t$ periods is played.

Note that G is SPE-invariant. Indeed, G has infinitely many subgame perfect equilibria (one for each first move of player 2), but each of them yields the outcome $(1,0)$. Moreover, each subgame of G is either a centipede game with $2t$ periods for some $t > 0$, or a subgame of such a game. So each subgame of G is a finite generic game and thus has a unique subgame perfect equilibrium.

By Theorem 11 we can reduce G by an infinite iterated elimination of weakly dominated strategies to a trivial game which contains all its subgame perfect equilibria. Note that the strategy elimination sequence constructed in the proof of this theorem consists of for more than ω steps. □

For finite extensive games, Theorem 11 extends the original result reported in [11, pages 108-109]. Namely, the authors prove the corresponding result for finite extensive games that are generic. In such games a unique subgame perfect equilibrium exists, while we only claim that the game is SPE-invariant.

To clarify the relevance of this relaxation let us mention two classes of well-founded extensive games that are SPE-invariant and that were studied for finite extensive games. Following [4] we say that an extensive game $(T, turn, p_1, \ldots, p_n)$ is ***without relevant ties*** if for all non-leaf nodes u in T the payoff function p_i, where $turn(u) = i$, is injective on the leaves of T^u. This is a more general property than being generic. The relevant property for finite extensive games is that a game without relevant ties has a unique subgame perfect equilibrium, see [2] for a straightforward proof. In the case of well-founded games a direct modification of this proof, that we omit, shows that every extensive game without relevant ties has at most one subgame perfect equilibrium. Further, if a game is without relevant ties, then so is every subgame of it, so we conclude that well-founded games without relevant ties are SPE-invariant.

Next, following [10] we say that an extensive game $(T, turn, p_1, \ldots, p_n)$ satisfies the ***transference of decisionmaker indifference (TDI)*** condition if:

$$\forall i \in \{1,\ldots,n\}\, \forall r_i, t_i \in S_i\, \forall s_{-i} \in S_{-i}$$
$$[p_i(leaf(r_i, s_{-i})) = p_i(leaf(t_i, s_{-i})) \to p(leaf(r_i, s_{-i})) = p(leaf(t_i, s_{-i}))].$$

where S_i is the set of strategies of player i. Informally, this condition states that whenever for some player i, two of his strategies r_i and t_i are indifferent w.r.t. some joint strategy s_{-i} of the other players then this indifference extends to all players.

Strategic games that satisfy the TDI condition are of interest because of the main result of [10] which states that in finite games that satisfy this condition iterated elimination of weakly dominated strategies is order independent.[4] The authors also give examples of natural games that satisfy this condition. Also strictly competitive games studied in the next section satisfy this condition.

The following result extends an implicit result of [10] to well-founded games.

Theorem 13 *Consider an extensive game G. Suppose that G has finitely many outcomes and G satisfies the TDI condition. Then G is SPE-invariant.*

Proof. We reduce the game G to a finite game H as follows. First, consider the set of all leaves of the game tree T of G that are the ends of the plays corresponding with a subgame perfect equilibrium. Next, for each outcome associated with a subgame perfect equilibrium retain in this set just one leaf with this outcome. By assumption the resulting set L is finite.

Next, order the leaves arbitrarily. Following this ordering remove all leaves with an outcome already associated with an earlier leaf, but ensuring that the leaves from L are retained. Let M be the resulting set of leaves. Finally, remove all nodes of T from which no leaf in M can be reached.

The resulting tree corresponds to a finite extensive game H in which all the outcomes possible in G are present. Further, all the leaves of H are also leaves of G, so H satisfies the TDI condition since G does. So by Theorem 12 of [2] (that is implicit in [10]) all subgame perfect equilibria of H are payoff equivalent.

Further, by Theorem 5 G has a subgame perfect equilibrium. Consider two subgame perfect equilibria s and t in G with the outcomes $p(s)$ and $p(t)$. By construction two subgame perfect equilibria s' and t' in H exist such that $p(s) = p(s')$ and $p(t) = p(t')$. We conclude that all subgame perfect equilibria of G are payoff equivalent.

To complete the proof it suffice to note that if an extensive game G satisfies the TDI condition, then so does every subgame of it. Indeed, consider a subgame G^w of G. Let $i = turn(w)$ and take $r_i^w, t_i^w \in S_i^w$ and $s_{-i}^w \in S_{-i}^w$. Extend these strategies to the strategies r_i, t_i and s_{-i} in the game G in such a way that w lies both on $play(r_i, s_{-i})$ and on $play(t_i, s_{-i})$. Then $p(r_i^w, s_{-i}^w) = p(r_i, s_{-i})$ and $p(t_i^w, s_{-i}^w) = p(t_i, s_{-i})$, so the claim follows. □

Corollary 14 *The claim of Theorem 11 holds for extensive games with finitely many outcomes that satisfy the TDI condition.*

Conjecture Every extensive game that satisfies the TDI-condition is SPE-invariant.

If the conjecture is true, Theorem 11 holds for all extensive games that satisfy the TDI condition. An example of a game with infinitely many outcomes that satisfies the TDI condition is the infinite version of the centipede game from Example 12.

5 Strictly competitive extensive games

In some games, for instance, the infinite version of the centipede game from Example 12, infinite rounds of elimination of weakly dominated strategies are needed to solve the game. In this section, we focus on maximal elimination of weakly dominated strategies and identify a subclass of extensive games for which we can provide a finite bound on the number of elimination steps required to solve the game. The outcome is our second main result which is a generalization of the following result due to [6] to a class of well-founded games.

[4] Alternative proofs of this result were given in [1] and [17].

Theorem Every finite extensive zero-sum game with n outcomes can be reduced to a trivial game by the maximal iterated elimination of weakly dominated strategies in $n-1$ steps.

We first present some auxiliary results. Their proofs follow our detailed exposition in [2] of the proofs in [6] generalized to strictly competitive games, now appropriately modified to infinite games.

5.1 Preliminary results

We denote by H^1 the subgame of H obtained by the elimination of all strategies that are weakly dominated in H, and put $H^0 := H$ and $H^{k+1} := (H^k)^1$, where $k \geq 1$. Abbreviate the phrase 'iterated elimination of weakly dominated strategies' to IEWDS. If for some k, H^k is a trivial game we say that H ***can be solved by the IEWDS***.

In infinite strategic games with finitely many outcomes it is possible that all strategies of a player are weakly dominated as shown in the Example 15. Then by definition, H^1 is an empty game. We define a class of games, called WD-admissible games in which this does not happen.

Example 15 Consider the following infinite zero-sum strategic game with two outcomes:

	A	B	C	D	$\dots$
A	$0,0$	$0,0$	$0,0$	$0,0$	$\dots$
B	$0,0$	$1,-1$	$0,0$	$0,0$	$\dots$
C	$0,0$	$1,-1$	$1,-1$	$0,0$	$\dots$
D	$0,0$	$1,-1$	$1,-1$	$1,-1$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

This game has a Nash equilibrium, namely (A,A), but each strategy of the row player is weakly dominated. So after one round of elimination the empty game is reached. □

Consider a strategic game H. We say that a strategy is ***undominated*** if no strategy weakly dominates it. Next, we say that H is ***WD-admissible*** if for all subgames H' of it the following holds: *each strategy is undominated or is weakly dominated by an undominated strategy*. Intuitively, a strategic game H is WD-admissible if in every subgame H' of it, for every strategy s_i in H' the relation 'is weakly dominated' in H' has a maximal element above s_i. The crucial property of WD-admissible games is formalised in the following lemma whose proof follows directly by induction.

Lemma 16 *Let $H := (H_1,\dots,H_n,p_1,\dots,p_n)$ be a WD-admissible strategic game and for $k \geq 1$, let $H^k := (H^k_1,\dots,H^k_n,p_1,\dots,p_n)$. Then $\forall i \in \{1,\dots,n\}\ \forall s_i \in H_i\ \exists t_i \in H^k_i\ \forall s_{-i} \in H^k_{-i} : p_i(t_i,s_{-i}) \geq p_i(s_i,s_{-i})$.*

A two player strategic game $H = (H_1,H_2,p_1,p_2)$ is called ***strictly competitive*** if $\forall i \in \{1,2\}\ \forall s,s' \in S$: $p_i(s) \geq p_i(s')$ iff $p_{-i}(s) \leq p_{-i}(s')$. For $i \in \{1,2\}$ we define $maxmin_i(H) := \max_{s_i \in H_i} \min_{s_{-i} \in H_{-i}} p_i(s_i,s_{-i})$. We allow $-\infty$ and ∞ as minima and maxima, so $maxmin_i(H)$ always exists. When $maxmin_i(H)$ is finite we call any strategy s^*_i such that $\min_{s_{-i} \in H_{-i}} p_i(s^*_i,s_{-i}) = maxmin_i(H)$ a ***security strategy*** for player i in H.

We shall reuse the following auxiliary results from [2].

Note 17 *Let $H = (H_1,H_2,p_1,p_2)$ be a strictly competitive strategic game. Then*

$$\forall i \in \{1,2\}\ \forall s,s' \in S : p_i(s) = p_i(s') \textit{ iff } p_{-i}(s) = p_{-i}(s').$$

This simply means that every strictly competitive strategic game satisfies the TDI condition.

Lemma 18 *Consider a strictly competitive strategic game H with a Nash equilibrium s. Suppose that for some $i \in \{1,2\}$, t_i weakly dominates s_i. Then (t_i,s_{-i}) is also a Nash equilibrium.*

Lemma 19 *Consider a strictly competitive strategic game H with two outcomes that has a Nash equilibrium. Then H^1 is a trivial game.*

The following result is standard (for the used formulation see, e.g., [14, Theorem 5.11, page 235]).

Theorem 20 *Consider a strictly competitive strategic game H.*

(i) All Nash equilibria of H yield the same payoff for player i, namely $maxmin_i(H)$.

(ii) All Nash equilibria of H are of the form (s_1^*, s_2^*) where each s_i^* is a security strategy for player i.

By modifying the proof of Corollary 5 from [2] appropriately, we have the following.

Lemma 21 *Consider a WD-admissible strictly competitive strategic game H that has a Nash equilibrium. Then H^1 has a Nash equilibrium, as well, and for all $i \in \{1,2\}$, $maxmin_i(H) = maxmin_i(H^1)$.*

5.2 A bound on IEWDS

We now move on to a discussion of extensive games. We say that an extensive game G is ***WD-admissible*** (respectively, ***strictly competitive***) if $\Gamma(G)$ is WD-admissible (respectively, strictly competitive). We write $\Gamma^k(G)$ instead of $(\Gamma(G))^k$, $\Gamma_i(G)$ instead of $(\Gamma(G))_i$, and $\Gamma_i^k(G)$ instead of $(\Gamma^k(G))_i$. So $\Gamma^0(G) = \Gamma(G)$. Further, for a strictly competitive game $H = (H_1, H_2, p_1, p_2)$ with finitely many outcomes for each player i we define the following three sets: $p_i^{max}(H) := \max_{s \in S} p_i(s)$, $win_i(H) := \{s_i \in H_i \mid \forall s_{-i} \in H_{-i}\, p_i(s_i, s_{-i}) = p_i^{max}(H)\}$ and $lose_{-i}(H) = \{s_{-i} \in H_{-i} \mid \exists s_i \in H_i\, p_i(s_i, s_{-i}) = p_i^{max}(H)\}$. By the assumption about H, $p_i^{max}(H)$ is finite.

We can then prove the following generalization of the crucial Lemma 1 and Theorem 1 from [6], where the proofs are analogous to that of Lemma 18 and Theorem 19 in [2].

Lemma 22 *Let G be a WD-admissible strictly competitive extensive game with finitely many outcomes. For all $i \in \{1,2\}$ and for all $k \geq 0$, if $win_i(\Gamma^k(G)) = \emptyset$ then $lose_{-i}(\Gamma^k(G)) \cap \Gamma_{-i}^{k+2}(G) = \emptyset$.*

Lemma 22 implies that if for all $i \in \{1,2\}$, $win_i(\Gamma^k(G)) = \emptyset$ then two further rounds of eliminations of weakly dominated strategies remove from $\Gamma^k(G)$ at least two outcomes.

This allows us to establish the following result. The proof is almost the same as the one given in [2, Theorem 19] for the finite extensive games. We reproduce it here for the convenience of the reader.

Theorem 23 *Let G be a WD-admissible strictly competitive extensive game with at most m outcomes. Then $\Gamma^{m-1}(G)$ is a trivial game.*

Proof. We prove a stronger claim, namely that for all $m \geq 1$ and $k \geq 0$ if $\Gamma^k(G)$ has at most m outcomes, then $\Gamma^{k+m-1}(G)$ is a trivial game.

We proceed by induction on m. For $m = 1$ the claim is trivial. For $m = 2$ we first note that by Theorem 5 and Lemma 21 each game $\Gamma^k(G)$ has a Nash equilibrium. So the claim follows by Lemma 19. For $m > 2$ two cases arise.

Case 1. For some $i \in \{1,2\}$, $win_i(\Gamma^k(G)) \neq \emptyset$.

For player i every strategy $s_i \in win_i(\Gamma^k(G))$ weakly dominates all strategies $s_i' \notin win_i(\Gamma^k(G))$ and no strategy in $win_i(\Gamma^k(G))$ is weakly dominated. So the set of strategies of player i in $\Gamma^{k+1}(G)$ equals $win_i(\Gamma^k(G))$ and consequently $p_i^{max}(\Gamma^k(G))$ is his unique payoff in this game. By Note 17 $\Gamma^{k+1}(G)$, and hence also $\Gamma^{k+m-1}(G)$, is a trivial game.

Case 2. For all $i \in \{1,2\}$, $win_i(\Gamma^k(G)) = \emptyset$.

Take joint strategies s and t such that $p_1(s) = p_1^{\max}(\Gamma^k(G))$ and $p_2(t) = p_2^{\max}(\Gamma^k(G))$. By Note 17 the outcomes $(p_1(s), p_2(s))$ and $(p_1(t), p_2(t))$ are different since $m > 1$.

We have $s_2 \in lose_2(\Gamma^k(G))$ and $t_1 \in lose_1(\Gamma^k(G))$. Hence by Lemma 22 for no joint strategy s' in $\Gamma^{k+2}(G)$ we have $p_1(s') = p_1^{\max}(\Gamma^k(G))$ or $p_2(s') = p_2^{\max}(\Gamma^k(G))$.

So $\Gamma(G^{k+2})$ has at most $m-2$ outcomes. By the induction hypothesis $\Gamma(G^{k+m-1})$ is a trivial game. □

We now show that Theorem 23 holds for a large class of natural games. Call an extensive game ***almost constant*** if for all but finitely many leaves the outcome is the same. Note that every almost constant game has finitely many outcomes, but the converse does not hold. Indeed, it suffices to take a game with two outcomes, each associated with infinitely many leaves. The following general result holds.

Theorem 24 *Every almost constant extensive game is WD-admissible.*

Proof. We begin with two unrelated observations. Call a function $p : A \to B$ ***almost constant*** if for some b we have $p(a) = b$ for all but finitely many $a \in A$.

Observation 1. Consider two sequences of some elements $(v_0, v_1, \ldots)$ and $(w_0, w_1, \ldots)$ such that $v_j \neq v_k$, $v_j \neq w_k$, and $w_j \neq w_k$ for all $j \geq 0$ and $k > j$, and a function $p : \{v_0, v_1, \ldots\} \cup \{w_0, w_1, \ldots\} \to B$ such that $p(v_j) \neq p(w_j)$ for all $j \geq 0$. Then p is not almost constant.
Indeed, otherwise for some $k \geq 0$ the function $p : \{v_k, v_{k+1}, \ldots\} \cup \{w_k, w_{k+1}, \ldots\} \to B$ would be constant.

Observation 2. Take an extensive game. For some player i, consider two joint strategies (s_i, s_{-i}) and (s'_i, s'_{-i}). If $leaf(s_i, s_{-i}) = leaf(s'_i, s'_{-i})$ then $leaf(s_i, s_{-i}) = leaf(s'_i, s_{-i})$.

Indeed, consider any node w in $play(s_i, s_{-i})$ such that $turn(w) = i$. Then by assumption $s_i(w) = s'_i(w)$. This implies that $play(s_i, s_{-i}) = play(s'_i, s_{-i})$, which yields the claim.

Now consider an almost constant extensive game G. Take an arbitrary subgame H of $\Gamma(G)$. Suppose by contradiction that for some player i there exists an infinite sequence of strategies $s_i^0, s_i^1, s_i^2, \ldots$ such that for all $j \geq 0$, s_i^{j+1} weakly dominates s_i^j in H. By definition of weak dominance, for all $j \geq 0$ there exists $s_{-i}^j \in H_{-i}$ such that $p_i(s_i^j, s_{-i}^j) < p_i(s_i^{j+1}, s_{-i}^j)$. Let for $j \geq 0$, $v_j = leaf(s_i^j, s_{-i}^j)$ and $w_j = leaf(s_i^{j+1}, s_{-i}^j)$. By the above inequalities $p_i(v_j) \neq p_i(w_j)$ for all $j \geq 0$.

We now argue that $v_j \neq v_k$, $v_j \neq w_k$, and $w_j \neq w_k$ for all $j \geq 0$ and $k > j$. First, note that by the transitivity of the 'weakly dominates' relation we have the following.

- $p_i(s_i^j, s_{-i}^j) < p_i(s_i^{j+1}, s_{-i}^j) \leq p_i(s_i^k, s_{-i}^j)$,
- $p_i(s_i^j, s_{-i}^j) < p_i(s_i^{j+1}, s_{-i}^j) \leq p_i(s_i^{k+1}, s_{-i}^j)$,
- $p_i(s_i^{j+1}, s_{-i}^k) \leq p_i(s_i^k, s_{-i}^k) < p_i(s_i^{k+1}, s_{-i}^k)$.

This implies in turn, $leaf(s_i^j, s_{-i}^j) \neq leaf(s_i^k, s_{-i}^j)$, $leaf(s_i^j, s_{-i}^j) \neq leaf(s_i^{k+1}, s_{-i}^j)$, and $leaf(s_i^{j+1}, s_{-i}^k) \neq leaf(s_i^{k+1}, s_{-i}^k)$. So by Observation 2 we have the following.

- $v_j = leaf(s_i^j, s_{-i}^j) \neq leaf(s_i^k, s_{-i}^k) = v_k$,
- $v_j = leaf(s_i^j, s_{-i}^j) \neq leaf(s_i^{k+1}, s_{-i}^k) = w_k$,
- $w_j = leaf(s_i^{j+1}, s_{-i}^j) \neq leaf(s_i^{k+1}, s_{-i}^k) = w_k$.

By Observation 1, p_i is not almost constant, which contradicts the assumption that G is almost constant. By the transitivity of the 'weakly dominates' relation we conclude that G is WD-admissible. □

Corollary 25 *Let G be an almost constant strictly competitive extensive game with at most m outcomes. Then $\Gamma^{m-1}(G)$ is a trivial game.*

Acknowledgments

We thank the reviewers for their helpful comments. The second author was partially supported by the grant CRG/2022/006140.

References

[1] K.R. Apt (2004): *Uniform Proofs of Order Independence for Various Strategy Elimination Procedures*. The B.E. Journal of Theoretical Economics 4(1), doi:10.2202/1534-5971.1141. Available at `https://arxiv.org/abs/cs/0403024`. Article 5, 48 pages.

[2] K.R. Apt & S. Simon (2021): *A tutorial for computer scientists on finite extensive games with perfect information*. Bulletin of the EATCS 135. Available at `https://arxiv.org/abs/2204.08740`. 40 pages.

[3] K.R. Apt & S. Simon (2021): *Well-founded extensive games with perfect information*. In: Proceedings 18th Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2021, 335, Electronic Proceedings in Theoretical Computer Science (EPTCS), pp. 7–21, doi:10.4204/EPTCS.335.2.

[4] P. Battigalli (1997): *On rationalizability in extensive games*. Journal of Economic Theory 74, pp. 40–61, doi:10.1006/jeth.1996.2252.

[5] B.D. Bernheim (1984): *Rationalizable Strategic Behavior*. Econometrica 52(4), pp. 1007–1028, doi:10.2307/1911196.

[6] C. Ewerhart (2002): *Iterated Weak Dominance in Strictly Competitive Games of Perfect Information*. Journal of Economic Theory 107(2), pp. 474–482, doi:10.1006/jeth.2001.2958.

[7] A. Heifetz (2012): *Game Theory: Interactive Strategies in Economics and Management*. Cambridge University Press, doi:10.1017/CBO9781139049344.

[8] P. Hummel (2008): *Iterative elimination of weakly dominated strategies in binary voting agendas with sequential voting*. Social Choice and Welfare 31(2), pp. 257–269, doi:10.1007/s00355-007-0278-4.

[9] B. L. Lipman (1994): *A Note on the Implications of Common Knowledge of Rationality*. Games and Economic Behavior 6, pp. 114–129, doi:10.1006/game.1994.1006.

[10] L.M. Marx & J.M. Swinkels (1997): *Order Independence for Iterated Weak Dominance*. Games and Economic Behaviour 18, pp. 219–245, doi:10.1006/game.1997.0525.

[11] M.J. Osborne & A. Rubinstein (1994): *A Course in Game Theory*. The MIT Press.

[12] D.G. Pearce (1984): *Rationalizable Strategic Behavior and the Problem of Perfection*. Econometrica 52(4), pp. 1029–1050, doi:10.2307/1911197.

[13] A. Perea (2014): *Belief in the opponents' future rationality*. Games and Economic Behaviour 83, pp. 231–254, doi:10.1016/j.geb.2013.11.008.

[14] K. Ritzberger (2001): *Foundations of Non-cooperative Game Theory*. Oxford University Press, Oxford, UK.

[15] R. Rosenthal (1981): *Games of perfect information, predatory pricing and the chain-store paradox*. Journal of Economic Theory 25(1), pp. 92–100, doi:10.1016/0022-0531(81)90018-1.

[16] J. Sobel (2019): *Iterated weak dominance and interval-dominance supermodular games*. Theoretical Economics 14(1), pp. 71–102, doi:10.3982/TE2904.

[17] L.P. Østerdal (2005): *Iterated weak dominance and subgame dominance*. Journal of Mathematical Economics 41(6), pp. 637–645, doi:10.1016/j.jmateco.2003.11.013.

Resilient Information Aggregation

Itai Arieli
Technion
Haifa, Israel
iarieli@technion.ac.il

Ivan Geffner
Technion
Haifa, Israel
ieg8@cornell.edu

Moshe Tennenholtz [*]
Technion
Haifa, Israel
moshet@technion.ac.il

In an information aggregation game, a set of senders interact with a receiver through a mediator. Each sender observes the state of the world and communicates a message to the mediator, who recommends an action to the receiver based on the messages received. The payoff of the senders and of the receiver depend on both the state of the world and the action selected by the receiver. This setting extends the celebrated cheap talk model in two aspects: there are many senders (as opposed to just one) and there is a mediator. From a practical perspective, this setting captures platforms in which strategic experts advice is aggregated in service of action recommendations to the user. We aim at finding an optimal mediator/platform that maximizes the users' welfare given highly resilient incentive compatibility requirements on the equilibrium selected: we want the platform to be incentive compatible for the receiver/user when selecting the recommended action, and we want it to be resilient against group deviations by the senders/experts. We provide highly positive answers to this challenge, manifested through efficient algorithms.

1 Introduction

Experts' opinions aggregation platforms are crucial for web monetizing. Indeed, in sites such as Reddit or Google, comments and reviews are aggregated as an answer to a user query about items observed or studied by others. We refer to these reviewers as *experts*. The platform can aggregate these experts' inputs or filter them when providing a recommendation to the user, which will later lead to a user action. An ideal platform should maximize the users' social welfare. In an economic setting, however, the different experts may have their own preferences. Needless to say, when commenting on a product or a service, we might not know if the expert prefers the user to buy the product or accept the service, or if the expert prefers otherwise. This is true even when all experts observe exactly the same characteristics of a product or service.

Interestingly, while the study of economic platforms is rich [21, 12, 24, 4, 20, 25, 7, 15, 23], there is no rigorous foundational and algorithmic setting for the study of aggregation and filtering of strategic experts opinions in service of the platform users. In this paper, we initiate such a study, which we believe to be essential. This study can be viewed as complementary to work on platform incentives [21], issues of dishonesty [12], and issues of ranking/filtering [7], by putting these ingredients in a concrete foundational economic setting dealing with recommendations based on inputs from strategic experts. The model we offer extends the classical cheap talk model in two fundamental directions. First, by having several strategic senders (experts) rather than only one; second, by introducing a platform that acts as a mediator in an information design setting.

Our work is related to the literature on information design that studies optimal information disclosure policies for informed players. The two leading models of information design are cheap talk [6] and

[*] The work by Ivan Geffner and Moshe Tennenholtz was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 740435).

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 31–45, doi:10.4204/EPTCS.379.6

Bayesian persuasion [10]. The main distinction between these models is the underlying assumption that, in the Bayesian persuasion model, the sender has commitment power in the way she discloses the information, while in the cheap talk model she has not.

Bayesian persuasion models emphasize commitment power, and while it may hold in some real-world situations, it is often considered strong. In addition, in Bayesian persuasion, the informed agent (the sender) is also the one who designs the information revelation policy. In practice, however, information revelation can be determined by other external or legal constraints. A commerce platform, for example, determines what information about a product is revealed to a potential customer based on information submitted by different suppliers.

In our model there is a finite state space of size n, several informed players (senders), an uninformed player (the receiver) that determines the outcome of the game by playing a binary action from the set $A := \{0,1\}$ (this could represent buying a product or not, passing a law or not, etc.), and a mediator that acts as a communication device between the senders and the receiver (the mediator can be seen as the platform used by all parties). The utility of each player is determined by the state and by the action played by the receiver. The incentives of the senders may not necessarily be aligned (e.g., senders can be a car seller and a technician that tested the car, two independent parties who studied the monetary value of law, two suppliers of a product, etc.). The state is drawn from a prior distribution that is commonly known among players, but only the senders know its realized value. Thus, the senders' purpose is to reveal information to the receiver in such a way that the receiver plays the action that benefits them the most. Since the senders have no commitment power, we are interested in a mediated cheap talk equilibrium, in which it is never in the interest of the senders to be dishonest, and it is always in the interest of the receiver to play the action suggested by the protocol.

The most common notions of equilibrium, such as Nash equilibrium, require that each individual player cannot increase its utility by deviating from the proposed strategy. However, notions of equilibria that are resilient to group deviations are currently gaining traction [3, 9, 2], in particular because of their Web applications. Indeed, on the Internet, it is not only fairly easy to collude, but it is also relatively simple to create proxy pseudo-identities and defect in a coordinated way (this is known as a Sybil attack [8]). Nowadays, in Web applications and in distributed systems, resilience against individual deviations is generally considered insufficient for practical purposes. For instance, blockchain protocols are required to tolerate coordinated deviations from up to a fraction of their user base. In this work, we focus on k-resilient equilibria, which are strategies profiles in which no coalition of up to k players can increase their utility by deviating.

Our main goal in the paper is to characterize, given the incentives of the senders and the receiver, which maps from states to distributions over actions result from playing k-resilient equilibria. More precisely, each cheap talk protocol $\vec{\sigma}$ induces a map M from states to distributions over actions, where $M(\omega)$ is mapped to the distribution over actions resulting from playing $\vec{\sigma}$ in state ω. Our aim is to characterize which of these maps (or *outcomes*, as we call them) can be implemented by a k-resilient equilibrium, and to efficiently construct a concrete k-resilient equilibrium whenever a given outcome is implementable. We first show that, if there are more than two senders, even if one of them defects and misreports the state, a majority of the senders would still report the truth, and thus the mediator will always be able to compute the correct state. Therefore, if there are at least three senders, outcomes are implementable by a 1-resilient equilibrium (i.e., a Nash equilibrium) if and only if they are incentive-compatible for the receiver. That is, an outcome is implementable by a 1-resilient equilibrium if and only if it improves the utility of the receiver relative to the case where no information is revealed to her. This result implies that the set of implementable distributions is independent of the utilities of the senders and only depends on that of the receiver, and thus that the senders have no bargaining power. It is also easy to check that this result extends to the case of k-resilient equilibria for $k < n/2$, where n is the number of senders. However, we show that if a majority of the players can collude, the set of implementable outcomes is defined by a system of linear equations that depend both on the utilities of the senders and the receiver. It may seem at first that computing such characterization may be highly inefficient since the number of possible coalitions of size at most $k \geq n/2$ grows exponentially over the number of players, and each of these possible coalitions imposes a constraint on the outcome. By contrast, our main result shows that, if the number of states is m, then the aforementioned linear system can be written with only m^2 inequality constraints, and all such inequalities can be computed in polynomial time over m and the number of senders n. This means that the best receiver k-resilient equilibrium or the k-resilient equilibrium that maximizes social welfare can be computed efficiently. We also provide, given a solution of the system of equations, an efficient way to construct a concrete k-resilient equilibrium that implements the desired outcome.

Our results so far assume that all senders have full information about the realized state. However, in some cases it is realistic to assume that senders only have partial information about it and, moreover, that each sender's information might be different. We show in Section 6 that our techniques generalize to this model as long as the senders's preferences are not influenced by their coalition, a condition that we call *k-separability*. This means that, assuming k-separability, we provide a characterization of all outcomes that are implementable by a k-resilient equilibrium, and an algorithm that construct a concrete k-resilient equilibrium that implements a desired (implementable) outcome. Both the characterization and the algorithm are efficient relative to the size of the game's description.

1.1 Related Work

The literature on information design is too vast to address all the related work. We will therefore mention some key related papers. Krishna and Morgan [14] consider a setting similar to that considered by Crawford and Sobel [6], where a real interval represents the set of states and actions. In this setting, the receiver's and the senders' utilities are *biased* by some factor that affects their incentives and utility. Similarly to the current paper where the sender is not unique, Krishna and Morgan consider two informed senders that reveal information sequentially to the receiver. They consider the best receiver equilibrium and show that, when both senders are *biased* in the same direction, it is never beneficial to consult both of them. By contrast, when senders are biased in opposite directions, it is always beneficial to consult them both.

In another work, Salamanca [22] characterizes the optimal mediation for the sender in a sender-receiver game. Lipnowski and Ravid [16], and Kamenica and Gentzkow [10] provide a geometric characterization of the best cheap talk equilibrium for the sender under the assumption that the sender's utility is state-independent. The geometric characterization of Lipnowski and Ravid is no longer valid for the case where there are two or more senders.

Kamenika and Gentzkow [11] consider a setting with two senders in a Bayesian persuasion model. The two senders, as in the standard Bayesian persuasion model (and unlike ours), have commitment power and they compete over information revelation. The authors characterize the equilibrium outcomes in this setting.

In many game-theoretical works, mediators are incorporated into strategic settings [5, 18]. Kosenko [13] also studied the information aggregation problem. However, their model assumed that the mediator had incentives of its own and selected its policy at the same time as the sender. Monderer and Tennenholtz [17] studied the use of mediators to enhance the set of situations where coalition deviance is stable. They show that using mediators in several classes of settings can produce stable behaviors that are resistant to coalition deviations. In our setting, the existence of a k-resilient equilibrium is straightforward (e.g., playing a constant action). Instead, the strength of our result follows from efficiently characterising the set of all outcomes that are implementable using k-resilient mediated equilibria.

2 Model

In an information aggregation game $\Gamma = (S, A, \Omega, p, u)$, there is a finite set of possible states $\Omega = \{\omega^1, \dots, \omega^m\}$, a commonly known distribution p over Ω, a set of possible actions $A = \{0, 1\}$, a set $S = \{1, 2, \dots, n\}$ of senders, a receiver r, a mediator d, and a utility function $u : (S \cup \{r\}) \times \Omega \times A \longrightarrow \mathbb{R}$ such that $u(i, \omega, a)$ gives the utility of player i when action a is played at state ω. Each information aggregation game instance is divided into four phases. In the first phase, a state ω is sampled from Ω following distribution p and this state is disclosed to all senders $i \in S$. During the second phase, each sender i sends a message m_i to the mediator. In the third phase (after receiving a message from each sender) the mediator must send a message $m_d \in A$ to the receiver, and in the last phase the receiver must play an action $a \in A$ and each player $i \in S \cup \{r\}$ receives $u(i, \omega, a)$ utility.

The behavior of each player i and is determined by its strategy σ_i and the behavior of the mediator is determined by its strategy σ_d. A strategy σ_i for a player $i \in S$ can be represented by a (possibly randomized) function $m_i : \Omega \longrightarrow \{0, 1\}^*$ such that $m_i(\omega)$ indicates what message i is sending to the mediator given state $\omega \in \Omega$. The strategy σ_d of the mediator can be represented by a function $m_d : (\{0, 1\}^*)^n \longrightarrow A$ that indicates, given the message received from each player, what message it should send to the receiver. The strategy σ_r of the receiver can be represented by a function $a_r : A \rightarrow A$ that indicates which action it should play given the message received from the mediator.

In summary, a game instance goes as follows:

1. A state ω is sampled from Ω following distribution p, and ω is disclosed to all senders $i \in S$.
2. Each sender $i \in S$ sends message $m_i(\omega)$ to the mediator.
3. The mediator sends message $m_d(m_1, \dots, m_n)$ to the receiver.
4. The receiver plays action $a_r(m_d)$ and each player $i \in S \cup \{r\}$ receives $u(i, \omega, a_r(m_d))$ utility.

Note that, in order to simplify the notation, we use a slight notation overload since m_i is both the message sent by player i and a function that depends on the state. This is because the message sent by i always depend on the state, even if it is not explicitly written. A similar situation happens with a_r.

2.1 Game mechanisms

Given a game $\Gamma = (S, A, \Omega, p, u)$, a *mechanism* $M = (m_1, m_2, \dots, m_n, m_d, a_r)$ uniquely determines a map $o^\Gamma_M : \Omega \longrightarrow \Delta A$ (where ΔA is the set of probability distributions of A) that maps each state ω to the distribution of actions obtained by playing Γ when the senders, the mediator and the receiver play the strategies represented by the components of M. We say that M *implements* o^Γ_M and that o^Γ_M is the *outcome* of M.

A mechanism M is *incentive-compatible* if it is not in the interest of the receiver or any of the senders to deviate from the proposed mechanism (note that the mediator has no incentives). We also say that M is *honest* if (a) $m_i \equiv Id_\Omega$, where $Id_\Omega(\omega) = \omega$ for all $\omega \in \Omega$, and (b) $a_r \equiv Id_A$. Moreover, we say that M is *truthful* if it is both honest and incentive-compatible. Intuitively, a mechanism is truthful if sending the true state to the mediator is a dominant strategy for the senders and playing the state suggested by the mediator is a dominant strategy for the receiver.

Example 1. *Consider a game $\Gamma = (S, A, \Omega, p, u)$ where $S = \{1,2,3\}$, $A = \{0,1\}$, $\Omega = \{\omega_1, \dots, \omega_k\}$, p is the uniform distribution over Ω and $u : (S \cup \{r\}) \times \Omega \times A \longrightarrow \mathbb{R}$ is an arbitrary utility function. Consider the truthful mechanism in which senders disclose the true state to the mediator, the mediator chooses the state $\omega \in \Omega$ sent by the majority of the senders and sends to the receiver the action a that maximizes $u(r, \omega, a)$, and the receiver plays the action sent by the mediator. It is easy to check that this mechanism is incentive-compatible: no individual sender can influence the outcome by deviating since the mediator chooses the state sent by the majority of the senders. Moreover, by construction, this mechanism gives the receiver the maximum possible utility among all mechanisms.*

Our first goal is to characterize the set of possible outcomes that can be implemented by truthful mechanisms. Note that, because of Myerson's revelation principle [19], characterizing the set of outcomes implemented by truthful mechanisms is the same as characterizing the set of outcomes implemented by any incentive-compatible mechanisms (not necessarily truthful):

Proposition 1. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game. Then, for any incentive-compatible mechanism M for Γ there exists a truthful mechanism M' such that M' implements o^{Γ}_{M}.*

Proof. Given $M = (m_1, m_2, \dots, m_n, m_d, a)$, consider a mechanism $M' = (m'_1, m'_2, \dots, m'_n, m'_d, a')$ such that $m'_i \equiv Id_{\Omega}$ for all $i \in S$, $m'_a \equiv Id_A$, and the mediator does the following. After receiving a message ω_j from each sender j, it computes $a(m_d(m_1(\omega_1),, m_2(\omega_2), \dots, m_n(\omega_n)))$ and sends this action to the receiver (if the message from some player j is inconsistent, the mediator takes ω_j to be an arbitrary element of Ω). By construction, M' is a truthful mechanism in which the mediator simulates everything the players would have sent or played with M. It is easy to check that, with M', for any possible deviation for player $j \in S \cup \{r\}$, there exists a deviation for j in M that produces the same outcome. Thus, if M is incentive-compatible, so is M'. □

This proposition shows that we can restrict our search to truthful mechanisms. Moreover, the construction used in the proof shows that we can assume without loss of generality that the senders can only send messages in Ω since sending any other message is equivalent to sending an arbitrary element of Ω. To simplify future constructions, we'll use this assumption from now on.

2.2 Resilient equilibria

Traditionally, in the game theory and mechanism design literature, the focus has always been on devising strategies or mechanisms such that no individual agent is incentivized to deviate. However, in the context of multi-agent Bayesian persuasion, this approach is not very interesting. The reason is that, if $n > 2$, the mediator can always compute the true state by taking the one sent by a majority of the senders (as seen in Example 1), and thus the mediator can make a suggestion to the receiver as a function of the true state while individual senders cannot influence the outcome by deviating. In fact, given action $a \in A$, let $U_a := \mathbb{E}_{\omega \leftarrow p}[u(r, \omega, a)]$ be the expected utility of the receiver when playing action a regardless of the mediator's suggestion and, given outcome o^{Γ}, let

$$E_i(o^{\Gamma}) := \mathbb{E}_{\substack{\omega \leftarrow p, \\ a \leftarrow o^{\Gamma}(\omega)}} [u(i, \omega, a)]$$

be the expected utility of player $i \in S \cup \{r\}$ with outcome o^{Γ}. The following proposition characterizes outcomes implementable by truthful mechanisms.

Proposition 2. *If $\Gamma = (S, A, \Omega, p, u)$ is an information aggregation game with $|S| > 2$, an outcome $o^{\Gamma} : \Omega \longrightarrow \Delta A$ of Γ is implementable by a truthful mechanism if and only if $E_r\left(o^{\Gamma}\right) \geq U_a$ for all $a \in A$.*

Intuitively, proposition 2 states that, if there are at least three senders, the only condition for an outcome to be implementable by a truthful incentive-compatible mechanism is that the receiver gets a better expected utility than the one it gets with no information. Before proving it, we need the following lemma, which will also be useful for later results.

Lemma 1. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game. An honest mechanism $M = (Id_\Omega, \dots, Id_\Omega, m_d, Id_A)$ for Γ is incentive-compatible for the receiver if and only if $E_r\left(o_M^\Gamma\right) \geq U_a$ for all $a \in A$.*

Proof. ($\Longrightarrow$) Let M be an honest mechanism for Γ that is incentive-compatible for the receiver. Then, if $E_r\left(o_M^\Gamma\right) < U_a$ for some $a \in A$, the receiver can increase its utility ignoring the mediator's suggestion and playing always action a. This would contradict the fact that M is incentive-compatible.

($\Longleftarrow$) Suppose that $E_r\left(o_M^\Gamma\right) \geq U_a$ for all $a \in A$. If M is not incentive-compatible, it means that the receiver can strictly increase its payoff either (a) by playing 1 when the mediator sends 0 and/or (b) playing 0 when the mediator sends 1. Suppose that (a) is true, then the receiver can strictly increase its payoff by playing 1 in all scenarios, which would contradict the fact that its expected payoff with M is greater or equal than U_1. The argument for (b) is analogous. □

With this we can prove Proposition 2. The mechanism used in the proof is very similar to the one in Example 1.

Proof of Proposition 2. Let M be a truthful mechanism. Then, by Lemma 1, o_m^Γ satisfies that $E_r\left(o_M^\Gamma\right) \geq U_a$ for all $a \in A$.

Conversely, suppose that an outcome o^Γ satisfies that $E_r\left(o_M^\Gamma\right) \geq U_a$ for all $a \in A$. Consider a mechanism $M = (Id_\Omega, \dots, Id_\Omega, m_d, Id_A)$ such that the mediator takes the state ω sent by the majority of the senders and sends $o^\Gamma(\omega)$ to the receiver. By construction, M implements o^Γ. Moreover, as in Example 1, M is incentive-compatible for the senders since, if $n > 2$, they cannot influence the outcome by individual deviations. By Lemma 1 M is also incentive-compatible for the receiver. Thus, M is a truthful mechanism that implements o^Γ. □

The construction used in the proof shows how easily we can implement any desired outcome as long as it is better for the sender than playing a constant action. However, Proposition 2 is only valid under the assumption that senders cannot collude and deviate in a coordinated way (an assumption that many times is unrealistic, as pointed out in the introduction). If we remove this assumption, the *next best thing* is to devise mechanisms such that all coalitions up to a certain size do not get additional utility by deviating. We focus mainly on the following notions of equilibrium:

Definition 1 ([1])**.** *Let Γ be any type of game for n players with strategy space $A = A_1 \times \ldots \times A_n$ and functions $u_i : S \longrightarrow \mathbb{R}$ that give the expected utility of player i when players play a given strategy profile. Then,*

- *A strategy profile $\vec{\sigma} \in A$ is a k-*resilient Nash equilibrium *if, for all coalitions K up to k players and all strategy profiles $\vec{\tau}_K$ for players in K, $u_i(\vec{\sigma}) \geq u_i(\vec{\sigma}_{-K}, \vec{\tau}_K)$ for some $i \in K$.*
- *A strategy profile $\vec{\sigma} \in A$ is a* strong k-resilient Nash equilibrium *if, for all coalitions K up to k players and all strategy profiles $\vec{\tau}_K$ for players in K, $u_i(\vec{\sigma}) \geq u_i(\vec{\sigma}_{-K}, \vec{\tau}_K)$ for all $i \in K$.*

Intuitively, a strategy profile is k-resilient if no coalition of up to k players can deviate in such a way that all members of the coalition strictly increase their utility, and a strategy profile is strongly k-resilient if no member of any coalition of up to k players can strictly increase its utility by deviating, even at the expense of the utility of other members of the coalition. We can construct analogous definitions in the context of information aggregation:

Definition 2. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game. A mechanism $M = (m_1, \ldots, m_n, m_d, a_r)$ for Γ is k-resilient incentive-compatible (resp., strong k-resilient incentive-compatible) if*

(a) The receiver cannot increase its utility by deviating from the proposed protocol.

(b) Fixing m_d and a_r beforehand, the strategy profile of the senders determined by M is a k-resilient Nash equilibrium (resp., strong k-resilient Nash equilibrium).

A mechanism M is k-resilient truthful if it is honest and k-resilient incentive-compatible. Strong k-resilient truthfulness is defined analogously.

3 Main Results

For the main results of this paper we need the following notation. Given an outcome $o : \Omega \to \Delta A$, we define by $o^* : \Omega \to [0,1]$ the function that maps each state ω to the probability that $o(\omega) = 0$. Note that, since $|A| = 2$, o is uniquely determined by o^*. The following theorem gives a high level characterization of all k-resilient truthful mechanisms (resp., strong k-resilient truthful mechanisms).

Theorem 1. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game with $\Omega = \{\omega^1, \ldots, \omega^m\}$. Then, there exists a system E of $O(m^2)$ equations over variables $x_1, \ldots, x_m$, such that each equation of E is of the form $x_i \leq x_j$ for some $i, j \in [m]$, and such that an outcome o of Γ is implementable by a k-resilient truthful mechanism (resp., strong k-resilient truthful mechanism) if and only if*

(a) $x_1 = o^*(\omega^1), \ldots, x_m = o^*(\omega^m)$ is a solution of E.

(b) $E_r(o) \geq U_a$ for all $a \in A$.

Moreover, the equations of E can be computed in polynomial time over m and the number of senders n.

Note that condition (b) is identical to the one that appears in Lemma 1. In fact, condition (b) is the necessary and sufficient condition for a mechanism that implements o to be incentive-compatible for the receiver, and condition (a) is the necessary and sufficient condition for this mechanism to be k-resilient incentive-compatible (resp., strong k-resilient incentive-compatible) for the senders. Theorem 1 shows that the set of outcomes implementable by k-resilient truthful mechanisms (resp., strong k-resilient truthful mechanisms) is precisely the set of solutions of a system of equations over $\{o^*(\omega^i)\}_{i \in [m]}$. This means that the solution that maximizes any linear function over $\{o^*(\omega^i)\}_{i \in [m]}$ can be reduced to an instance of linear programming. In particular, the best implementable outcome for the receiver or for each of the senders can be computed efficiently.

Corollary 1. *There exists a polynomial time algorithm that computes the outcome that could be implemented by a k-resilient truthful mechanism (resp., strong k-resilient truthful mechanism) that gives the most utility to the receiver or that gives the most utility to a particular sender.*

Our last result states that not only we can characterize the outcomes implementable by truthful mechanisms, but that we can also efficiently compute a truthful mechanism that implements a particular outcome. Before stating this formally, it is important to note that all truthful mechanisms can be encoded by a single function m_d^* from message profiles $\vec{m} = (m_1, \ldots, m_n)$ to $[0,1]$. Intuitively, the mechanism m_d defined by m_d^* is the one that maps $(\vec{m})$ to the distribution such that 0 has probability $m_d^*(\vec{m})$ and 1 has probability $1 - m_d^*(\vec{m})$. Moreover, note that the description of a k-resilient truthful mechanism for a game with m possible states is exponential over k since the mechanism must describe what to do if k players misreport their state, which means that the mechanism should be defined over at least m^k inputs. Clearly, no polynomial algorithm over n and m can compute this mechanism just because of the sheer size of the output. However, given a game Γ and an output o, it is not necessary to compute the whole description of the resilient truthful mechanism m_d^* for Γ that implements o, we only need to be able to compute $m_d^*(\vec{m})$ in polynomial time for each possible message profile $\vec{m}$. We state this as follows.

Theorem 2. *There exists an algorithm π that receives as input the description of an information aggregation game $\Gamma = (S, A, \Omega, p, u)$, an outcome o for Γ implementable by a k-resilient mechanism (resp., strong k-resilient mechanism), and a message input $\vec{m}$ for the mediator, and π outputs a value $q \in [0,1]$ such that the function m_d^* defined by $m_d^*(\vec{m}) := A(\Gamma, o, \vec{m})$ determines a k-resilient truthful mechanism (resp., strong k-resilient truthful mechanism) for Γ that implements o. Moreover, π runs in polynomial time over $|\Omega|$ and $|S|$.*

The proofs of Theorems 1 and 2 are detailed in Sections 4 and 5 respectively. Intuitively, each coalition imposes a constraint over the space of possible messages that the mediator may receive, implying that the mediator should suggest action 0 more often for some message inputs than others. These constraints induce a partial order over *pure inputs* (i.e., messages such that all senders report the same state), which is precisely the order defined by E in Theorem 1. It can be shown that, even though there may be exponentially many possible coalitions of size at most k, this partial order can be computed in polynomial time over the number of states and senders.

4 Proof of Theorem 1

In this section we prove Theorem 1. Note that, because of Lemma 1, we only have to show that, given a game $\Gamma = (S, A, \Omega, p, u)$ with $|\Omega| = m$ and $|S| = n$, there exists a system of equations E as in Theorem 1 such that an outcome o is implementable by an honest mechanism that is k-resilient incentive-compatible (resp., strong k-resilient) for the senders if and only if $(o^*(\omega^1), \dots, o^*(\omega^m))$ is a solution of E.

To understand the key idea, let us start with an example in which $\Omega = \{\omega^1, \omega^2\}$, $S = \{1,2,3,4\}$, senders $1,2$ and 3 prefer action 0 in ω^2, senders $2,3$ and 4 prefer action 1 in ω^1, and in which we are trying to characterize all outcomes that could be implemented by a mechanism that is 2-resilient incentive-compatible for the senders. If all senders are honest, then the mediator could only receive inputs $(\omega^1, \omega^1, \omega^1, \omega^1)$ or $(\omega^2, \omega^2, \omega^2, \omega^2)$ (where the ith component of an input represents the message sent by sender i). However, since senders could in principle deviate, the mediator could receive, for instance, an input of the form $(\omega^1, \omega^1, \omega^2, \omega^2)$. This input could originate in two ways, either the true state is ω^1 and senders 3 and 4 are misreporting the state, or the state is ω^2 and senders 1 and 2 are misreporting. Even though a mechanism is honest, the mediator's message function m_d should still be defined for inputs with different components, and it must actually be done in such a way that players are not incentivized to misreport.

Let m_d^* be the function that maps each message (m_1, m_2, m_3, m_4) to the probability that $m_d(m_1, \dots, m_4) = 0$. If the honest mechanism determined by m_d^* is 2-resilient incentive-compatible for the senders, the probability of playing 0 should be lower with $(\omega^1, \omega^1, \omega^2, \omega^2)$ than with $(\omega^2, \omega^2, \omega^2, \omega^2)$. Otherwise, in ω^2, senders 1 and 2 can increase their utility by reporting 1 instead of 2. Thus, m_d^* must satisfy that $m_d^*(\omega^1, \omega^1, \omega^2, \omega^2) \leq m_d^*(\omega^2, \omega^2, \omega^2, \omega^2)$. Moreover, $m_d^*(\omega^1, \omega^1, \omega^2, \omega^2) \geq m_d^*(\omega^1, \omega^1, \omega^1, \omega^1)$, since otherwise, in state ω^1, senders 3 and 4 can increase their utility by reporting 2 instead of 1. These inequalities together imply that $m_d^*(\omega^1, \omega^1, \omega^1, \omega^1) \leq m_2^*(\omega^2, \omega^2, \omega^2, \omega^2)$, and therefore that $o^*(\omega^1) \leq o^*(\omega^2)$. In fact, we can show that this is the only requirement for o to be implementable by a mechanism that is k-resilient incentive compatible for the senders. Given o such that $o^*(\omega^1) \leq o^*(\omega^2)$, consider an honest mechanism determined by m_d^*, in which $m_d^*(m_1, m_2, m_3, m_4)$ is defined as follows:

- If at least three players sent the same message ω, then $m^*_d(m_1,m_2,m_3,m_4) := o^*(\omega)$.
- Otherwise, $m^*_d(m_1,m_2,m_3,m_4) := (o^*(\omega^1)+o^*(\omega^2))/2$.

We can check that the honest mechanism M determined by m^*_d is indeed 2-resilient incentive-compatible for the senders. Clearly, no individual sender would ever want to deviate since it cannot influence the outcome by itself (still three messages would disclose the true state). Moreover, no pair of senders can increase their utility by deviating since, in both ω^1 and ω^2, at least one of the senders in the coalition would get the maximum possible utility by disclosing the true state. This shows that, in this example, $o^*(\omega^1) \leq o^*(\omega^2)$ is the only necessary and sufficient condition for o to be implementable by a mechanism that is 2-resilient incentive-compatible for the senders.

4.1 Theorem 1, general case

The proof of the general case follows the same lines as the previous example. We show the generalization for the case of k-resilient incentive-compatibility, the proof for strong k-resilience is analogous, with the main differences highlighted in Section 4.2. In the example, note that we could argue that $m^*_d(\omega^1,\omega^1,\omega^2,\omega^2)$ should be greater than $m^*_d(\omega^1,\omega^1,\dots,\omega^1)$ since, otherwise, senders 3 and 4 could increase their utility in state ω^1 by reporting ω^2 instead of ω^1. More generally, suppose that in some state ω there exists a subset C of at most k senders such that all senders in C prefer action 1 to action 0. Then, all k-resilient truthful mechanisms must satisfy that $m^*_d(\omega,\dots,\omega) \geq m^*_d(\vec{m})$ for all inputs $\vec{m}$ such that $m_i = \omega$ for all $i \notin C$.

Following this intuition, we make the following definitions. Let $\Gamma = (S,A,\Omega,p,u)$ be an information aggregation game with $\Omega = \{\omega^1,\dots,\omega^m\}$ and $|S| = n$. We say that a possible input $\vec{m} = (m_1,\dots,m_n)$ for m_d is *ω-pure* if $m_1 = m_2 = \ldots = m_n = \omega$ (i.e., if all m_j are equal to ω). We also say that an input is pure if it is ω-pure for some ω. Additionally, if $\omega \in \Omega$, we denote by $\vec{\omega}$ the ω-pure input $(\omega,\dots,\omega)$. Moreover, given two inputs $\vec{m} = (m_1,\dots,m_n)$ and $\vec{m}' = (m'_1,\dots,m'_n)$ for m_d, we say that $\vec{m} \prec_k \vec{m}'$ if the subset C of senders such that their input differs in $\vec{m}$ and $\vec{m}'$ has size at most k, and such that

(a) $\vec{m}$ is ω-pure for some ω and all senders in C strictly prefer action 1 to action 0 in state ω, or

(b) $\vec{m}'$ is ω-pure for some ω and all senders in C strictly prefer action 0 to action 1 in state ω.

By construction we have the following property of $\prec_k$.

Lemma 2. *A honest mechanism is k-resilient incentive-compatible for the senders if and only if*

$$\vec{m} \prec_k \vec{m}' \Longrightarrow m^*_d(\vec{m}) \leq m^*_d(\vec{m}')$$

for all inputs $\vec{m}$ and $\vec{m}'$.

Note that Lemma 2 completely characterizes the honest mechanisms that are k-resilient incentive-compatible for the senders. However, this lemma is of little use by itself since mechanisms have an exponential number of possible inputs. Let $\leq_k$ be the partial order between pure states induced by $\prec_k$. More precisely, we say that two states ω and ω' satisfy $\omega \leq_k \omega'$ if there exists a sequence of inputs $\vec{m}^1,\dots,\vec{m}^t$ such that

$$\vec{\omega} \prec_k \vec{m}^1 \prec_k \ldots \prec_k \vec{m}^t \prec_k \vec{\omega}'.$$

For instance, in the example at the beginning of this section, we would have that $\omega^1 \leq_2 \omega^2$ since $(\omega^1,\omega^1,\omega^1,\omega^1) \prec_2 (\omega^1,\omega^1,\omega^2,\omega^2) \prec_2 (\omega^2,\omega^2,\omega^2,\omega^2)$. The following proposition shows that the $\leq_k$ relations completely determine the outcomes implementable by honest mechanisms that are k-resilient incentive-compatible for the senders.

Proposition 3. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game. Then, an outcome o of Γ is implementable by an honest mechanism that is k-resilient incentive-compatible for the senders if and only if*

$$\omega \leq_k \omega' \Longrightarrow o^*(\omega) \leq o^*(\omega')$$

for all $\omega, \omega' \in \Omega$.

Proof. The fact that any honest mechanism that is k-resilient incentive-compatible for the senders implies $\omega \leq_k \omega' \Longrightarrow o^*(\omega) \leq o^*(\omega')$ follows directly from Lemma 2.

To show the converse, given o satisfying $\omega \leq_k \omega' \Longrightarrow o^*(\omega) \leq o^*(\omega')$, define m_d^* as follows. If $\vec{m}$ is ω-pure for some ω, then $m_d^*(\vec{m}) := o^*(\omega)$. Otherwise, let $A^k_{\prec}(\vec{m})$ be the set of inputs $\vec{m}'$ such that $\vec{m} \prec_k \vec{m}'$ and $A^k_{\succ}(\vec{m})$ be the set of inputs $\vec{m}'$ such that $\vec{m}' \prec_k \vec{m}$. Then,

- If $A^k_{\prec}(\vec{m}) = \emptyset$, then $m_d^*(\vec{m}) := 1$.
- Otherwise, if $A^k_{\succ}(\vec{m}) = \emptyset$, then $m_d^*(\vec{m}) := 0$.
- Otherwise,
$$m_d^*(\vec{m}) := \frac{\min_{\vec{m}' \in A^k_{\prec}(\vec{m})}\{m_d^*(\vec{m}')\} + \max_{\vec{m}' \in A^k_{\succ}(\vec{m})}\{m_d^*(\vec{m}')\}}{2}.$$

Note that m_d^* is well-defined since all elements in $A^k_{\prec}(\vec{m})$ and $A^k_{\succ}(\vec{m})$ are pure, which means that $m_d^*(\vec{m}')$ is already defined for all these elements. Moreover, the honest mechanism M determined by m_d^* implements o. It remains to show that M is k-resilient incentive-compatible for the senders. By Lemma 2 this reduces to show that $\vec{m} \prec_k \vec{m}' \Longrightarrow m_d^*(\vec{m}) \leq m_d^*(\vec{m}')$ for all inputs $\vec{m}$ and $\vec{m}'$. To show this, take a pure input $\vec{\omega}$ and another input $\vec{m}$ such that $\vec{\omega} \prec_k \vec{m}$. If $\vec{m}$ is ω'-pure, then $\vec{\omega} \prec_k \vec{m} \Longrightarrow \vec{\omega} \leq_k \vec{\omega}'$ and thus $m_d^*(\vec{\omega}) \leq m_d^*(\vec{\omega}')$. If $\vec{m}$ is not pure and $A^k_{\prec}(\vec{m}) = \emptyset$ we have by construction that $m_d^*(\vec{m}) = 1$, which is greater than $m_d^*(\vec{\omega})$. Otherwise, for all ω' such that $\vec{\omega}' \in A^k_{\prec}(\vec{m})$, we have that $\omega \leq_k \omega'$ and thus by assumption that $m_d^*(\vec{\omega}) \leq m_d^*(\omega')$. Therefore,

$$\frac{\min_{\vec{m}' \in A^k_{\prec}(\vec{m})}\{m_d^*(\vec{m}')\}}{2} \geq \frac{m_d^*(\vec{\omega})}{2}$$

Moreover, we have that

$$\frac{\max_{\vec{m}' \in A^k_{\succ}(\vec{m})}\{m_d^*(\vec{m}')\}}{2} \geq \frac{m_d^*(\vec{\omega})}{2}$$

since $\vec{\omega} \in A^k_{\succ}(\vec{m}')$. Hence

$$m_d^*(\vec{m}) \geq m_d^*(\vec{\omega})$$

as desired. An analogous argument can be used for the case in which $\vec{m} \prec_k \vec{\omega}$. □

It remains to show that the partial order between the states in Ω defined by $\leq_k$ can be computed with a polynomial algorithm. To do this, note that, by definition, any chain

$$\vec{\omega} \prec_k \vec{m}^1 \prec_k \ldots \prec_k \vec{m}^t \prec_k \vec{\omega}'$$

between two pure inputs $\vec{\omega}$ and $\vec{\omega}'$ must satisfy that either $\vec{m}^1$ or $\vec{m}^2$ are also pure. This implies the following lemma:

Lemma 3. *Let $\Gamma = (S, A, \Omega, p, u)$ be an information aggregation game with $\Omega = \{\omega^1, \dots, \omega^m\}$. Let E a system of equations over $x_1, \dots, x_m$ such that equation $x_i \leq x_j$ appears in E if and only if $\vec{\omega}^i \prec_k \vec{\omega}_j$ or if there exists an input $\vec{m}$ such that $\vec{\omega}^i \prec_k \vec{m} \prec_k \vec{\omega}^j$. Then, $y_1, \dots, y_m$ is a solution of E if and only if*

$$\omega^i \leq_k \omega^j \Longrightarrow y_i \leq y_j$$

for all $i, j \in [m]$.

Intuitively, Lemma 3 says that the inequalities obtained from chains of length 2 or 3 *span* the partial order over Ω defined by $\leq_k$, and thus that we can take the system of equations E of Theorem 1 to be the one in the lemma above. Therefore, given two states ω and ω', it only remains to show that we can check in polynomial time if $\vec{\omega} \prec_k \vec{\omega}'$ or if there exists a state $\vec{m}$ such that $\vec{\omega} \prec_k \vec{m} \prec_k \vec{\omega}'$. Checking if $\vec{\omega} \prec_k \vec{\omega}'$ is equivalent to checking if $k = n$ and either all senders prefer 1 in ω or all senders prefer 0 in ω'. Finding an input $\vec{m}$ such that $\vec{\omega} \prec_k \vec{m} \prec_k \vec{\omega}'$ reduces to finding an input $\vec{m}$ such that

(a) the set C_ω of senders such that their message is not ω in $\vec{m}$ has size at most k, and all senders in C_ω strictly prefer 1 to 0 in ω.

(b) the set $C_{\omega'}$ of senders such that their message is not ω' in $\vec{m}$ has size at most k, and all of them strictly prefer 0 to 1 in ω'.

The high level idea of the algorithm is that, if $\vec{m}$ satisfies the above properties, all senders i that prefer 0 to 1 in ω must satisfy that $m_i = \omega$ (otherwise, it breaks property (a)), and all senders i that prefer 1 to 0 in ω' must satisfy that $m_i = \omega'$ (otherwise, it breaks property (b)). If there is a sender i that prefers 0 to 1 in ω and 1 to 0 in ω' then such an input $\vec{m}$ does not exist, and if there is a sender i that strictly prefers 1 to 0 in ω and 0 to 1 in ω', then m_i has no constraints. The only remaining restriction is that there can only be at most k values different than ω and at most k values different than ω' (note that this implies that if $2k < n$ such an input does not exist). The algorithm goes as follows:

1. Split the set of senders into four subsets $X_{0,1}^{0,1}, X_{0,1}^{1,0}, X_{1,0}^{0,1}, X_{1,0}^{1,0}$, in which $X_{i,j}^{i',j'}$ is the set of senders that prefer i to j in ω (resp., strictly prefer if $i = 1$) and prefer i' to j' in ω' (resp., strictly prefer if $i' = 0$).
2. If $X_{0,1}^{1,0} \neq \emptyset$ or $2k < n$, there is no solution.
3. If $|X_{0,1}^{0,1}| > k$ or $|X_{1,0}^{1,0}| > k$, there is no solution.
4. Otherwise, set $m_i = \omega$ for all $i \in X_{0,1}^{0,1}$, $m_i = \omega'$ for all $i \in X_{1,0}^{1,0}$. Then, set $k - |X_{0,1}^{0,1}|$ of the messages from $X_{1,0}^{0,1}$ to ω and the rest to ω'. Return $\vec{m}$.

Proof of Correctness: Because of the previous discussion, if $X_{0,1}^{1,0} \neq \emptyset$ or $2k < n$, there is no solution. If $|X_{0,1}^{0,1}| \geq k$ then, any input $\vec{m}$ that satisfies $\vec{\omega} \prec_k \vec{m} \prec_k \vec{\omega}'$ would require to have at least $|X_{0,1}^{0,1}|$ components equal to ω, which would break property (b). An analogous argument can be used when $|X_{1,0}^{1,0}| > k$. If none of these conditions hold, then we set all messages from $X_{0,1}^{0,1}$ to ω, all messages from $X_{1,0}^{1,0}$ to ω', and we split the messages sent by senders in $X_{1,0}^{0,1}$ between ω and ω' in such a way that no value appears more than k times. The resulting input satisfies properties (a) and (b).

4.2 Theorem 1, strong k-resilience

The proof of Theorem 1 for strong k-resilience is analogous to the one of k-resilience in the previous section. The main difference is the definition of $\prec_k$. In this case we say that two inputs $\vec{m}$ and $\vec{m}'$ satisfy

$\vec{m} \prec_k^s \vec{m}'$ if and only if the subset C of senders such that their input differs in $\vec{m}$ and $\vec{m}'$ has size at most k, and such that

(a) $\vec{m}$ is ω-pure for some ω and at least one sender in C strictly prefers action 1 to action 0 in state ω, or

(b) $\vec{m}'$ is ω-pure for some ω and at least one sender in C strictly prefers action 0 to action 1 in state ω.

We have that $\vec{\omega} \prec_k^s \vec{\omega}'$ if and only if $k = n$ and at least one sender in ω prefers action 1 to action 0, or at least one sender in ω' prefers action 0 to action 1. Given ω and ω', finding if there exists $\vec{m}$ such that $\vec{\omega} \prec_k^s \vec{m} \prec_k^s \vec{\omega}'$ can be reduced to finding if there exists a partition of the set of senders S into two sets S_ω and $S_{\omega'}$ such that $|S_\omega| \leq k$ and $|S_{\omega'}| \leq k$, and such that at least one sender of S_ω prefers action 0 to 1 in ω' and at least one sender of $S_{\omega'}$ prefers 1 to 0 in ω. This can easily be done in polynomial time.

For future reference, we define $\leq_k^s$ in the same way as $\leq_k$ except that we use $\prec_k^s$ instead of $\prec_k$.

5 Proof of Theorem 2

Most of the tools used to prove Theorem 2 have already appeared in the proof of Theorem 1. We prove the theorem for k-resilience, the case of strong k-resilience is analogous. Given a game Γ and an outcome o for Γ, we set $m_d^*(\vec{\omega}) := o^*(\omega)$ for each $\omega \in \Omega$. For every other input $\vec{m}$, we define $m_d^*(\vec{m})$ in the same way as in the proof of Proposition 3. As shown in the proof of Theorem 1, checking if $\vec{m} \prec_k \vec{m}'$ can be performed in polynomial time. Thus, $m_d^*(\vec{m})$ can also be computed in polynomial time.

6 Extended Model and Generalization of Main Results

An *extended information aggregation game* is defined in the same way as a standard information aggregation game (see Section 2) except that each sender starts the game with a private signal x_i (as opposed to all senders starting the game with the same input ω), and the utility function u takes as input the signals from each sender instead of just ω. More precisely, in an extended information aggregation game $\Gamma = (S, A, X, p, u)$ there is a set of senders $S = \{1, 2, 3, \ldots, n\}$, a receiver r, a mediator d, a set of actions A, a set $X = X_1 \times X_2 \times \ldots \times X_n$ of signals, a probability distribution p over X, and a utility function $u : (S \cup \{r\}) \times X \times A \longrightarrow \mathbb{R}$. Each game instance proceeds exactly the same way as in a standard information aggregation game except that, in phase 1, a signal profile $(x_1, \ldots, x_n) \in X$ is sampled following distribution p, and each signal x_i is disclosed only to sender i. In this context, an outcome o for Γ is just a function from signal profiles $\vec{x} \in X$ to distributions over A, and mechanisms for Γ are determined by functions m_d^* from X to $[0, 1]$.

Our aim is to generalize the results from Section 3 to the extended model. However, the main problem is that, for a fixed signal profile, the preferences of the agents may depend on their coalition. For instance, consider a game Γ for five players with uniformly distributed binary signals and binary actions such that the utility of each sender is 1 if the action that the receiver plays is equal to the majority of the signals, and their utility is 0 otherwise. Suppose that senders have signals $(0, 0, 0, 1, 1)$. It is easy to check that if players 1, 2 and 3 collude, player 1 would prefer action 0 to action 1. However, if players 1, 4 and 5 collude, player 1 would prefer action 1 since in this case it is more likely that the majority of the signals are 1.

We can avoid the issue above by assuming that the game is *k-separable*, which is that, for all signal profiles $\vec{x}$ and all senders i, there exists an action a such that the preference of sender i inside any coalition K of size at most k is a. Intuitively, an extended information aggregation game is *k-separable* if the preferences of the senders do not depend on the coalition they are in. With this, we can provide algorithms for the characterization and implementation of k-resilient truthful implementable outcomes that are efficient relative to the size of the description of the game Γ.

Theorem 3. *Let $\Gamma = (S, A, X, p, u)$ be a k-separable extended information aggregation game such that the support of signal profiles in distribution p is $\{(\vec{x})_1, \dots, (\vec{x})_m\}$. Then, there exists a system E of $O(m^2)$ equations over variables $x_1, \dots, x_m$, such that each equation of E is of the form $x_i \leq x_j$ for some $i, j \in [m]$, and such that an outcome o of Γ is implementable by a k-resilient truthful mechanism (resp., strong k-resilient truthful mechanism) if and only if*

(a) $x_1 = o^*((\vec{x})_1), \dots, x_m = o^*((\vec{x})_m)$ is a solution of E.

(b) $E_r(o) \geq U_a$ for all $a \in A$.

Moreover, the equations of E can be computed in polynomial time over m and the number of senders n.

Note that Theorem 3 states that E can be computed in polynomial time over the size of the support of signal profiles as opposed to $|X|$, which may be way larger. There is also a generalization of Theorem 2 in the extended model.

Theorem 4. *There exists an algorithm A that receives as input the description of a k-separable extended information aggregation game $\Gamma = (S, A, \Omega, p, u)$, an outcome o for Γ implementable by a k-resilient mechanism (resp., strong k-resilient mechanism), and a message input $\vec{m}$ for the mediator, and A outputs a value $q \in [0,1]$ such that the function m_d^* defined by $m_d^*(\vec{m}) := A(\Gamma, o, \vec{m})$ determines a k-resilient truthful mechanism (resp., strong k-resilient truthful mechanism) for Γ that implements o. Moreover, A runs in polynomial time over the size m of the support of signal profiles and $|S|$.*

The proofs of Theorems 3 and 4 are analogous to the ones of Theorems 1 and 2 with the following difference. Given two inputs $\vec{m}$ and $\vec{m}'$, we say that $\vec{m} \prec_k \vec{m}'$ if the subset C of senders such that their input differs in $\vec{m}$ and $\vec{m}'$ has size at most k, and such that

(a) $\vec{m}$ is in the support of p and all senders in C strictly prefer action 1 to action 0 given signal profile $\vec{m}$, or

(b) $\vec{m}'$ is is in the support of p and all senders in C strictly prefer action 0 to action 1 given signal profile $\vec{m}'$.

Intuitively, we replace the notion of *pure input* by the condition that the input is in the support of p. Note that the assumption of k-separability is crucial for this definition, since otherwise the preferences of the players may not be uniquely determined by the signal profile. With this definition, we can construct analogous statements for Lemmas 2, 3 and Proposition 3, and proceed identically as in the proofs of Theorems 1 and 2.

7 Conclusion

We provided an efficient characterization of all outcomes implementable by k-resilient and strong k-resilient truthful mechanisms in information aggregation games. We also gave an efficient construction of the k-resilient or strong k-resilient mechanism that implements a given implementable outcome. These techniques generalize to the extended model where senders may receive different signals, as long as the senders' preferences are not influenced by their coalition (k-separability). It is still an open problem to find if the techniques used in this paper generalize to other notions of coalition resilience as, for instance, the notion in which the sum of utilities of the members of a coalition cannot increase when defecting, or if we can get efficient algorithms in the extended model without the k-separability assumption. It is also an open problem to find if we can get similar results in partially synchronous or asynchronous systems in which the messages of the senders are delayed arbitrarily.

References

[1] I. Abraham, D. Dolev, R. Gonen & J. Y. Halpern (2006): *Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation*. In: *Proc. 25th ACM Symposium on Principles of Distributed Computing*, pp. 53–62, doi:10.1145/1146381.1146393.

[2] Ittai Abraham, Lorenzo Alvisi & Joseph Y Halpern (2011): *Distributed computing meets game theory: combining insights from two fields*. *ACM Sigact News* 42(2), pp. 69–76, doi:10.1145/1998037.1998055.

[3] Amitanand S Aiyer, Lorenzo Alvisi, Allen Clement, Mike Dahlin, Jean-Philippe Martin & Carl Porth (2005): *BAR fault tolerance for cooperative services*. In: *Proceedings of the twentieth ACM symposium on Operating systems principles*, pp. 45–58, doi:10.1145/1095810.1095816.

[4] Ali Aouad & Daniela Saban (2020): *Online assortment optimization for two-sided matching platforms*. *Available at SSRN 3712553*, doi:10.2139/ssrn.3712553.

[5] Robert J Aumann (1987): *Correlated equilibrium as an expression of Bayesian rationality*. *Econometrica: Journal of the Econometric Society*, pp. 1–18, doi:10.2307/1911154.

[6] Vincent P Crawford & Joel Sobel (1982): *Strategic information transmission*. *Econometrica: Journal of the Econometric Society*, pp. 1431–1451, doi:10.2307/1913390.

[7] Mahsa Derakhshan, Negin Golrezaei, Vahideh Manshadi & Vahab Mirrokni (2022): *Product ranking on online platforms*. *Management Science*, doi:10.1287/mnsc.2021.4044.

[8] John R. Douceur (2002): *The Sybil Attack*. In: *Peer-to-Peer Systems*, Springer Berlin Heidelberg, pp. 251–260, doi:10.1007/3-540-45748-8_24.

[9] Joseph Y Halpern (2008): *Beyond Nash equilibrium: Solution concepts for the 21st century*. In: *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, pp. 1–10, doi:10.1145/1400751.1400752.

[10] Emir Kamenica & Matthew Gentzkow (2011): *Bayesian persuasion*. *American Economic Review* 101(6), pp. 2590–2615, doi:10.3386/w15540.

[11] Emir Kamenica & Matthew Gentzkow (2017): *Competition in persuasion*. *Review of Economic Studies* 84(1), p. 1, doi:10.1093/restud/rdw052.

[12] Shehroze Khan & James R Wright (2021): *Disinformation, Stochastic Harm, and Costly Effort: A Principal-Agent Analysis of Regulating Social Media Platforms*. doi:10.48550/arXiv.2106.09847. Available at https://arxiv.org/abs/2106.09847.

[13] Andrew Kosenko (2018): *Mediated persuasion*. *SSRN Electronic Journal*, doi:10.2139/ssrn.3276453.

[14] Vijay Krishna & John Morgan (2001): *A model of expertise*. *The Quarterly Journal of Economics* 116(2), pp. 747–775, doi:10.2139/ssrn.150589.

[15] Hannah Li, Geng Zhao, Ramesh Johari & Gabriel Y Weintraub (2022): *Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms*. In: *Proceedings of the ACM Web Conference 2022*, pp. 182–192, doi:10.1145/3485447.3512063.

[16] Elliot Lipnowski & Doron Ravid (2020): *Cheap talk with transparent motives*. *Econometrica* 88(4), pp. 1631–1660, doi:10.3982/ecta15674.

[17] Dov Monderer & Moshe Tennenholtz (2009): *Strong mediated equilibrium*. *Artificial Intelligence* 173(1), pp. 180–195, doi:10.1016/j.artint.2008.10.005.

[18] Mary S Morgan & Margaret Morrison (1999): *Models as mediators*. Cambridge University Press Cambridge, doi:10.1017/CBO9780511660108.

[19] Roger B. Myerson (1979): *Incentive compatibility and the bargaining problem*. Econometrica 47(1), p. 61, doi:10.2307/1912346.

[20] Marcelo Olivares, Andres Musalem & Daniel Yung (2020): *Balancing agent retention and waiting time in service platforms*. In: Proceedings of the 21st ACM Conference on Economics and Computation, pp. 295–313, doi:10.2139/ssrn.3502469.

[21] Christos Papadimitriou, Kiran Vodrahalli & Mihalis Yannakakis (2022): *The Platform Design Problem*. In: Web and Internet Economics, Springer International Publishing, pp. 317–333, doi:10.1007/978-3-030-94676-0_18.

[22] Andrés Salamanca (2021): *The value of mediated communication*. Journal of Economic Theory 192, p. 105191, doi:10.1016/j.jet.2021.105191.

[23] Zheyuan Ryan Shi, Leah Lizarondo & Fei Fang (2021): *A Recommender System for Crowdsourcing Food Rescue Platforms*. In: Proceedings of the Web Conference 2021, pp. 857–865, doi:10.1145/3442381.3449787.

[24] Amandeep Singh, Jiding Zhang & Senthil K Veeraraghavan (2021): *Fulfillment by platform: Antitrust and upstream market power*. Available at SSRN 3859573, doi:10.2139/ssrn.3859573.

[25] Konstantinos I Stouras, Sanjiv Erat & Kenneth C Lichtendahl Jr (2020): *Prizes on crowdsourcing platforms: An equilibrium analysis of competing contests*. In: Proceedings of the 21st ACM Conference on Economics and Computation, pp. 875–876, doi:10.2139/ssrn.3485193.

Depth-bounded Epistemic Logic

Farid Arthaud
MIT
Cambridge, Massachusetts
farto@csail.mit.edu

Martin Rinard
MIT
Cambridge, Massachusetts
rinard@csail.mit.edu

Epistemic logics model how agents reason about their beliefs and the beliefs of other agents. Existing logics typically assume the ability of agents to reason perfectly about propositions of unbounded modal depth. We present **DBEL**, an extension of **S5** that models agents that can reason about epistemic formulas only up to a specific modal depth. To support explicit reasoning about agent depths, **DBEL** includes depth atoms E_a^d (agent a has depth exactly d) and P_a^d (agent a has depth at least d). We provide a sound and complete axiomatization of **DBEL**.

We extend **DBEL** to support public announcements for bounded depth agents and show how the resulting **DPAL** logic generalizes standard axioms from public announcement logic. We present two alternate extensions and identify two undesirable properties, amnesia and knowledge leakage, that these extensions have but **DPAL** does not. We provide axiomatizations of these logics as well as complexity results for satisfiability and model checking.

Finally, we use these logics to illustrate how agents with bounded modal depth reason in the classical muddy children problem, including upper and lower bounds on the depth knowledge necessary for agents to successfully solve the problem.

1 Introduction

Epistemic logics model how agents reason about their beliefs and the beliefs of other agents. These logics generally assume the ability of agents to perfectly reason about propositions of unbounded modal depth, which can be seen as unrealistic in some contexts [7, 19].

To model agents with the ability to reason only to certain preset modal depths, we extend the syntax of epistemic logic **S5** [8] to depth-bounded epistemic logic (DBEL). The **DBEL** semantics assigns each agent a depth in each state. For an agent to know a formula ψ in a given state of a model, the assigned depth of the agent must be at least the modal depth of ψ, i.e. $d(\psi)$. To enable agents to reason about their own and other agents' depths, **DBEL** includes **depth atoms** E_a^d (agent a has depth exactly d) and P_a^d (agent a has depth at least d). For example, the formula $K_a(P_b^5 \to K_b p)$ expresses the fact that, "agent a knows that whenever agent b is depth at least 5, agent b knows the fact p." Depth atoms enable agents to reason about agent depths and their consequences in contexts in which each agent may have complete, partial, or even no information about agent depths (including its own depth).

We provide a sound and complete axiomatization of **DBEL** (Section 2), requiring a stronger version of the LINDENBAUM lemma which ensures each agent can be assigned a depth (proven in Appendix B). Its satisfiability problem for two or more agents is immediately PSPACE-hard (because **DBEL** includes **S5** as a syntactic fragment). We provide a depth satisfaction algorithm for **DBEL** in PSPACE (Section 5), establishing that the **DBEL** satisfiability problem is PSPACE-complete for two or more agents.

Public announcement logic (PAL) [9] extends epistemic logic with public announcements. PAL includes the following public announcement and knowledge axiom (PAK), which characterizes agents' knowledge after public announcements,

$$[\varphi]K_a\psi \leftrightarrow (\varphi \to K_a[\varphi]\psi). \tag{PAK}$$

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 46–65, doi:10.4204/EPTCS.379.7

We extend **DBEL** to include public announcements (Section 3). The resulting depth-bounded public announcement logic (DPAL) provides a semantics for public announcements in depth-bounded epistemic logic, including a characterization of how agents reason when public announcements exceed their epistemic depth. We prove the soundness of several axioms that generalize (PAK) to **DPAL**, first in a setting where each agent has exact knowledge of its own depth, then in the general setting where each agent may have partial or even no knowledge of its own depth. We provide a sound axiom set for **DPAL** as well as an upper bound on the complexity of its model checking problem [1]

We also present two alternate semantics that extend **DBEL** with public announcements (Section 3.3). The resulting logics verify simpler generalizations of (PAK) in the context of depth-bounded agents, but each has one of two undesirable properties that we call *amnesia* and *knowledge leakage*. Amnesia causes agents to completely forget about all facts they knew after announcements, whereas knowledge leakage means shallow agents can infer information from what deeper agents have learned from a public announcement. **DPAL** suffers from neither of these two undesirable properties. We provide a sound and complete axiomatization of the first of the two alternate semantics (Section 4). We also prove the PSPACE-completeness of its satisfiability problem and show that its model checking problem remains P-complete (Section 5).

Finally, we use these logics to illustrate how agents with bounded depths reason in the muddy children reasoning problem [8]. We prove a lower bound and an upper bound on the structure of knowledge of depths required for agents to solve this problem (Section 6).

Related work Logical omniscience, wherein agents are capable of deducing any fact deducible from their knowledge, is a well-known property of most epistemic logics. The ability of agents to reason about facts to unbounded modal depth is a manifestation of logical omniscience. Logical omniscience has been viewed as undesirable or unrealistic in many contexts [8] and many attempts have been made to mitigate or eliminate it [8, 15, 17]. To the best of our knowledge, only Kaneko and Suzuki [11] below have involved modal depth in the treatment of logical omniscience in epistemic logic.

Kaneko and Suzuki [11] define the logic of shallow depths GL_{EF}, which relies on a set E of chains of agents $(i_1,\ldots,i_k)$ for which chains of modal operators $K_{i_1}\cdots K_{i_m}$ can appear. A subset $F \subseteq E$ restricts chains of modal operators along which agents can perform deductions about other agents' knowledge. An effect of bounding agents' depths in **DPAL** is creating a set of allowable chains of modal operators $\cup_a\{(a,i_1,\ldots,i_{d_a}),\ (i_1,\ldots,i_{d_a}) \in \mathscr{A}^{d_a}\}$. Unlike GL_{EF}, the bound on an agent's depth is not global in **DPAL**, it can also be a function of the worlds in the Kripke possible-worlds semantics [8]. In particular, **DPAL**, unlike GL_{EF}, enables agents to reason about their own depth, the depth of other agents, and (recursively) how other agents reason about agent depths. **DPAL** also includes public announcements, which to the best of our knowledge has not been implemented in GL_{EF}.

Kline [12] uses GL_{EF} to investigate the 3-agent muddy children problem, specifically by deriving minimal epistemic structures F that solve the problem. The proof relies on a series of belief sets with atomic updates called "resolutions," with the nested length of the chains in F providing epistemic bounds on the required reasoning. **DPAL**, in contrast, includes depth atoms and public announcements as first-class features. We leverage these features to directly prove theorems expressing that for k muddy children, *(i)* (Theorem 6.2) if the problem is solvable by an agent, that agent must have depth at least $k-1$ and know that it has depth at least $k-1$ (this theorem provides a lower bound on the agent depths required to solve the problem) and *(ii)* (Theorem 6.1) if an agent has depth at least $k-1$, knows it, knows another agent is depth at least $k-2$, knows that the other agent knows of another agent of depth $k-2$, *etc.*, then it

[1] Arthaud and Rinard [3] present a lower bound for this problem, as well as additional results, proofs and content.

can solve the problem (this theorem provides an upper bound on the agent depths necessary to solve the problem). Our depth bounds match the depth bounds of Kline [12] for 3 agents (Theorems 3.1 and 3.3 in [12]), though our bounds also provide conditions on recursive knowledge of depths for the agents as described above.

Dynamic epistemic logic (DEL) [6, 18] introduces more general announcements. Private announcements are conceptually similar to public announcements in **DPAL** in that they may be perceived by only some of the agents. In DEL, model updates depend only on the relation between states in the initial model and the relations in the action model. But in **DPAL**, model updates must also take into account the agent depths in the entire connected components of each state (see Definition 3).

Resource-bounded agents in epistemic logics have been explored by Balbiani et. al [5] (limiting perceptive and inferential steps), Artemov and Kuznets [2] (limiting the computational complexity of inferences), and Alechina et. al [1] (bounding the size of the set of formulas an agent may believe at the same time and introducing communication bounds). Alechina et. al [1] also bound the modal depth of formulas agents may believe, but all agents share the same depth bound and they leave open the question of whether inferences about agent depth or memory size could be implemented, which **DPAL** does.

2 Depth-bounded epistemic logic

The modal depth $d(\varphi)$ of a formula φ, defined as the largest number of modal operators on a branch of its syntactic tree, is the determining factor of the complexity of a formula in depth-bounded epistemic logic (DBEL). Modal operators are the main contributing factor to the complexity of model checking a formula; the recursion depth when checking satisfiability of a formula is equivalent to its modal depth [14]; and bounding modal depth often greatly simplifies the complexity of the satisfiability problem in epistemic logics [16]. Humans are believed to reason within limited modal depth [7, 19].

We extend the syntax of classical epistemic logic by assigning to each agent a in a set of agents $\mathscr{A}$ a depth $d(a,s)$ in each possible world s. The language also includes **depth atoms** E_a^d and P_a^d to respectively express that agent a has depth exactly d and agent a has depth at least d.

To know a formula φ, agents are required to be at least as deep as $d(\varphi)$ and also know that the formula φ is true in the usual possible-worlds semantics sense [8]. We translate the classical modal operator K_a from multi-agent epistemic logic into the operator K_a^∞ with the same properties, therefore $K_a^\infty\varphi$ can be interpreted as "agent a would know φ if a were of infinite depth". The operator $K_a\varphi$ will now take the meaning described above, i.e. $P_a^{d(\varphi)} \wedge K_a^\infty \varphi$.

Definition 1. The language of **DPAL** is inductively defined as, for all agents $a \in \mathscr{A}$ and depths $d \in \mathbb{N}$,

$$\mathscr{L}^\infty := \varphi = p \mid E_a^d \mid P_a^d \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid K_a^\infty\varphi \mid [\varphi]\varphi.$$

The K_a^∞ operator is used mainly as a tool in axiomatization proofs, we call $\mathscr{L}$ the fragment of our logic formulas without any K_a^∞ operators, which will be used in most of our theorems. We further define $\mathscr{H}^\infty$ and $\mathscr{H}$ to respectively be the syntactic fragments of $\mathscr{L}^\infty$ and $\mathscr{L}$ without public announcements $[\varphi]\psi$.

The modal depth d of a formula in $\mathscr{L}^\infty$ is inductively defined as,

$$d(p) = d\left(E_a^d\right) = d\left(P_a^d\right) = 0 \qquad d(\neg\varphi) = d(\varphi) \qquad d([\varphi]\psi) = d(\varphi) + d(\psi)$$
$$d(\varphi \wedge \psi) = \max(d(\varphi), d(\psi)) \qquad d(K_a\varphi) = 1 + d(\varphi) \qquad d(K_a^\infty\varphi) = 1 + d(\varphi).$$

We defer treatment of public announcements $[\varphi]\psi$ to Section 3. We work in the framework of **S5** [8], assuming each agent's knowledge relation to be an equivalence relation, unless otherwise specified—however, our work could be adapted to weaker epistemic logics [8] by removing the appropriate axioms.

All propositional tautologies	$p \to p$, *etc.*
Deduction	$(K_a\varphi \wedge K_a(\varphi \to \psi)) \to K_a\psi$
Truth	$K_a\varphi \to \varphi$
Positive introspection	$(K_a\varphi \wedge P_a^{d(\varphi)+1}) \to K_a(P_a^{d(\varphi)} \to K_a\varphi)$
Negative introspection	$(\neg K_a\varphi \wedge P_a^{d(\varphi)+1}) \to K_a\neg K_a\varphi$
Depth monotonicity	$P_a^d \to P_a^{d-1}$
Exact depths	$P_a^d \leftrightarrow \neg(E_a^0 \vee \cdots \vee E_a^{d-1})$
Unique depth	$\neg(E_a^{d_1} \wedge E_a^{d_2})$ for $d_1 \neq d_2$
Depth deduction	$K_a\varphi \to P_a^{d(\varphi)}$
Modus ponens	From φ and $\varphi \to \psi$, deduce ψ
Necessitation	From φ deduce $P_a^{d(\varphi)} \to K_a\varphi$

Table 1: Sound and complete axiomatization for **DBEL** over $\mathscr{H}$.

Definition 2. A model in **DBEL** is defined as a tuple $M = (\mathscr{S}, \sim, V, d)$ where $\mathscr{S}$ is a set of states, $V : \mathscr{S} \to 2^{\mathscr{P}}$ is the valuation function for atoms and $d : \mathscr{A} \times \mathscr{S} \to \mathbb{N}$ is a depth assignment function. For each agent a, $\sim_a$ is an equivalence relation on $\mathscr{S}$ modeling which states are seen as equivalent in the eyes of a. The semantics are inductively defined over $\mathscr{H}^\infty$ by,

$$(M,s) \models p \iff p \in V(s) \qquad (M,s) \models E_a^d \iff d(a,s) = d \qquad (M,s) \models P_a^d \iff d(a,s) \geq d$$

$$(M,s) \models \neg\varphi \iff (M,s) \not\models \varphi \qquad (M,s) \models \varphi \wedge \psi \iff (M,s) \models \varphi \text{ and } (M,s) \models \psi$$

$$(M,s) \models K_a^\infty\varphi \iff (\forall s',\ s \sim_a s' \implies (M,s') \models \varphi) \qquad (M,s) \models K_a\varphi \iff (M,s) \models P_a^{d(\varphi)} \wedge K_a^\infty\varphi.$$

Note that this definition does not require agents to have any (exact or approximate) knowledge of their own depth. On the other hand, it does not prohibit agents agents from having exact knowledge of their own depths, for instance we could model each agent carrying out some 'meta-reasoning' about its own depth [2] leading each agent to know its own depth exactly. These models are a subset of the class of the models we consider, which we study in more detail in Section 3.1.

As **DBEL** is an extension of **S5** up to renaming of the modal operators, one can expect for it to have a similar axiomatization: one new axiom is needed to axiomatize K_a and three others for depth atoms.

Theorem 2.1. *Axiomatization from Table 1 is sound and complete with respect to* **DBEL** *over* $\mathscr{H}$.

Proof. Rather than directly showing soundness and completeness, we show it is equivalent to the axiomatization of Table 3 in Appendix A on the fragment $\mathscr{H}$, which is shown to be sound and complete over $\mathscr{H}^\infty$ in Theorem A.1. We begin by proving any proposition in $\mathscr{H}$ that can be shown using Table 1 can be shown using Table 3 and then that any proof of a formula in $\mathscr{H}$ using the axioms in Table 3 can be shown using those in Table 1.

For the first direction, we prove that the axioms in Table 1 can be proven using those from Table 3. Most of them are immediate applications of bounded knowledge within the axioms of Table 3, along with tautologies when necessary. For positive and negative introspection, see equation (6) below in the proof of the opposite direction of the equivalence. We prove the least evident axiom, the deduction axiom, here as an example:

$$\text{Deduction} \quad (K_a^\infty\varphi \wedge K_a^\infty(\varphi \to \psi)) \to K_a^\infty\psi \tag{1}$$

[2] For instance deducing $P_a^{d(\varphi)}$ from the fact that it knows φ, or deducing $\neg P_a^n$ from the fact that it does not know $K_a^n\top$.

$$\begin{array}{rlr}
\text{Bounded knowledge in (1)} & (K_a^\infty \varphi \land K_a^\infty(\varphi \to \psi)) \to P_a^{d(\psi)} \to K_a\psi & (2)\\
\text{Tautology in (2)} & P_a^{\max(d(\varphi),d(\psi))} \to K_a^\infty \varphi \to K_a^\infty(\varphi \to \psi) \to P_a^{d(\psi)} \to K_a\psi & (3)\\
\text{Repeated depth consistency} & P_a^{\max(d(\varphi),d(\psi))} \to (P_a^{d(\varphi)} \land P_a^{d(\psi)}) & (4)\\
\text{Bounded knowledge and (3) and (4)} & P_a^{\max(d(\varphi),d(\psi))} \to K_a\varphi \to K_a(\varphi \to \psi) \to K_a\psi & (5)\\
\text{Bounded knowledge in (5)} & K_a\varphi \to K_a(\varphi \to \psi) \to K_a\psi. &
\end{array}$$

In the other direction, we will show by induction over a proof of a valid formula in $\mathscr{H}$ using Table 3 that it can be transformed into a proof with the same conclusion, using only axioms from Table 1. The transformation of a proof in the first axiomatization is as follows,

- If an item of the proof is a propositional tautology, replace all $K_a^\infty \varphi$ subformulas by $P_a^{d(\varphi)} \to K_a\varphi$, clearly the tautology still holds and it is in Table 1.
- If an item is an instance of the bounded knowledge axiom, replace it with the formula $K_a\varphi \leftrightarrow (P_a^{d(\varphi)} \land P_a^{d(\varphi)} \to K_a\varphi)$ which is a consequence of depth deduction and a tautology (and therefore can be added to the proof with two extra steps).
- If it uses any of the other axioms, replace it with the corresponding axiom (with the same name) from Table 1.

We now have a sequence that has the same conclusion (since the conclusion is in $\mathscr{H}$) and only uses axioms from Table 1. The last thing to show for this to be a proof in this axiomatization is that all applications of *modus ponens* and necessitation are still correct within this sequence. To this end, we show by induction that each step of the sequence is the same as the original proof where every $K_a^\infty \varphi$ subformula in each step has been replaced by $P_a^{d(\varphi)} \to K_a\varphi$.

First, note that this is the case for the two first bullet points of our transformation rules above. This is also true of each axiom in the table after our transformation: a proof similar to the one in equation (1) will yield the equivalence for deduction, the only remaining non-trivial cases are positive and negative introspection. For positive introspection, performing the substitution yields,

$$(P_a^{d(\varphi)} \to K_a\varphi) \to P_a^{d(\varphi)+1} \to K_a(P_a^{d(\varphi)} \to K_a\varphi). \tag{6}$$

Through application of a tautology and the depth monotonicity axiom we find it to be equivalent to, $P_a^{d(\varphi)+1} \to K_a\varphi \to K_a(P_a^{d(\varphi)} \to K_a\varphi)$. Therefore, up to adding steps to the proof and using tautologies, we can prove the axiom from Table 1 from the axiom in Table 3 after the substitution. The same can be said of negative introspection through a similar transformation.

Finally, since *modus ponens* and necessitation also maintain the property of replacing $K_a^\infty \varphi$ subformulas in each step by $P_a^{d(\varphi)} \to K_a\varphi$, it is true that the transformed proof is indeed a proof of the same conclusion in Table 1's axiomatization. □

3 Depth-bounded public announcement logic

We next present how to incorporate depth announcements in **DBEL**, which are a key challenge in defining depth-bounded public announcement logic (DPAL). Recall the axiom (PAK) of public announcement logic, $[\varphi]K_a\psi \leftrightarrow (\varphi \to K_a[\varphi]\psi)$. For the right-hand side to be true, agent a must be of depth $d([\varphi]\psi) = d(\varphi) + d(\psi)$ according to **DBEL**. This suggests that an agent must "consume" $d(\varphi)$ of its depth every time an announcement φ is made, meaning that an agent's depth behaves like a depth budget with respect to public announcements.

Moreover, to model that some agents might be too shallow for the announcement φ, each possible world is duplicated in a *negative* version where the announcement has not taken place and a *positive* version where the announcement takes place in the same way as in PAL. Agents who are not deep enough to perceive the announcement see the negative and positive version of the world as equivalent.

Definition 3. Models in **depth-bounded public announcement logic** (DPAL) are defined the same way as in **DBEL** and the semantics is extended to $\mathscr{L}^\infty$ by $(M,s) \models [\varphi]\psi \iff ((M,s) \models \varphi \implies (M \mid \varphi, (1,s)) \models \psi)$, where we define $M \mid \varphi$ to be the model $(\mathscr{S}', \sim', V', d')$, where,

$$\mathscr{S}' = (\{0\} \times \mathscr{S}) \cup \{(1,s),\ s \in \mathscr{S},\ (M,s) \models \varphi\}$$

$$\sim'_a \text{ is the transitive symmetric closure of } R_a \text{ such that,}$$

$$\begin{aligned}
(i,s)\,R_a\,(i,s') &\iff s \sim_a s' && \text{for } i = 0,1\\
(1,s)\,R_a\,(0,s) &\iff (M,s) \not\models P_a^{d(\varphi)}\\
V'((i,s)) &= V(s) && \text{for } i = 0,1\\
d'(a,(0,s)) &= d(a,s)\\
d'(a,(1,s)) &= \begin{cases} d(a,s) & \text{if } d(a,s) < d(\varphi)\\ d(a,s) - d(\varphi) & \text{otherwise.}\end{cases}
\end{aligned} \tag{7}$$

Since public announcements are no longer unconditionally and universally heard by all agents, we revisit the axiom (PAK) in **DPAL**. The determining factor is **depth ambiguity**: agents that are unsure about their own depth introduce uncertainty about which agents have perceived the announcement.

3.1 Unambiguous depths setting

A model verifies the **unambiguous depths** setting whenever each agent knows its own depth exactly:

$$\forall a,s,s', \quad s \sim_a s' \implies d(a,s) = d(a,s'). \tag{8}$$

The proof of the following theorem is given as Proposition C.1 in Appendix C.

Theorem 3.1. *For all $\varphi \in \mathscr{L}^\infty$, the following two properties, respectively called **knowledge preservation** and **traditional announcements**, are valid in **DPAL** in the unambiguous depths setting,*

$$\forall \psi \in \mathscr{L}_a^\infty,\ \neg P_a^{d(\varphi)} \to ([\varphi]K_a\psi \leftrightarrow (\varphi \to K_a\psi)) \tag{KP}$$

$$\forall \psi \in \mathscr{L}^\infty,\ P_a^{d(\varphi)} \quad \to ([\varphi]K_a\psi \leftrightarrow (\varphi \to K_a[\varphi]\psi)), \tag{TA}$$

where $\mathscr{L}_a^\infty$ is the fragment of $\mathscr{L}^\infty$ without depth atoms or modal operators for agents other than a.

Discussion Knowledge preservation (KP) means that an agent who is not deep enough to perceive an announcement φ must not change its knowledge of a formula ψ. However, such a property could not be true of all formulas ψ, for instance if $\psi = K_aK_bp$ but b is deep enough to perceive φ, then the depth adjustment formula (7) could mean that b's depth is now 0, making ψ no longer hold. Even when a is certain about b's depth, its uncertainty about what the announcement entails could also mean that formulas such as $\neg K_bp$ could no longer be true if $P_b^{d(\varphi)}$ and $\varphi \to p$ in the model. This demonstrates that in depth-bounded logics public announcements must introduce uncertainty: if a is unsure what b has perceived, it can no longer hold any certainties about what b does not know. This is not the case in PAL since all agents perceive all announcements. Our treatment of the depth-ambiguous case in Section 3.2 generalizes (KP) to obtain a property (KP') that holds on all formulas in $\mathscr{L}^\infty$.

Traditional announcements (TA) ensures that announcements behave the same as in PAL when the agent is deep enough for the announcement. The caveats from the discussion of (KP) no longer apply here, as any K_b operator that appears in ψ will still appear after the same public announcement operator, meaning that depth variations or knowledge variations are accounted for.

3.2 Ambiguous depths setting

We now abandon the depth unambiguity assumption from equation (8), and explore how properties (KP) and (TA) generalize to settings without depth unambiguity. We find a condition that ensures that sufficient knowledge about other agents' depths is given to a in order to maintain its recursive knowledge about other agents. The proof to the following theorem is given as Proposition C.2 in Appendix C.

Theorem 3.2. *For any $\varphi \in \mathscr{L}^\infty$, let $\mathscr{F}_\varphi : \mathscr{L}^\infty \to \mathscr{L}^\infty$ be inductively defined as,*

$$\mathscr{F}_\varphi(p) = \mathscr{F}_\varphi(E_a^d) = \mathscr{F}_\varphi(P_a^d) = \top \qquad \mathscr{F}_\varphi(\neg\psi) = \mathscr{F}_\varphi(\psi) \qquad \mathscr{F}_\varphi(\psi \wedge \chi) = \mathscr{F}_\varphi(\psi) \wedge \mathscr{F}_\varphi(\chi)$$

$$\mathscr{F}_\varphi(K_a\psi) = \neg K_a^\infty(\varphi \to P_a^{d(\varphi)}) \wedge K_a^\infty(\varphi \to \neg P_a^{d(\varphi)} \vee P_a^{d(\varphi)+d(\psi)}) \wedge K_a^\infty \mathscr{F}_\varphi(\psi)$$

$$\mathscr{F}_\varphi(K_a^\infty\psi) = \neg K_a^\infty(\varphi \to P_a^{d(\varphi)}) \wedge K_a^\infty \mathscr{F}_\varphi(\psi) \qquad \mathscr{F}_\varphi([\psi_1]\psi_2) = \mathscr{F}_\varphi(\psi_1) \wedge \mathscr{F}_\varphi(\psi_2).$$

*For all $\varphi \in \mathscr{L}^\infty$, the following two properties are valid in **DPAL**,*

$$\forall\psi \in \mathscr{L}^\infty,\ \mathscr{F}_\varphi(K_a\psi) \quad \to ([\varphi]K_a\psi \leftrightarrow (\varphi \to K_a\psi)) \tag{KP'}$$

$$\forall\psi \in \mathscr{L}^\infty,\ K_a^\infty(\varphi \to P_a^{d(\varphi)}) \to ([\varphi]K_a\psi \leftrightarrow (\varphi \to K_a[\varphi]\psi)). \tag{TA'}$$

3.3 Alternate treatments of model updates for public announcements

One question is whether using a definition of public announcements closer to PAL would produce a version of the above axioms closer to (PAK). Eager depth-bounded public announcement logic (EDPAL) below unconditionally decrements the depth value of all agents after public announcements.

Definition 4 (EDPAL). EDPAL extends the **DBEL** semantics to include public announcements by defining $(M,s) \models [\varphi]\psi \iff ((M,s) \models \varphi \implies (M \mid \varphi, s) \models \psi)$, where $M \mid \varphi$ is the model $(\mathscr{S}', \sim', V, d')$ in which $\mathscr{S}' = \{s \in \mathscr{S},\ (M,s) \models \varphi\}$, $\sim'_a$ is the restriction of $\sim_a$ to $\mathscr{S}'$, $d'(a,s) = d(a,s) - d(\varphi)$, and d may take values in $\mathbb{Z}$.

EDPAL has a sound and complete axiomatization based on the axiomatization of **DBEL** (Theorem 4.1), which also allows us to prove the complexity result of Theorem 5.1.

However, another consequence of its definition is that excessive public announcements in **EDPAL** can lead an agent to a state in which it cannot reason anymore, as it has consumed its entire depth budget.

Proposition 3.3 (Amnesia)**.** *In **EDPAL**, the formula $\neg P_a^{d(\varphi)} \to [\varphi]\neg K_a\psi$ is valid for all φ and ψ.*

Proof. If $(M,s) \not\models \varphi$ then the implicand is true. If $(M,s) \models \varphi \wedge \neg P_a^{d(\varphi)}$ then the depth of a in $(M \mid \varphi, s)$ will be at most -1, meaning that $(M \mid \varphi, s) \not\models K_a\psi$ for all ψ. □

In particular, for $\psi = \top$ one notices that standard intuitions about knowledge fail in **EDPAL**. This property is undesirable: *(i)* one may expect agents to maintain some knowledge even after public announcements that they are not deep enough to understand and *(ii)* deeper agents should be able to continue to benefit from the state of knowledge of shallower agents even after the shallower agents have exceeded their depth.

One way to try to remedy this property is to change model updates in **EDPAL** to make agents perceive announcements only when they are deep enough to understand them. The resulting asymmetric depth-bounded public announcement logic (ADPAL) removes depth from an agent's budget only when it is deep enough for an announcement, and only updates its equivalence relation in states where it is deep enough for the announcement.

Definition 5 (**ADPAL**). **ADPAL** extends the **DBEL** semantics to include public announcements by defining $(M,s) \models [\varphi]\psi \iff ((M,s) \models \varphi \implies (M \mid \varphi, s) \models \psi)$, where $M \mid \varphi$ is the model $(\mathscr{S}, \sim', V, d')$,

$$s \not\sim'_a s' \iff s \not\sim_a s' \text{ or } \begin{cases} (M,s) \models P_a^{d(\varphi)} \\ (M,s) \models \varphi \iff (M,s') \not\models \varphi, \end{cases}$$

$$d'(a,s) = \begin{cases} d(a,s) & \text{if } d(a,s) < d(\varphi) \\ d(a,s) - d(\varphi) & \text{otherwise.} \end{cases}$$

The relations $\sim_a$ are only assumed to be reflexive (as opposed to equivalence relations earlier).

Unfortunately, in **ADPAL** an agent that is too shallow for an announcement could still learn positive information that was learned by another agent who is deep enough to perceive the announcement. We call this property *knowledge leakage* as reflected in the following proposition.

Proposition 3.4 (Knowledge leakage). ***ADPAL** does not verify the $\rightarrow$ direction of* (KP').

Proof. Consider three worlds, $\{0,1,2\}$ and three agents a,b,c. The relations for a and c are identity, the relation for b is the symmetric reflexive closure of, $0 \sim_b 1 \sim_b 2$. The depth of a is 1 everywhere, b's depth is $0,2,0$ in each respective state and the depth of c is 2 everywhere. The atom p_0 is true only in 0 and 1. Consider $\varphi = K_c K_c p_0$, it is true in 0 and 1 only, and consider $\psi = K_b p_0$. $K_a \psi$ is not true in state 1, however $[\varphi] K_a \psi$ is. Moreover, one can easily check that $\mathscr{F}_\varphi(K_a \psi)$ is true in that state. □

The proof provides a practical example of such leakage in **ADPAL** and we further demonstrate knowledge leakage in Proposition 6.4 in the muddy children reasoning problem (see Section 6).

Note how each direction of the equivalence in (KP') expresses ($\rightarrow$) that no knowledge leakage occurs and ($\leftarrow$) no amnesia occurs. As shown in Theorem 3.2, **DPAL** verifies both directions and thus has neither amnesia nor knowledge leakage. As reflected in the following proposition, although EDPAL has amnesia, it doesn't have knowledge leakage and verifies (TA).

Proposition 3.5. *[3] **EDPAL** verifies* (TA) *and the $\rightarrow$ direction in* (KP) *over $\psi \in \mathscr{L}^\infty$, but not the converse.*

4 Axiomatizations

Theorem 4.1. *The axiomatization in Table 2 is sound and complete with respect to **EDPAL** (Definition 4) over the fragment $\mathscr{L}$.*

All axioms from Table 1	
Atomic permanence	$[\varphi]p \leftrightarrow (\varphi \to p)$
Depth adjustment	$\forall d \in \mathbb{Z},\ [\varphi]E_a^d \leftrightarrow \left(\varphi \to E_a^{d(\varphi)+d}\right)$
Negation announcement	$[\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$
Conjunction announcement	$[\varphi](\psi \wedge \chi) \leftrightarrow ([\varphi]\psi \wedge [\varphi]\chi)$
Knowledge announcement	$[\varphi](P_a^{d(\psi)} \to K_a\psi) \leftrightarrow (\varphi \to P_a^{d(\varphi)+d(\psi)} \to K_a[\varphi]\psi)$
Announcement composition	$[\varphi][\psi]\chi \leftrightarrow ([\varphi \wedge [\varphi]\psi]\chi)$
Modus ponens	From φ and $\varphi \to \psi$, deduce ψ
Necessitation	From φ deduce $P_a^{d(\varphi)} \to K_a\varphi$

Table 2: Sound and complete axiomatization of **EDPAL** over $\mathscr{L}$.

Proof. Similarly to the proof of Proposition 2.1, rather than directly showing soundness and completeness we show it is equivalent to the axiomatization of Table 4, which is shown to be sound and complete for **EDPAL** in Theorem A.2 in Appendix A.

In the first direction, all axioms in Table 2 can be shown using those in Table 4 immediately, either from the proof of Proposition 2.1 or because they are the same. The only difficulty lies in knowledge announcement, but a proof similar to equation (1) shows it is sound.

The other direction also follows the exact same proof as in Proposition 2.1: the public announcement axioms are direct translations of the same axioms in Table 4 by replacing the $K_a^\infty\varphi$ subformulas with $P_a^{d(\varphi)} \to K_a\varphi$. The proof transformation from Proposition 2.1 therefore still yields a proof of the same formula in this axiomatization, which proves completeness. □

We now present a sound set of axioms for **DPAL**. The main missing axioms for a sound and complete axiomatization are knowledge and public announcements, which we explored in the previous section, and announcement composition. In fact, announcement composition cannot exist in **DPAL**, since making a single announcement of depth $d_1 + d_2$ can behave very differently from making an announcement of depth d_1 followed by another of depth d_2, for instance when an agent's depth is between d_1 and $d_1 + d_2$.

Theorem 4.2. *Replacing knowledge announcement by* (KP') *and* (TA') *and depth adjustment by,*

$$\forall d \in \mathbb{N}, \quad [\varphi]E_a^d \leftrightarrow \left(\varphi \to \left((P_a^{d(\varphi)} \wedge E_a^{d+d(\varphi)}) \vee (\neg P_a^{d(\varphi)} \wedge E_a^d)\right)\right)$$

in Table 2 produces a set of sound axioms with respect to ***DPAL*** [3].

Proof. Theorem 3.2 verifies the two axioms (KP') and (TA'). The proofs for most axioms follows from Theorem 4.1 and that knowledge is defined the same way in both semantics. In particular, atomic permanence and conjunction announcement axioms are proven in Theorem 3.1's induction for (KP).

We are left to show depth adjustment,

$$\begin{aligned}
(M,s) \models [\varphi]E_a^d &\iff (M,s) \models \varphi \implies (M \mid \varphi, (1,s)) \models E_a^d \\
&\iff (M,s) \models \varphi \implies \begin{cases} d(a,s) = d + d(\varphi) & \text{if } d(a,s) \geq d(\varphi) \\ d(a,s) = d & \text{if } d(a,s) < d(\varphi) \end{cases} \\
&\iff (M,s) \models \varphi \to \left((P_a^{d(\varphi)} \wedge E_a^{d+d(\varphi)}) \vee (\neg P_a^{d(\varphi)} \wedge E_a^d)\right).
\end{aligned}$$

□

[3] One could also easily add axioms for K_a^∞ modal operators, for instance using those from Table 4 in Appendix A.

5 Complexity

We first state that adding depth bounds does not change the complexity of **S5** and PAL respectively.

Theorem 5.1. *The satisfiability problems for **DBEL** with $n \geq 2$ agents and for **EDPAL** are PSPACE-complete.*

Proof. The lower bound results from PSPACE-completeness of $\mathbf{S5}_n$ for $n \geq 2$ [10] and PAL [14], respective syntactic fragments of **DBEL** and **EDPAL**.

For both logics, we begin by translating $K_a\varphi$ subformulas into $P_a^{d(\varphi)} \wedge K_a^\infty\varphi$, which only increases formula size at most linearly. Then, in the case of **EDPAL**, using the same translation as Lemma 9 of [14], we translate formulas with public announcement φ into equivalent formulas $t(\varphi)$ without public announcement such that $|t(\varphi)|$ is at most polynomial in $|\varphi|$ (this is possible because the axiomatization of K_a^∞ with relation to public announcements is the same).

We have therefore transformed our formula φ into an equivalent formula in the syntactic fragment without K_a operators or public announcements of polynomial size relative to the initial formula φ's size.

We can then use the ELE-World procedure from Figure 6 of [14] by re-defining types to accommodate for depth atoms. As a reminder, we define $\mathbf{cl}(\Gamma)$ for any set of formulas Γ to be the smallest set of formulas containing Γ and closed by single negation and sub-formulas. We then say that $\gamma \subseteq \mathbf{cl}(\Gamma)$ is a type if all of the following are true,

1. $\neg\psi \in \gamma$ if and only if $\psi \notin \gamma$ when ψ is not a negation
2. if $\psi \wedge \chi \in \mathbf{cl}(\Gamma)$ then $\psi \wedge \chi \in \gamma$ if and only if $\psi \in \gamma$ and $\chi \in \gamma$
3. if $K_a^\infty\psi \in \gamma$ then $\psi \in \gamma$
4. if $P_a^d \in \gamma$ then $\neg P_a^{d'} \notin \gamma$ and $E_a^{d'} \notin \gamma$ for all $d' < d$
5. if $E_a^d \in \gamma$ then $E_a^{d'} \notin \gamma$ for all $d' \neq d$ and $\neg P_a^{d'} \notin \gamma$ for $d' < d$
6. if $\neg P_a^d \in \gamma$ then there exists $d' < d$ such that $\neg E_a^{d'} \notin \gamma$
7. $\neg P_a^0 \notin \gamma$

Clearly, checking that a subset of $\mathbf{cl}(\Gamma)$ is not a type does not increase the space complexity of the algorithm. Lemma 18 from [14] remains true here, i.e. the procedure ELE-World returns true if and only if the formula is satisfiable. It is sufficient for this to show that any type has a consistent depth assignment for all agents, as it is clear that if any of the new rules introduced for depths are violated the formula is not satisfiable.

If the type contains E_a^d then it contains only one such depth atom per rule 5, the only $P_a^{d'}$ it contains are for $d' \leq d$ per rule 4, and it does not contain $\neg P_a^{d'}$ for $d' \leq d$ per rule 5, therefore $d(a) = d$ is a consistent setting. If it does not contain any E_a^d, it may contain a number of inequalities polynomial in $|\varphi|$, that admit a solution in $\mathbb{N}$ by rule 7. Therefore a possible algorithm is $d_0 = \max\{d',\ P_a^{d'} \in \gamma\}$ and then $d(a) = \min\{d',\ d' \geq d_0,\ \neg E_a^{d'} \notin \gamma\}$. If no P_a^d are in the type, then $d_0 = \min\{d',\ \neg P_a^{d'} \in \gamma\}$ and $d(a) = \max\{d',\ d' \leq d_0, \neg E_a^{d'} \notin \gamma\}$ are a possible choice (this choice will always be greater or equal to 0 because of rules 7 and 6 above). Finally, if there are no depth atoms in the type, the formula is clearly satisfiable for any choice of $d(a)$. □

The model checking problem remains P-complete in **DBEL**, using the same algorithm as for **S5** [8]. For **EDPAL** and **ADPAL**, the model checking problem is P-complete, as the same algorithm as PAL can be used, relying on the fact that model size can only decrease after announcements [13] (the lower bounds results from the fact that PAL is a fragment of both). This is however not the case of **DPAL**, where model size grows after announcements, potentially exponentially, in fact model checking in **DPAL** is NP-hard [3].

Theorem 5.2. *The complexity of model checking for finite models in* **DPAL** *is in* EXPTIME*. An upper bound in time complexity for checking φ in M is $O(2^{2|\varphi|}\|M\|)$, where $\|M\|$ is the sum of the number of states and number of pairs in each relation of M.*

Proof. The model-checking algorithm is the same as the one for public announcement logic [13]: a tree is built from subformulas φ, with splits introduced only for subformulas of the form $[\psi]\chi$, with ψ to the left and χ to the right. Treating a node labeled ψ means labeling each state in M with either ψ or $\neg\psi$. The tree is treated from bottom-left to the top, always going up first except when a node of the type $[\psi]\chi$ is found. In that case, since the nodes in the left sub-tree have been treated, we can build $M \mid \psi$ easily in time $O(\|M\|)$ from the truth value of ψ and the depth functions of M. Moreover, the size of $M \mid \psi$ is at most $4\|M\|$.

To see this, consider an equivalence class for $\sim_a$ in M of size k, it has exactly k^2 connections within it. The number of states it creates in $M \mid \psi$ is at most $2k$, and the number of connections it creates is at most $4k^2$. Each connection being in exactly one connected component means the bound holds.

Therefore we can recurse in the right sub-tree with $M \mid \varphi$ to check χ in time $O(2^{2|\chi|} \times 4\|M\|)$. Writing $O(\|M\|) \leq c\|M\|$ the time necessary to build $M \mid \varphi$, we find that checking $[\psi]\chi$ takes time at most $O((c + 2^{2|\psi|} + 2^{2|\chi|+2})\|M\|) = O(2^{2|[\psi]\chi|}\|M\|)$. □

6 Muddy children

Consider the well-known muddy children reasoning problem, where n children convene after playing outside with mud. $k \geq 1$ of them have mud on their foreheads, but have no way of knowing it. The father, an external agent, announces that at least one child has mud on their forehead. Then, he repeatedly asks if any child would like to go wash themselves. After exactly $k-1$ repetitions of the father's question, all muddy children understand they are muddy and go wash themselves. Readers unfamiliar with the reasoning problem and its solution are directed to Van Ditmarsch et. al [18]'s treatment using PAL.

Consider the set of states $\{0,1\}^n$, where each tuple contains n entries indicating for each child if they are muddy (1) or not (0). For the sake of simplicity and since it is of depth 0, we assume the father's announcement has taken place and therefore define the Kripke structure M_n with states $\{0,1\}^n \setminus \{0\}^n$ with the usual definition of the agents' knowledge relations [8]. We define the **DPAL** class of muddy children models to be models $\hat{M}_n$ extending M_n with any depth function. We name m_i the atom expressing that child i is muddy.

We number the agents in $[|0;n-1|]$, where the first k are muddy, and focus on the reasoning of one agent (without loss of generality agent 0) to understand that it is muddy. Recall the definition of the dual of public announcements, $\langle\varphi\rangle\psi := \neg[\varphi]\neg\psi$ and define the following series of formulas for $i \leq k$,

$$\varphi_i = \langle\neg K_{i-1}m_{i-1}\rangle\langle\neg K_{i-2}m_{i-2}\rangle\cdots\langle\neg K_1 m_1\rangle K_0 m_0.$$

Here φ_k states that if each of the children from $k-1$ to 1 announce one after the other they don't know they are muddy, then child 0 knows that they (child 0) are muddy [4] It is well known this formula is true for unbounded agents in M_n in PAL (it is also a consequence of Theorem 6.1 below). The following two

[4] These announcements are a sufficient subset of the full announcements $\wedge_{j=1,\ldots,n}\neg(K_j m_j \vee K_j\neg m_j)$ in the usual formulation.

theorems define a sufficient structure of knowledge of depths for the formula to be true and a necessary condition on the structure of knowledge of depths for it to be true.

Theorem 6.1 (Upper bound)**.** *For all three semantics,* $K_0\left(P_0^{k-1} \wedge K_1(P_1^{k-2} \wedge \cdots K_{k-1}(P_{k-1}^0)\cdots)\right) \to \varphi_k$ *is true in all muddy children models* $\hat{M}_n$ *in the initial state.*

Note that this formula directly provides an upper bound on the structure of depths and knowledge about depths: it shows a sufficient condition on the knowledge of depths for the problem to be solvable by agent 0. Moreover, the upper bound for one child readily generalizes to a sufficient condition for all children to understand they are muddy: each muddy child must know they are of depth at least $k-1$, know at least some other muddy child knows they are of depth at least $k-2$, and know that that other child knows some other muddy child knows they are of depth at least $k-3$, *etc*.

Proof. For the sake of simplicity and since it does not change the treatment of the problem, we assume $n = k$. We show the result for **DPAL**, as the treatments for **EDPAL** and **ADPAL** are similar.

We will show the result by induction over k. Denote $s_k = (1,\ldots,1)$ the true state of the world where all the children are muddy.

For $k = 2$, we assume $K_0P_0^1$ and want to show $\neg K_1m_1 \wedge [\neg K_1m_1]K_0m_0$. First notice that $(\hat{M}_2, s_2) \models \neg K_1m_1$, simply because it considers the state $(1,0)$ to also be possible. In the state $(0,1)$, child 1 knows it is muddy. Therefore, the set of states for the successful part of the model update will be $(1,(1,1))$ and $(1,(1,0))$. Moreover, since $K_0P_0^1$, it is deep enough in s_2 to not have any links to the unsuccessful part of the model update, therefore it knows m_0.

Consider some $k > 2$, we denote S_i the set of states that are "active" when considering φ_i. More precisely, we set $S_i = \{0,1\}^i \times \{1\}^{k-i} \setminus \{0\}^k$. We will show that after $k-i$ announcements, the remainder of the problem is equivalent to checking φ_i on the subgraph induced by the states S_i. This is evident for $i = k$ by definition, we now show by descending induction that it is equivalent to checking φ_2 on S_2, which we have just verified to be true.

Firstly, it is true that $(\hat{M}_n, s_k) \models \neg K_{k-1}m_{k-1}$ since child $k-1$ considers possible the state $(1,\ldots,1,0)$. The set of states in which $K_{k-1}m_{k-1}$ holds is exactly $(0,\ldots,0,1)$. Therefore, the model update will create a copy of all other states. We then notice that the set of states whose last component is 0 can be ignored in the rest of the problem: they are not reachable from s_k by any sequence of $\sim_i$ that does not contain $\sim_{k-1}$ and the rest of the formula φ_{k-1} to be checked does not use any modal operators for agent $k-1$ any more. These states will never be reached and can therefore be removed without altering the result of the rest of the execution.

We are therefore restricting ourselves, after the model update, to the set of states S_{k-1} in the positive part of the model. Note however there are still possibly links between the negative part of the model and S_{k-1} in the positive part of the model. We will show that these links have no effect on the checking of the rest of the formula, by showing that links for child i find themselves in $S_{k-1} \setminus S_i$: therefore, by the time we query modal operator i, the set of ignored states will contain all states with a link for child i.

For child $i < k-1$, the information we have about its depth is $K_0K_1\cdots K_iP_i^{k-1-i}$ before the model update. Therefore, we in particular know it is deep enough for the announcement (which is of depth $1 \leq k-1-i$) in the set of states in which the i first components might have changed compared to s_k but the last $k-1-i$ are all fixed to 1: this is exactly S_i.

We have shown that the recursive check in $M \mid \neg K_{k-1}m_{k-1}$ will take place on a set of states for which the execution is equivalent to S_{k-1} and on which we will have to check the formula φ_{k-1}. Finally, since the depths of each agent other than $k-1$ was at least 1 on S_{k-2}, they are reduced by 1 and the induction hypothesis on depths for $k-2$ is also verified. □

Theorem 6.2 (Lower bound)**.** *For* **DPAL***, the formula* $\varphi_k \to K_0 P_0^{k-1}$ *is true in all models* $\hat{M}_n$*.*

Proof. We use the notations from the proof of Theorem 6.1 above. Notice first that all of the announcements remain true when they are performed, because $\neg K^{\infty}_{k-1} m_{k-1} \to \neg K_{k-1} m_{k-1}$ and the implicant is true by the usual lower bound for muddy children (it takes k announcements for any child to know they are muddy).

Assume by contraposition that $d(0,s_k) = i < k-1$ or $d(0,\tilde{s}_k) = i < k-1$ initially, where $\tilde{s}_k$ is the state $(0,1,\dots,1)$ of $\hat{M}_n$. After i public announcements, it will be true that $\neg K_0 m_0$ still, as well as $\neg K_0 \neg E_0^0$ since each public announcement is of depth 1. The former is a consequence of the usual lower bound for muddy children, and can be derived from the proof in Theorem 6.1 using symmetry between 0 and $k-1-i$ after the i announcements and monotonicity of knowledge of atoms: if the depths are lower than they were in the previous proof, there are more states and more links in the updated model and therefore $\neg K_{k-1-i} m_{k-1-i}$ remains true.

Therefore in this model after i announcements, either s_k or $\tilde{s}_k$ sees agent 0 of depth 0 and both states are still connected by $\sim_0$. This means that for the next announcement, since $\neg K_0 m_0$ after each announcement except potentially the last using the same argument as above, we will have the chain of connections $(1,s_k) \sim'_0 (0,s_k) \sim'_0 (0,s'_k)$ or $(1,s_k) \sim'_0 (1,\tilde{s}_k) \sim'_0 (0,\tilde{s}_k)$. This means that by an immediate induction, after the $k-i$ announcements it is still true that $\neg K_0 m_0$: this is a contradiction with φ_k. □

A stronger lower bound for each child is available [3], with recursive conditions on the depth of all agents similarly to Theorem 6.1. This formula provides a lower bound on the knowledge of depths of the agent 0 to be able to solve the problem: it must be depth at least $k-1$ and know so. By symmetry, this generalizes to any child or any set of children solving the problem.

Finally, we present propositions that illustrate how *amnesia* in **EDPAL** (Proposition 3.3) and *knowledge leakage* in **ADPAL** (Proposition 3.4) manifest in the muddy children problem. These propositions are easily verified by computing explicitly the models after updates.

Proposition 6.3 (Amnesia in **EDPAL**)**.** *Consider the instance of muddy children* M_3*, where child* i *is unambiguously of depth* $2-i$*, i.e.* $d(i,\cdot) = 2-i$*. The formula* $\langle \neg K_2 m_2 \rangle \langle \neg K_1 m_1 \rangle \neg K_2 \top$ *is true in* **EDPAL** *but not in* **DPAL** *or* **ADPAL***. This means that in* **EDPAL***, after the first two announcements, agent* 2 *does not know anything anymore.*

Proposition 6.4 (Knowledge leakage in **ADPAL**)**.** *The formula* $\langle K_1 \neg K_2 m_2 \rangle K_1 K_0 m_0$ *is true in* **ADPAL** *but not in* **DPAL** *or* **EDPAL***. In* **ADPAL***, agent* 1 *has deduced the conclusion of agent* 0*'s reasoning, despite not being deep enough to perceive the announcement. Moreover, if agent* 0 *were of depth* 1 *it would not be true that* $\langle K_1 \neg K_2 m_2 \rangle K_0 m_0$*: agent* 0 *would not be able to deduce what agent* 1 *has deduced.*

Library Alongside this paper, we publish code for a library for multi-agent epistemic logic model checking and visualization in Python. It implements depth-unbounded PAL models as well as **DPAL**, **EDPAL** and **ADPAL**. The code is available in an online repository [4]. The code can also be used to generate illustrations of model updates in the muddy children reasoning problem [3] under the assumptions of Theorem 6.1 above.

Conclusion We have shown how **S5** and public announcement logic (PAL) can be extended to incorporate bounded-depth agents. We have shown completeness results for several of the resulting logics and explored the relationship between public announcements and knowledge in **DPAL**, as well as complexity bounds for these logics. We finally illustrated the behavior of depth-bounded agents in the muddy children reasoning problem, where we showed upper and lower bounds on depths (and recursive knowledge of depths) necessary and sufficient to solve the problem. These results extend epistemic logics to support formal reasoning about agents with limited modal depth.

References

[1] Natasha Alechina, Brian Logan, Nguyen Hoang Nga & Abdur Rakib (2008): *Reasoning about Other Agents' Beliefs under Bounded Resources*. In John-Jules Ch. Meyer & Jan M. Broersen, editors: *Knowledge Representation for Agents and Multi-Agent Systems, KRAMAS 2008, Sydney, Australia, September 17, 2008*, *Lecture Notes in Computer Science* 5605, Springer, pp. 1–15, doi:10.1007/978-3-642-05301-6_1.

[2] Sergei N. Artëmov & Roman Kuznets (2009): *Logical omniscience as a computational complexity problem*. In Aviad Heifetz, editor: *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2009), Stanford, CA, USA, July 6-8, 2009*, pp. 14–23, doi:10.1145/1562814.1562821.

[3] Farid Arthaud & Martin Rinard (2023): *Depth-bounded epistemic logic*. *CoRR* abs/2305.08607, doi:10.48550/arXiv.2305.08607.

[4] Farid Arthaud & Martin Rinard (2023): *Library for depth-bounded epistemic logic*. https://gitlab.com/farid-fari/depth-bounded-epistemic-logic.

[5] Philippe Balbiani, David Fernández-Duque & Emiliano Lorini (2016): *A Logical Theory of Belief Dynamics for Resource-Bounded Agents*. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, ACM, pp. 644–652, doi:10.5555/2936924.2937020.

[6] Alexandru Baltag & Lawrence S. Moss (2004): *Logics for Epistemic Programs*. *Synth.* 139(2), pp. 165–224, doi:10.1023/B:SYNT.0000024912.56773.5e.

[7] Thomas Bolander, Robin Engelhardt & Thomas S. Nicolet (2020): *The Curse of Shared Knowledge: Recursive Belief Reasoning in a Coordination Game with Imperfect Information*. *CoRR* abs/2008.08849, doi:10.48550/arXiv.2008.08849.

[8] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (1995): *Reasoning About Knowledge*. MIT Press, doi:10.7551/mitpress/5803.001.0001.

[9] Jelle Gerbrandy & Willem Groeneveld (1997): *Reasoning about Information Change*. *J. Log. Lang. Inf.* 6(2), pp. 147–169, doi:10.1023/A:1008222603071.

[10] Joseph Y. Halpern & Yoram Moses (1992): *A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief*. *Artif. Intell.* 54(2), pp. 319–379, doi:10.1016/0004-3702(92)90049-4.

[11] Mamoru Kaneko & Nobu-Yuki Suzuki (2000): *Epistemic Logic of Shallow Depths and Game Theoretical Applications*. In Frank Wolter, Heinrich Wansing, Maarten de Rijke & Michael Zakharyaschev, editors: *Advances in Modal Logic 3, papers from the third conference on "Advances in Modal logic," held in Leipzig, Germany, 4-7 October 2000*, World Scientific, pp. 279–298, doi:10.1142/9789812776471_0015.

[12] J Jude Kline (2013): *Evaluations of epistemic components for resolving the muddy children puzzle*. *Economic Theory* 53(1), pp. 61–83, doi:10.1007/s00199-012-0735-x.

[13] Barteld Kooi & Johan Benthem (2004): *Reduction axioms for epistemic actions*. In: *Advances in Modal Logic 5, papers from the fifth conference on "Advances in Modal logic," held in Manchester, UK, 9-11 September 2004*, King's College Publications, pp. 197–211. Available at https://core.ac.uk/download/pdf/148195794.pdf.

[14] Carsten Lutz (2006): *Complexity and succinctness of public announcement logic*. In Hideyuki Nakashima, Michael P. Wellman, Gerhard Weiss & Peter Stone, editors: *5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan, May 8-12, 2006*, ACM, pp. 137–143, doi:10.1145/1160633.1160657.

[15] John-Jules Ch Meyer (2003): *Modal epistemic and doxastic logic*. *Handbook of philosophical logic*, pp. 1–38, doi:10.1007/978-94-017-4524-6_1.

[16] Linh Anh Nguyen (2004): *On the Complexity of Fragments of Modal Logics*. In Renate A. Schmidt, Ian Pratt-Hartmann, Mark Reynolds & Heinrich Wansing, editors: *Advances in Modal Logic 5, papers from the fifth conference on "Advances in Modal logic," held in Manchester, UK, 9-11 September 2004*, King's College Publications, pp. 249–268. Available at `http://www.aiml.net/volumes/volume5/Nguyen.ps`.

[17] Kwang Mong Sim (1997): *Epistemic logic and logical omniscience: A survey*. *Int. J. Intell. Syst.* 12(1), pp. 57–81, doi:10.1002/(SICI)1098-111X(199701)12:1<57::AID-INT3>3.0.CO;2-X.

[18] Hans Van Ditmarsch, Wiebe van Der Hoek & Barteld Kooi (2007): *Dynamic epistemic logic*. 337, Springer Science & Business Media, doi:10.1007/978-1-4020-5839-4.

[19] Rineke Verbrugge & Lisette Mol (2008): *Learning to Apply Theory of Mind*. *J. Log. Lang. Inf.* 17(4), pp. 489–511, doi:10.1007/s10849-008-9067-4.

A Axiomatization proofs

All propositional tautologies	$p \to p$, *etc.*
Deduction	$(K_a^\infty \varphi \wedge K_a^\infty(\varphi \to \psi)) \to K_a^\infty \psi$
Truth	$K_a^\infty \varphi \to \varphi$
Positive introspection	$K_a^\infty \varphi \to K_a^\infty K_a^\infty \varphi$
Negative introspection	$\neg K_a^\infty \varphi \to K_a^\infty \neg K_a^\infty \varphi$
Depth monotonicity	$P_a^d \to P_a^{d-1}$
Exact depths	$P_a^d \leftrightarrow \neg(E_a^0 \vee \cdots \vee E_a^{d-1})$
Unique depth	$\neg(E_a^{d_1} \wedge E_a^{d_2})$ for $d_1 \neq d_2$
Bounded knowledge	$K_a\varphi \leftrightarrow P_a^{d(\varphi)} \wedge K_a^\infty \varphi$
Modus ponens	From φ and $\varphi \to \psi$, deduce ψ
Necessitation	From φ deduce $K_a^\infty \varphi$

Table 3: Sound and complete axiomatization of **DBEL** over $\mathscr{H}^\infty$.

Theorem A.1. *Axiomatization from Table 3 is sound and complete with respect to* **DBEL** *over* $\mathscr{H}^\infty$.

Proof. Soundness of all of these axioms is immediate: the definition of K_a^∞ follows that of **S5** and so do the axioms, those concerning depth atoms are consequences of linear arithmetic, and the bounded knowledge axiom follows immediately from the definition of K_a in the semantics.

For completeness, first note we can translate any formula φ in $\mathscr{H}^\infty$ into an equivalent formula $t(\varphi)$ that does not contain any P_a^d atoms or K_a modal operators using the exact depths and bounded knowledge axioms (which we know to be sound). Call **S5D** this fragment of **DBEL**.

We will use a proof through the LINDENBAUM lemma and the truth lemma, to this end we need to complete the definition for the canonical model to add a depth function. As a reminder, the proof is as follows: if φ cannot be shown within the axiomatization in Table 1, i.e. $\not\vdash \varphi$, then we show that $\not\models \varphi$ by showing there is a state in the canonical model in which it does not hold.

The canonical model M^c is the model whose states are maximally consistent sets Γ of formulas for our axiomatization and whose states are related by $\sim_a$ if the set of formulas a knows is the same in both states. Its valuation function for atoms $V(\Gamma)$ is simply the set of axioms in Γ, i.e. $\Gamma \cap \mathscr{P}$.

We restrict M^c to sets Γ that contain at least some E_a^d for each agent $a \in \mathscr{A}$ and by the unique depth axiom we define $d(a,\Gamma) = \max\{d,\ E_a^d \in \Gamma\}$, since Γ contains exactly one depth to be consistent. This completes M^c into a **DBEL** model.

All propositional tautologies	$p \to p$, *etc.*
Deduction	$(K_a^\infty \varphi \wedge K_a^\infty(\varphi \to \psi)) \to K_a^\infty \psi$
Truth	$K_a^\infty \varphi \to \varphi$
Positive introspection	$K_a^\infty \varphi \to K_a^\infty K_a^\infty \varphi$
Negative introspection	$\neg K_a^\infty \varphi \to K_a^\infty \neg K_a^\infty \varphi$
Atomic permanence	$[\varphi]p \leftrightarrow \varphi \to p$
Depth adjustment	$\forall d \in \mathbb{Z},\ [\varphi]E_a^d \leftrightarrow \left(\varphi \to E_a^{d(\varphi)+d}\right)$
Negation announcement	$[\varphi]\neg\psi \leftrightarrow (\varphi \to \neg[\varphi]\psi)$
Conjunction announcement	$[\varphi](\psi \wedge \chi) \leftrightarrow ([\varphi]\psi \wedge [\varphi]\chi)$
Knowledge announcement	$[\varphi]K_a^\infty \psi \leftrightarrow (\varphi \to K_a^\infty[\varphi]\psi)$
Announcement composition	$[\varphi][\psi]\chi \leftrightarrow ([\varphi \wedge [\varphi]\psi]\chi)$
Depth monotonicity	$P_a^d \to P_a^{d-1}$
Exact depths	$P_a^d \leftrightarrow \neg(E_a^0 \vee \cdots \vee E_a^{d-1})$
Unique depth	$\neg(E_a^{d_1} \wedge E_a^{d_2})$ for $d_1 \neq d_2$
Bounded knowledge	$K_a\varphi \leftrightarrow P_a^{d(\varphi)} \wedge K_a^\infty \varphi$
Modus ponens	From φ and $\varphi \to \psi$, deduce ψ
Necessitation	From φ deduce $K_a^\infty \varphi$

Table 4: Sound and complete axiomatization of **EDPAL**.

Since $\nvdash \varphi$, the set $\{\neg\varphi\}$ is consistent for our axiomatization. We must now show we can extend this set into a maximal consistent set of formulas that contains a depth atom E_a^d for each agent a.

However, this stronger requirement is not satisfied by the usual LINDENBAUM lemma, since a consistent set of formulas could be $\{P_a^d,\ d \in \mathbb{N}\}$ (which is not consistent with any E_a^d). Note however we only need it to hold for a finite set of formulas (namely $\{\neg\varphi\}$): Lemma B.1 below proves this version of the LINDENBAUM lemma, by showing there must exist some E_a^d that is consistent with any finite set for each a, and then a maximally consistent set can be derived using the traditional LINDENBAUM lemma.

Finally, the truth lemma shows that $\varphi \in \Gamma \iff (M^c, \Gamma) \models \varphi$ by induction on φ and is enough to conclude (since the maximal consistent set containing $\neg\varphi$ will not verify φ). Most induction cases are the same as for **S5**, the only new symbols left in our formula φ are the E_a^d atoms, and the truth lemma is immediately true for them by definition of the depth function of M^c.

Finally, if $\models \varphi$, then $\models t(\varphi)$ by the soundness of the axiomatization and definition of the transformation, then **S5D** $\vdash t(\varphi)$ since we have just shown the completeness of this fragment. Finally, this must mean **DBEL** $\vdash t(\varphi)$ and then $\vdash \varphi$ since the transformations of t can be performed using equivalences in our axiomatization: we have shown completeness. □

Theorem A.2. *The axiomatization in Table 4 is sound and complete with respect to* ***EDPAL****.*

Proof. Soundness of the axioms of **DBEL** is proven in Theorem A.1. Soundness of all axioms for public announcement is also a consequence of their definition in PAL with which they share their definition, except for depth adjustment for which the proof is relatively immediate.

For completeness, we translate any formula φ into $t(\varphi)$ by removing public announcements, K_a modal operators and P_a^d atoms by using the sound axioms from Table 4. The formula $t(\varphi)$ is in the syntactic fragment **S5D**, thus we can use completeness shown in Theorem 2.1 to show $\vdash t(\varphi)$, which implies $\vdash \varphi$ within the axiomatization of Table 4 by using the same axioms in the opposite direction. □

B LINDENBAUM lemma with depth assignments

Lemma B.1. *For every agent a and finite consistent set of formulas Γ without public announcement, P_b^d literals or K_b operators for all b, there exists some $d \in \mathbb{N}$ such that $\Gamma \cup \{E_a^d\}$ is a consistent set.*

Proof. Fix agent a. As Γ is a finite set of finite formulas, the set of exact depth atoms for a that appear in its formulas is included in a finite set $F = \{E_a^0, \dots, E_a^D\}$ for some $D \in \mathbb{N}$.

We can add to Γ instances of the unique depth axiom for each pair of integers in $[|0;D|]$ while maintaining consistency. The set Γ can then be seen as a consistent set of formulas for **S5** over the set of atoms $F \cup \mathscr{P}$, i.e. consistent in the axiomatization of Table 3 without depth axioms or bounded knowledge (or tautologies involving symbols not in the language of **S5**). Therefore there is an **S5** model (M,s) that satisfies it by the usual LINDENBAUM lemma and the truth lemma (the canonical model here).

In (M,s), if any of the E_a^d are valued to $\top$, then at most one of them is satisfied (since we added the unique depth axiom for all pair of depths). If all of the E_a^d are valued to $\bot$, then we can introduce a new atom E_a^{D+1} and set its value to $\top$ in all states of the model. All of the unique depths axioms for $D+1$ and $d \leq D$ can be added to Γ without making it inconsistent.

In both cases, let d_0 be the value of the unique $E_a^{d_0}$ valued to $\top$ in this final model. We claim that $\{\varphi, E_a^{d_0}\}$ must be a consistent set. Indeed, a proof of its inconsistency with the axioms from Table 3 must only involve axioms from **S5** and unique depths axioms for the set F, since none of the symbols P_a^d or K_a are necessary in a proof (they can be replaced by their equivalents with E_a^d and K_a^∞ without changing the conclusion) and any occurrence of E_a^d for $d > D+1$ can be replaced by $\bot$ while maintaining the truthfulness and conclusion of the proof.

Therefore, such an inconsistency proof would also hold within **S5**, which is a contradiction with soundness since these formulas are verified in a consistent set (the set of true formulas in (M,s)). □

C Proofs for Section 3

Proposition C.1. *Formulas* (KP) *and* (TA) *are valid for* ***DPAL*** *in the unambiguous depths setting.*

Proof. To prove (KP), suppose without loss of generality that $(M,s) \models \neg P_a^{d(\varphi)} \wedge \varphi$. In particular, this means that in $M \mid \varphi$, we have $(0,s) \sim_a' (1,s)$ and therefore the equivalence class of $(1,s)$ in $M \mid \varphi$ contains all $(0,s')$ whenever $s' \sim_a s$. Then,

$$\begin{aligned}
(M,s) \models [\varphi]K_a\psi &\iff (M,s) \models \varphi \implies (M \mid \varphi,(1,s)) \models K_a\psi \\
&\iff (M \mid \varphi,(1,s)) \models P_a^{d(\psi)} \text{ and } \forall s', j, (j,s') \sim_a' (1,s) \implies (M \mid \varphi,(j,s')) \models \psi \\
&\iff (M \mid \varphi,(1,s)) \models P_a^{d(\psi)} \text{ and } \forall s' \sim_a s, \begin{cases} (M \mid \varphi,(0,s')) \models \psi \\ (M,s') \models \varphi \implies (M \mid \varphi,(1,s')) \models \psi. \end{cases}
\end{aligned} \tag{9}$$

On the other hand,

$$(M,s) \models K_a\psi \iff (M,s) \models P_a^{d(\psi)} \text{ and } \forall s', s' \sim_a s \implies (M,s') \models \psi. \tag{10}$$

We prove by structural induction over $\psi \in \mathscr{H}_a$ the stronger equivalence,

$$\forall s' \sim_a s, \quad \begin{cases} (M,s') \models \psi \iff (M \mid \varphi,(0,s')) \models \psi \\ (M,s') \models \varphi \implies ((M,s') \models \psi \iff (M \mid \varphi,(1,s')) \models \psi). \end{cases} \tag{11}$$

Given that $(M,s) \not\models P_a^{d(\varphi)}$, we have $(M \mid \varphi,(1,s)) \models P_a^{d(\psi)} \iff (M,s) \models P_a^{d(\psi)}$. Therefore the depth conditions in equations (9) and (10) are the same and since both sides are true if $(M,s) \not\models \varphi$, equation (11) is enough to prove (KP).

For $\psi \in \mathscr{P}$, it is true because $V'((j,s')) = V(s')$ for all j and s' (note that $\mathscr{P}$ does not include depth atoms). For depth atoms about a, it is a consequence of $(M,s) \models K_a \neg P_a^{d(\varphi)}$ by the depth unambiguity condition (8), which means the depth of a is unchanged in all $s' \sim_a s$ after the model update.

The cases where $\psi = \psi_1 \wedge \psi_2$ and $\psi = \neg\chi$ are immediate, by the way these operators coincide with the usual propositional logic definition on both sides of the equivalences.

If $\psi = K_a\chi$ and $s' \sim_a s$, recall that by the depth unambiguity condition (8) we have $(M,s') \models \neg P_a^{d(\varphi)}$. Therefore, if $(M,s') \models \varphi$,

$$\begin{aligned}
(M \mid \varphi,(1,s')) \models \psi &\iff d(a,s') \geq d(\chi) \text{ and } \forall (j,s'') \sim_a' (1,s'),\ (M \mid \varphi,(j,s'')) \models \chi \\
&\iff d(a,s') \geq d(\chi) \text{ and } \forall s'' \sim_a s,\ \begin{cases} (M \mid \varphi,(0,s'')) \models \chi \\ (M,s'') \models \varphi \implies (M \mid \varphi,(1,s'')) \models \chi \end{cases} \\
&\iff d(a,s') \geq d(\chi) \text{ and } \forall s'' \sim_a s',\ (M,s'') \models \chi \\
&\iff (M,s') \models \psi,
\end{aligned}$$

where we have used the induction hypothesis (11) for χ once in each direction. The first equivalence in equation (11) is even easier to verify, by the same technique. The case for $\psi = K_a^\infty\chi$ is directly implied by this proof, as there are no depth conditions to verify.

To prove public announcements, we will need a stronger induction hypothesis than (11). Write for any s, $1_0(s) = s$ and $1_n(s) = (1, 1_{n-1}(s)) = (1,\ldots,(1,s))$. We posit,

$$\begin{aligned}
&\forall n \in \mathbb{N},\ \forall \psi_1,\ldots,\psi_n,\ \forall s' \sim_a s,\ (M,s') \models P_a^{d(\psi_1)+\cdots d(\psi_n)+d(\psi)} \text{ and } (M,s') \models \neg P_a^{d(\varphi)} \implies \\
&(M,s') \models \psi_1 \text{ and } (M \mid \psi_1,(1,s')) \models \psi_2 \text{ and } \ldots \text{ and } (M \mid \psi_1 \mid \cdots \mid \psi_{n-1}, 1_{n-1}(s')) \models \psi_n \implies \\
&\begin{cases} (M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi \iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((0,s'))) \models \psi \\ (M,s') \models \varphi \implies ((M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi \iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((1,s'))) \models \psi). \end{cases}
\end{aligned} \tag{12}$$

Note we slightly abuse notation here and some of these states might not exist, the convention is that the equivalences need only hold when the states exist in the models on both sides. The implicant implies that the left-hand term always exists.

Taking this for $n = 0$ is sufficient to conclude on (KP), since both equations (9) and (10) will be false whenever $(M,s) \not\models P_a^{d(\psi)}$.

The cases for atoms, negations and conjunction are clear for the same reasons as they were in equation (11). The case for depth atoms for a is direct, since the assumption $(M,s') \models P_a^{d(\psi_1)+\cdots d(\psi_n)}$ implies that the depth of a after the $\psi_1,\ldots,\psi_n$ announcements is its initial depth minus the sum of the depths of all the announcements, and the assumption that it is not deep enough for φ means its depth does not change with the announcement of φ.

The case for modal operators K_a relies on the fact that depth atoms are preserved (by the induction hypothesis for depth atoms) and the relations verify in $M \mid \psi_1 \mid \cdots \mid \psi_n$ when these states exist,

$$(1,(1,\ldots(1,s_1))) \sim_a^{(n)} (1,(1,\ldots(1,s_2))) \iff s_1 \sim_a s_2,$$

by denoting $\sim_a^{(k)}$ the relation for a in a model after k announcements. And similarly in $M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n$,

$$(1,(1,\ldots(j,s_1))) \sim_a^{(n+1)} (1,(1,\ldots(k,s_2))) \iff (j,s_1) \sim_a' (k,s_2) \iff s_1 \sim_a s_2.$$

This also implies the case for K_a^∞, since the verification is the same without the depth condition.

Finally, if $\psi = [\psi']\chi$, we verify that for $s' \sim_a s$ such that $(M,s') \models \varphi$,

$$\begin{aligned}
&(M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi \\
&\iff (M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi' \implies (M \mid \psi_1 \mid \cdots \mid \psi_n \mid \psi', 1_{n+1}(s')) \models \chi \\
&\iff (M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi' \implies (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n \mid \psi', 1_{n+1}((1,s'))) \models \chi \\
&\iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((1,s'))) \models \psi' \implies (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n \mid \psi', 1_{n+1}((1,s'))) \models \chi \\
&\iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((1,s'))) \models \psi
\end{aligned} \tag{13}$$

when the latter state exists. Our first use of the induction hypothesis on χ is justified because the left-hand side of the implication is the $n+1$ term in the assumptions for the induction hypothesis in equation (12) (and $d(\psi) = d(\psi') + d(\chi)$). The second use of the induction hypothesis on ψ' is justified for the same depth reason and the other assumptions remain the same. Once more the case for $(0,s')$ is very similar.

For (TA), we assume without loss of generality that $(M,s) \models K_a(P_a^{d(\varphi)}) \wedge \varphi$ (using the depth unambiguity condition (8)), this means in particular the equivalence class of $(1,s)$ in $M \mid \varphi$ is $\{(1,s'),\ s' \sim_a s,\ (M,s') \models \varphi\}$ since no state equivalent to s by $\sim_a$ has a not deep enough for φ. Using the same reasoning as in equation (9), we have,

$$(M,s) \models [\varphi]K_a\psi \iff (M \mid \varphi,(1,s)) \models P_a^{d(\psi)} \text{ and } \forall s' \sim_a s,\ (M,s') \models \varphi \implies (M \mid \varphi,(1,s')) \models \psi.$$

Moreover, we have,

$$(M,s) \models K_a[\varphi]\psi \iff (M,s) \models P_a^{d(\varphi)+d(\psi)} \text{ and } \forall s' \sim_a s,\ (M,s') \models \varphi \implies (M \mid \varphi,(1,s')) \models \psi. \tag{14}$$

Since $(M,s) \models P_a^{d(\varphi)}$, the depth of a in $(M \mid \varphi,(1,s))$ is its depth in (M,s) minus $d(\varphi)$. This means that

$$(M \mid \varphi,(1,s)) \models P_a^{d(\psi)} \iff (M,s) \models P_a^{d(\varphi)+d(\psi)}. \qquad \square$$

Proposition C.2. ***DPAL** verifies* (KP') *and* (TA').

Proof. For (KP'), in light of equations (9) and (10), we use the following induction hypothesis,

$$\begin{aligned}
&\forall s,a, \quad (M,s) \models K_a^\infty \mathscr{F}_\varphi(\psi) \implies \\
&\forall s' \sim_a s, \quad \begin{cases} (M \mid \varphi,(0,s')) \models \psi \iff (M,s') \models \psi \\ (M,s') \models \varphi \implies ((M \mid \varphi,(1,s')) \models \psi \iff (M,s') \models \psi). \end{cases}
\end{aligned} \tag{15}$$

Assume that $(M,s) \models \varphi \wedge \mathscr{F}_\varphi(K_a\psi)$. In particular, $(M,s) \models \neg K_a^\infty(\varphi \to P_a^{d(\varphi)})$. First notice that this condition allows us to write, $(0,s') \sim_a' (1,s) \iff s' \sim_a s$. Indeed, since there exists some $s'' \sim_a s$ where a is of depth strictly less than $d(\varphi)$ and φ holds, we deduce the chain of connections, $(1,s) \sim_a' (1,s'') \sim_a' (0,s'') \sim_a' (0,s')$ for any $s' \sim_a s$ (and the direct implication is immediate).

Moreover, we have assumed $(M,s) \models \varphi \wedge (\varphi \to \neg P_a^{d(\varphi)} \vee P_a^{d(\varphi)+d(\psi)})$. In either case of the disjunction, the depth conditions of equations (9) and (10) become equivalent as they did in the proof of (KP). Therefore, proving the induction hypothesis (15) is sufficient to conclude (KP') here.

The cases for atoms, negations and conjunctions is the same as in the proof of Proposition C.1, as the induction hypothesis holds because $\mathscr{F}_\varphi(\neg\psi) = \mathscr{F}_\varphi(\psi)$, $\mathscr{F}_\varphi(\psi_1 \wedge \psi_2) = \mathscr{F}_\varphi(\psi_1) \wedge \mathscr{F}_\varphi(\psi_2)$, and by commutativity of K_a^∞ with conjunction.

If $\psi = K_b\chi$ for some agent $b \in \mathscr{A}$, for some fixed $s' \sim_a s$, we know that $(M,s') \models \neg K_b^\infty(\varphi \to P_b^{d(\varphi)})$ as well as $(M,s') \models K_b^\infty \mathscr{F}_\varphi(\chi)$. Moreover, the condition $(M,s') \models K_b^\infty(\varphi \to \neg P_b^{d(\varphi)} \vee P_b^{d(\varphi)+d(\chi)})$ implies that the depth of b will be greater or equal to $d(\chi)$ in $(M \mid \varphi, (1,s'))$ if and only if it was in (M,s'). If $(M,s') \models \varphi$, by once more using the induction hypothesis (15) for b in s', we obtain that,

$$\begin{aligned}
(M \mid \varphi, (1,s')) \models \psi &\iff d(b,s') \geq d(\chi) \text{ and } \forall (j,s'') \sim_b' (1,s'),\ (M \mid \varphi, (j,s'')) \models \chi \\
&\iff d(b,s') \geq d(\chi) \text{ and } \forall s'' \sim_b s',\ \begin{cases} (M \mid \varphi, (0,s'')) \models \chi \\ (M,s'') \models \varphi \implies (M \mid \varphi, (1,s'')) \models \chi \end{cases} \\
&\iff d(b,s') \geq d(\chi) \text{ and } \forall s'' \sim_b s',\ (M,s'') \models \chi \\
&\iff (M,s') \models \psi.
\end{aligned}$$

The case for $(0,s')$ is the same, since its equivalence class in $M \mid \varphi$ is the same and the depth condition is the same. The case for $\psi = K_b^\infty \chi$ is implied by this proof, as there are no depth conditions to verify.

Finally, checking public announcements involves performing the same induction hypothesis strengthening as in the proof of (KP) in its equation (12). The new induction hypothesis becomes,

$$\begin{aligned}
&\forall s,a, \quad (M,s) \models K_a^\infty \mathscr{F}_\varphi(\psi) \implies \\
&\forall n \in \mathbb{N},\ \forall \psi_1, \ldots, \psi_n,\ \forall s' \sim_a s,\ (M,s') \models P_a^{d(\psi_1)+\cdots d(\psi_n)+d(\psi)} \text{ and } (M,s') \models \neg P_a^{d(\varphi)} \implies \\
&(M,s') \models \psi_1 \text{ and } (M \mid \psi_1, (1,s')) \models \psi_2 \text{ and } \ldots \text{ and } (M \mid \psi_1 \mid \cdots \mid \psi_{n-1}, 1_{n-1}(s')) \models \psi_n \implies \\
&\begin{cases} (M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi \iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((0,s'))) \models \psi \\ (M,s') \models \varphi \implies ((M \mid \psi_1 \mid \cdots \mid \psi_n, 1_n(s')) \models \psi \iff (M \mid \varphi \mid \psi_1 \mid \cdots \mid \psi_n, 1_n((1,s'))) \models \psi). \end{cases}
\end{aligned}$$

Note we slightly abuse notation here and some of these states might not exist, the convention is that the equivalences need only hold when the states exist in the models on both sides. The implicant implies that the left-hand term always exists.

Checking atoms, depth atoms, negation and conjunction is the same as in the proof of (KP) once more. Checking modal operators K_a and K_a^∞ is similar to the proof of (KP) using induction hypothesis (12), but using the same reasoning as above for induction hypothesis (15): the induction hypothesis contained in $\mathscr{F}_\varphi$ tells us that the announcement is not perceived by the agent at each modal operator.

Finally, public announcements follow the exact same proof as they did in (KP) in equation (13), with the extra information that $\mathscr{F}_\varphi([\psi']\chi) = \mathscr{F}_\varphi(\psi') \wedge \mathscr{F}_\varphi(\chi)$, allowing us to obtain the assumption of the inductive hypothesis in both inductive hypothesis applications (one for ψ' and one for χ).

For (TA'), we assume without loss of generality that $(M,s) \models K_a^\infty(\varphi \to P_a^{d(\varphi)}) \wedge \varphi$, this means in particular the equivalence class of $(1,s)$ in $M \mid \varphi$ is $\{(1,s'),\ s' \sim_a s,\ (M,s') \models \varphi\}$ since no state equivalent to s by $\sim_a$ has a not deep enough for φ. Using once more the same re-writings as in equation (14), it is sufficient to prove that the depth conditions are the same. This is the case because $(M,s) \models \varphi$, therefore by the truth axiom, $(M,s) \models P_a^{d(\varphi)}$. □

Comparing Social Network Dynamic Operators

Edoardo Baccini
University of Groningen
e.baccini@rug.nl

Zoé Christoff
University of Groningen
z.l.christoff@rug.nl

Numerous logics have been developed to reason either about threshold-induced opinion diffusion in a network, or about similarity-driven network structure evolution, or about both. In this paper, we first introduce a logic containing different dynamic operators to capture changes that are 'asynchronous' (opinion change only, network-link change only) and changes that are 'synchronous' (both at the same time). Second, we show that synchronous operators cannot, in general, be replaced by asynchronous operators and vice versa. Third, we characterise the class of models on which the synchronous operator *can* be reduced to sequences of asynchronous operators.

1 Introduction

There are two main types of change affecting agents connected through a social network. First, the features of an agent, e.g., their opinions or behavior, can be influenced by its neighbors in the network: for instance, if one's entire social circle has adopted an opinion in favor of (or against) vaccines, one is unlikely to disagree with this opinion. Under this type of social influence, or *social conformity pressure*, network-neighbors tend to align their opinions (or any other feature that can change) and therefore become more similar. Second, in addition to changing their own state (opinion, or other feature), agents can also reshape their social environment by connecting with others. What generally drives the formation of new links between two agents is their *similarity*. Both types of changes relate to how similar agents are: social influence makes network neighbors become more similar while new links make similar agents become more connected [9, Ch. 4].

In social network analysis, a common way of representing both types of dynamics is to assume that certain *thresholds* drive the dynamics. On the one hand, a typical way of representing social influence is via *threshold models* [12, 17, 8, 9]: agents adopt a feature when a large enough proportion of their network neighbors has already adopted it. On the other hand, the formation of new links has been modelled in a similar way. In probabilistic models, it is usual to assume that agents who are more similar are more likely to connect than those who are less similar [23, 4]. In deterministic models, this has been translated by a similarity threshold: two agents get connected as soon as they are similar enough [18, 20, 19, 21].

Both types of changes have been addressed in logic. Indeed, a number of logical frameworks has flourished to reason about threshold-based social influence [2, 6, 5, 16, 14], and about threshold-based link formation [18, 20, 19, 21]. Yet, to the exception of [11, 15, 1], either the two aspects have been treated separately [22, 18, 20, 19, 21, 10], or the two types of changes have been taken to happen one after the other [19]. To our knowledge, only [1] provides a logic capturing specifically *simultaneous* changes of the network structure and the state of the agents.

In this paper, we introduce a closely related framework that, similarly to [1] combines three dynamic operators: one corresponding to the change of the network structure only, one corresponding to the change of the agents feature (opinion/behavior/state) only, and one corresponding to *both changes at the same time*, but restricting ourselves to monotonic changes. We then tackle for this monotonic setting an

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 66–81, doi:10.4204/EPTCS.379.8

open question by [19, 1] in the literature: Can different sequences of dynamic operators be reduced to one another?

We first introduce the framework in Section 2. We then discuss the (ir)replaceability of the three dynamic operators in Section 3. We show in particular that our 'synchronous' operator cannot always be replaced by any sequences of other operators (Theorem 2). We also show that, when it can be replaced, the sequence of operators replacing it can only be of four specific types (Theorem 3). Finally, we characterize the class of models on which the synchronous operator can be replaced (Theorem 4).

2 Logic of asynchronous and synchronous network changes

We introduce a logic to reason about asynchronous and synchronous changes in social networks. We use a propositional language (where atoms are parametrized by our sets of agents and features) extended with three dynamic operators $\triangle, \square, \bigcirc$, to capture, respectively, diffusion update, network update, and both updates happening simultaneously.

Definition 1 (Syntax $\mathscr{L}$)**.** *Let $\mathscr{A}$ be a non-empty finite set of agents, $\mathscr{F}$ be a non-empty finite set of features. Let $\Phi_{at} := \{N_{ab} : a,b \in \mathscr{A}\} \cup \{f_a : f \in \mathscr{F}, a \in \mathscr{A}\}$ be the set of atomic formulas. The syntax $\mathscr{L}$ is the following:*

$$\varphi := N_{ab} \mid f_a \mid \neg\varphi \mid \varphi \wedge \varphi \mid \triangle\varphi \mid \square\varphi \mid \bigcirc\varphi$$

where $f \in \mathscr{F}$ and $a,b \in \mathscr{A}$.

The connectors $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined as usual. N_{ab} is read as 'agent a is an influencer of agent b'; f_a as 'agent a has feature f'; $\triangle\varphi$ as 'after a diffusion update, φ holds'; $\square\varphi$ as 'after a network update, φ holds'; $\bigcirc\varphi$ as 'after a synchronous update, φ holds'.

We now introduce the models representing who is influencing whom and who has which features, and our three different types of updates.

Definition 2 (Model M)**.** *Let $\mathscr{A}$ be a non-empty finite set of agents, $\mathscr{F}$ be a non-empty finite set of features. A model M over $\mathscr{A}$ and $\mathscr{F}$ is a tuple $\langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$, where:*

- *$\mathscr{N} \subseteq \mathscr{A} \times \mathscr{A}$ is a social influence relation;*
- *$\mathscr{V} : \mathscr{A} \longrightarrow \mathscr{P}(\mathscr{F})$ is a valuation function, assigning to each agent a set of adopted features;*
- *$\omega, \tau \in \mathbb{Q}$ are two rational numbers such that $0 \leq \omega \leq 1$ and $0 < \tau \leq 1$, interpreted, respectively, as similarity threshold and influenceability threshold.*

We write $C^{\omega\tau}$ for the class of all models for given values of ω and τ.

We turn to defining the three types of model updates corresponding to our three dynamic operators. First, after the diffusion (only) update, the set of features each agent adopt is updated. Agents might start adopting new features if enough of their neighbors had already adopted them before the update. Note that, while [19, 1] consider updates in which agents might start abandoning previously adopted features, here we restrict ourselves to the case in which agents are not allowed to start unadopting features, similarly as in [2].

Definition 3 (Diffusion update - $M_\triangle$)**.** *Given a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$, the updated model $M_\triangle = \langle \mathscr{N}, \mathscr{V}', \omega, \tau \rangle$ is such that for any $a,b \in \mathscr{A}$ and any $f \in \mathscr{F}$:*

$$f \in \mathscr{V}'(a) \text{ iff } \left\{ \begin{array}{ll} f \in \mathscr{V}(a), & \text{if } N(a) = \emptyset \\ f \in \mathscr{V}(a) \text{ or } \frac{|N_f(a)|}{|N(a)|} \geq \tau, & \text{otherwise} \end{array} \right\}$$

where $N_f(a) := \{b \in A : (b,a) \in \mathscr{N}$ and $f \in \mathscr{V}(b)\}$ and $N(a) := \{b \in A : (b,a) \in \mathscr{N}\}$.

The diffusion update does not affect the network structure. In contrast, the network update only affects the connections, not the features adopted by any of the agents. After a network update, new links may have formed between agents that agree on sufficiently many features. Just as agents could not unadopt previously adopted features, agents cannot break old connections, which differs for instance from [15, 19]. In this respect, our network update is a monotonic version of that in [19].

Definition 4 (Network update - $M_\Box$)**.** *Given a model $M = \langle \mathcal{N}, \mathcal{V}, \omega, \tau \rangle$, the updated model $M_\Box = \langle \mathcal{N}', \mathcal{V}, \omega, \tau \rangle$ is such that for any $a, b \in \mathcal{A}$ and any $f \in \mathcal{F}$:*

$$(a,b) \in \mathcal{N}' \text{ iff } (a,b) \in \mathcal{N} \text{ or } \frac{|(\mathcal{V}(a) \cap \mathcal{V}(b)) \cup (\mathcal{F} \setminus (\mathcal{V}(a) \cup \mathcal{V}(b)))|}{|\mathcal{F}|} \geq \omega$$

Third, the synchronous update affects features and connections at once. Adoption of new features happens under the same conditions as with the diffusion update, and new links are created under the same conditions as with the network update.

Definition 5 (Synchronous update - $M_\bigcirc$)**.** *Let $M = \langle \mathcal{N}, \mathcal{V}, \omega, \tau \rangle$, the model resulting from synchronous update is $M_\bigcirc := \langle \mathcal{N}', \mathcal{V}', \omega, \tau \rangle$, where $\mathcal{N}'$ is as in Definition 4 and $\mathcal{V}'$ is as in Definition 3.*

Now that we have defined the model-updates, we can introduce the semantic clauses for formulas containing the corresponding operators.

Definition 6 (Satisfaction)**.** *For any model $M = \langle \mathcal{N}, \mathcal{V}, \omega, \tau \rangle$ and any formula $\varphi \in \mathcal{L}$, the truth of φ in M is inductively defined as follows:*

$M \models f_a$ *if and only if* $f \in \mathcal{V}(a)$

$M \models N_{ab}$ *if and only if* $(a,b) \in \mathcal{N}$

$M \models \neg\varphi$ *if and only if* $M \not\models \varphi$

$M \models \varphi \wedge \psi$ *if and only if* $M \models \varphi$ *and* $M \models \psi$

$M \models \Box\varphi$ *if and only if* $M_\Box \models \varphi$

$M \models \triangle\varphi$ *if and only if* $M_\triangle \models \varphi$

$M \models \bigcirc\varphi$ *if and only if* $M_\bigcirc \models \varphi$

where $M_\triangle$ is the updated model as in Definition 3, and $M_\Box$ is the updated model as in Definition 4, and $M_\bigcirc$ is the updated model as in Definition 5.

As usual, we say that a formula is valid in a class of models if it is true in all models of that class and valid (tout court) if it is valid in all models.

Observation 1. *Let $M_1 = \langle \mathcal{N}_1, \mathcal{V}_1, \omega, \tau \rangle$ and $M_2 = \langle \mathcal{N}_2, \mathcal{V}_2, \omega, \tau \rangle$ be two models. The following are equivalent:*

- *for all $\varphi_{at} \in \Phi_{at}$, $M_1 \models \varphi_{at}$ iff $M_2 \models \varphi_{at}$*
- *for all $\varphi \in \mathcal{L}$, $M_1 \models \varphi$ iff $M_2 \models \varphi$*
- *$M_1 = M_2$*

We introduce the following two abbreviations capturing, respectively, when an agent a has sufficient pressure to adopt a feature ($f^\tau_{N(a)}$), and when two agents a and b have sufficient similarity to connect (sim^ω_{ab}).

$$f^\tau_{N(a)} := \bigvee_{\{G \subseteq N \subseteq A,\, N \neq \emptyset \,:\, \frac{|G|}{|N|} \geq \tau\}} \Big(\bigwedge_{b \in N} N_{ba} \wedge \bigwedge_{b \notin N} \neg N_{ba} \wedge \bigwedge_{b \in G} f_b \Big)$$

$\Box N_{ab} \leftrightarrow N_{ab} \vee sim^{\omega}_{ab}$	$\triangle N_{ab} \leftrightarrow N_{ab}$	$\bigcirc N_{ab} \leftrightarrow N_{ab} \vee sim^{\omega}_{ab}$
$\Box f_a \leftrightarrow f_a$	$\triangle f_a \leftrightarrow f_a \vee f^{\tau}_{N(a)}$	$\bigcirc f_a \leftrightarrow f_a \vee f^{\tau}_{N(a)}$
$\Box(\varphi \wedge \psi) \leftrightarrow \Box\varphi \wedge \Box\psi$	$\triangle(\varphi \wedge \psi) \leftrightarrow \triangle\varphi \wedge \triangle\psi$	$\bigcirc(\varphi \wedge \psi) \leftrightarrow \bigcirc\varphi \wedge \bigcirc\psi$
$\Box\neg\varphi \leftrightarrow \neg\Box\varphi$	$\triangle\neg\varphi \leftrightarrow \neg\triangle\varphi$	$\bigcirc\neg\varphi \leftrightarrow \neg\bigcirc\varphi$

From $\varphi_1 \leftrightarrow \varphi_2$, infer that $\varphi \leftrightarrow \varphi[\varphi_1/\varphi_2]$, where $\varphi[\varphi_1/\varphi_2]$ is a formula
obtained by replacing one or more occurrences of φ_1 with φ_2

Table 1: Reduction Axioms and derivation rule for the dynamic modalities $\Box, \triangle, \bigcirc$.

$$sim^{\omega}_{ab} := \bigvee_{\{E\subseteq\mathscr{F}:\frac{|E|}{|\mathscr{F}|}\geq\omega\}} \bigwedge_{f\in E} (f_a \leftrightarrow f_b)$$

These abbreviations can then be used to obtain reduction axioms for each of the dynamic modalities in $\mathscr{L}$, which are shown in Table 1. The reduction axioms for the dynamic operators in $\mathscr{L}$ are very similar to those in other dynamic logics of social network change. Indeed, the reduction axioms for the operator $\triangle$ are the same as the those of the dynamic operator $[adopt]$ in [2], with the exception that our logic captures multiple diffusing features and thus contains reduction axioms for each spreading feature in $\mathscr{F}$. In this sense, they resemble the reduction axioms in [19] with the difference that in our setting features cannot be unadopted. Moreover, the reduction axioms for the operator $\Box$ are similar to those in [19], with the difference that our framework does not allow for link deletion. The reduction axioms for the operator $\bigcirc$ merely reflect the fact that both features and links are affected by a synchronous update.

We will investigate how and when operators can replace one another in the next section. Before that, by looking at our axioms, we can immediately observe that an operator can replace another when it precedes specific formulas:

Observation 2. *Let $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be a model. For all $a, b \in \mathscr{A}$, for all $f \in \mathscr{F}$:*

- *$M \models \bigcirc f_a$ iff $M \models \triangle f_a$*
- *if $M \models \Box f_a$, then $M \models \bigcirc f_a$*
- *$M \models \bigcirc N_{ab}$ iff $M \models \Box N_{ab}$*
- *if $M \models \triangle N_{ab}$ then $M \models \bigcirc N_{ab}$*

Definition 7 (Logic $L^{\omega\tau}$)**.** *Let $\omega \in [0,1]$ and $\tau \in (0,1]$ be two rational numbers. The Logic $L^{\omega\tau}$ consists of some complete axiomatisation and derivation rules of propositional logic, together with the reduction axioms and the derivation rule in Table 1.*

Theorem 1. *Let $\omega \in [0,1]$ and $\tau \in (0,1]$ be two rational numbers. For any $\varphi \in \mathscr{L}$: $\models_{\mathscr{C}^{\omega\tau}} \varphi$ iff $\vdash_{L^{\omega\tau}} \varphi$*

The proof uses standard techniques and is very similar to that of the related settings in [2, 19, 1]: a sketch is included in the Appendix.

3 Irreplaceability of synchronous operators

Given that our dynamic formulas are reducible to the static fragment of our language, the question of comparing the expressivity of fragments of our language excluding one or two of the dynamic operators is uninteresting. In contrast, what is interesting, as suggested already in [19, 1], is to compare whether formulas containing some (specific combinations of) dynamic operators could be translated into formulas containing other (combinations of) dynamic operators. Another way to put it, closer to the way [19] first introduces the question, is to ask when different sequences of different model updates result in the same model.

To be able to investigate the extent to which our dynamic operators are inter-translatable or not (beyond the atomic preceding cases mentioned in Observation 2), we first have to introduce some notation and define the relevant type of expressivity criteria.

Definition 8 (Notation for sequences of operators). *Let $D = \{\bigcirc, \triangle, \Box\}$. For $O \subseteq D$, S_O denotes the set of all non-empty finite sequences of operators in O. We write $d_1 d_2 ... d_n$ for the sequence $\langle d_1, d_2, ..., d_n \rangle$ and d^n for the sequence consisting of $n \in \mathbb{N}$ repetitions of $d \in D$. We denote by $s^{j:k}$ the subsequence of s starting with the j-th element of s and ending with the k-th element of s. Given two sequences s_1, s_2 of lengths $n, m \in \mathbb{N}$, respectively, we write $s_1 s_2$ for the sequence of length $n+m$ obtained by prefixing s_1 to s_2.*

Definition 9 (Equivalence of sequences). *Two sequences $s_1, s_2 \in S_D$ are equivalent on a model M when $M_{s_1} = M_{s_2}$, or, equivalently (by Observation 1), when for all $\varphi \in \mathscr{L}$, $M \models s_1 \varphi$ if and only if $M \models s_2 \varphi$. Two sequences s_1 and s_2 are equivalent over a class of models when they are equivalent over all models in the class. Two sequences are equivalent (tout court) when they are equivalent over the class of all models.*

We start by making some observations about sequences of $\triangle$ and $\Box$ operators.

Observation 3. *Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given.*

- *Any sequence $s \in S_{\{\Box\}}$ is equivalent to the sequence $\Box$ on M.*
- *There exists an $n < |\mathscr{A}|$, such that, for any $m > n$, $\triangle^m$ is equivalent to $\triangle^n$ on M.*

The first point follows from the fact that the model update in Definition 4 is idempotent, and therefore $M \models \Box^n \varphi$ if and only if $M \models \Box \varphi$. A proof of the second point can be found in [2].

We then lift this notion of equivalence between sequences to an existential notion between sets of sequences, so that we can compare the different dynamic fragments of our language.

Definition 10 (Replaceability of sets). *Let $S_1, S_2 \subseteq S_D$ be two sets of sequences. The set S_1 is replaceable with the set S_2 in a model M, when, for all sequences $s_1 \in S_1$, there exists a sequence $s_2 \in S_2$ that is equivalent to s_1 in M. S_1 is replaceable with S_2 over a class of model when it is repleaceable with S_2 in all models of the class. S_1 is replaceable with S_2 (tout court) when it is repleaceble with S_2 over the class of all models.*

When comparing our dynamic operators, it is easy to see that $S_{\{\Box\}}$ (and therefore any superset of it) is not replaceable by $S_{\{\triangle\}}$ and, vice versa, that $S_{\{\triangle\}}$ (and therefore any superset of it) is not replaceable by $S_{\{\Box\}}$ and similarly for $S_{\{\Box\}}$ and $S_{\{\bigcirc\}}$, and $S_{\{\triangle\}}$ and $S_{\{\bigcirc\}}$, which implies that $S_{\{\Box,\triangle\}}$ is not replaceable with $S_{\{\bigcirc\}}$. The only interesting question is: can we replace our synchronous operator?

Theorem 2. *$S_{\{\bigcirc\}}$ is not replaceable with $S_{\{\Box,\triangle\}}$.*

Proof. We show that there is no sequence in $S_{\{\triangle,\Box\}}$ that is equivalent to the sequence $\bigcirc$ on all models. Assume, towards a contradiction that there exists a sequence $s \in S_{\{\triangle,\Box\}}$ equivalent to $\bigcirc$ in the model M given in Fig. 1. Let $n \in \mathbb{N}$ be the length of s. One of two cases must hold:

[Case 1: s starts with $\triangle$.] We can rewrite s as $\triangle s^{2:n}$. From Fig.1, we know that $M \not\models \triangle N_{ac}$, whereas $M \models \bigcirc N_{ac}$ and therefore $s \neq \triangle$. Note that M is such that any number of successive triangles reduce to one: for any φ, $M \models \triangle \varphi$ iff $M \models \triangle s' \varphi$ for every $s' \in S_{\{\triangle\}}$. Since $M \not\models \triangle N_{ac}$,

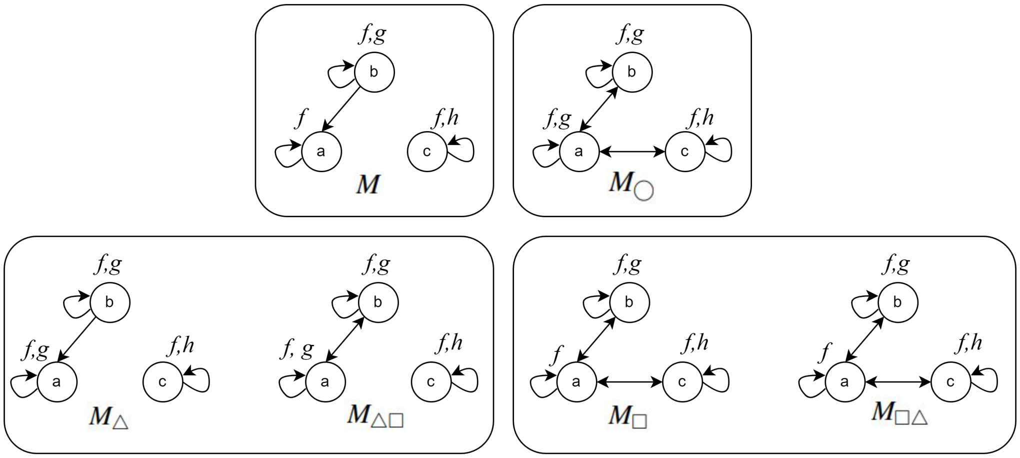


Figure 1: The figure represents the model M, and its model updates $M_{\bigcirc}, M_{\triangle}, M_{\Box}, M_{\triangle\Box}, M_{\Box\triangle}$ considered in Theorem 2. The model M is as follows: $\mathscr{A} = \{a,b,c\}$, $\mathscr{F} = \{f,g,h\}$, $\omega = \tau = \frac{1}{2}$. For each model, each agent is represented as a node, and the influence of one agent on another agent is represented as a directed arrow from the influencing node to the influenced node. Next to each node, all and only the features in $\mathscr{F}$ that an agent possess are reported.

then $M \not\models \triangle s' N_{ac}$, for every $s' \in S_{\{\triangle\}}$. Thus, $s \notin S_{\{\triangle\}}$. The sequence s must therefore contain at least one $\Box$, and can be rewritten as $s^{1:m}\Box s^{(m+2):n}$ where $s^{1:m} \in S_{\{\triangle\}}$, with $1 \leq m < n$. Given that, as mentioned above, all triangles can be reduced to one, for all $\varphi \in \mathscr{L}$, $M \models s^{1:m}\Box s^{(m+2):n}\varphi$ iff $M \models \triangle\Box s^{(m+2):n}\varphi$. As illustrated in Figure 1, $M \not\models \triangle\Box N_{ac}$. Furthermore, note that $M \models \triangle\Box\varphi$ iff $M \models \triangle\Box s'\varphi$ for all $s' \in S_{\{\triangle,\Box\}}$, i.e., $M_{\triangle\Box}$ is stable. Hence, $M \not\models \triangle\Box s' N_{ac}$ for all $s' \in S_{\{\triangle,\Box\}}$. Thus, in particular: $M \not\models \triangle\Box s^{(m+2):n} N_{ac}$. Hence, using again the fact that the number of initial triangles is irrelevant, $M \not\models s^{1:m}\Box s^{(m+2):n} N_{ac}$ which is $M \not\models sN_{ac}$. But $M \models \bigcirc N_{ac}$, and hence s is not equivalent to $\bigcirc$ in M.

[Case 2: s starts with $\Box$.] s is of the kind $\Box s^{2:n}$. As illustrated in Fig.1, $M \not\models \Box g_a$, whereas $M \models \bigcirc g_a$. Note that M is such that $M \models \Box\varphi$ iff $M \models \Box s'\varphi$ for all $s' \in S_{\{\Box,\triangle\}}$. Hence, $M \not\models \Box s' g_a$ for all $s' \in S_{\{\Box,\triangle\}}$. Thus, $M \not\models \Box s^{2:n} g_a$ which is the same as $M \not\models sg_a$. Hence, s is not equivalent to $\bigcirc$ in M.

Hence, in both cases, s is not equivalent to $\bigcirc$ in M. Contradiction. Thus, there is no sequence in $S_{\{\triangle,\Box\}}$ equivalent to $\bigcirc$, which implies that $S_{\{\bigcirc\}}$ is not replaceable by $S_{\{\triangle,\Box\}}$.

$\Box$

From the proof of Theorem 2, we know that there is no sequence in $S_{\{\Box,\triangle\}}$ that is equivalent to $\bigcirc$ in the class of *all* models. However, we will show in Proposition 1 that there are classes of models on which $\bigcirc$ does have equivalent sequences in $S_{\{\triangle,\Box\}}$. We first need to introduce the following additional abbreviations, where we already name ψ_s the formula that captures the conditions under which $\bigcirc$ is equivalent to s.

Definition 11 (Abbreviations $\psi_{\triangle}$, $\psi_{\Box}$, $\psi_{\triangle\Box}$ and $\psi_{\Box\triangle^n}$ **).**

- $\psi_{\triangle} := \bigwedge_{a,b\in\mathscr{A}}(N_{ab} \vee \neg sim^{\omega}_{ab})$
- $\psi_{\Box} := \bigwedge_{a\in\mathscr{A}}\bigwedge_{f\in F}(f_a \vee \neg f^{\tau}_{N(a)})$

- $\psi_{\triangle\Box} := \bigwedge_{a,b\in\mathscr{A}}(\neg N_{ab} \to (sim^{\omega}_{ab} \leftrightarrow \triangle sim^{\omega}_{ab}))$
- *For all* $n > 0$*:* $\psi_{\Box\triangle^n} := \bigwedge_{a\in\mathscr{A}} \bigwedge_{f\in F}(\neg f_a \to (f^{\tau}_{N(a)} \leftrightarrow \bigvee_{0\leq i\leq n-1} \Box\triangle^i f^{\tau}_{N(a)}))$

We can now show that these formulas indeed define four classes of models in which $\bigcirc$ has an equivalent sequence in $S_{\{\triangle\Box\}}$:

Proposition 1. *Let M be a model. $\bigcirc$ is equivalent on M to:*

- $\triangle$ *iff* $M \models \psi_{\triangle}$
- $\Box$ *iff* $M \models \psi_{\Box}$
- $\triangle\Box$ *iff* $M \models \psi_{\triangle\Box}$
- $\Box\triangle^n$ *iff* $M \models \psi_{\Box\triangle^n}$*, for* $n > 0$

For space reasons, the proof of Proposition 1 is provided in the Appendix.

We can now show that if $\bigcirc$ has an equivalent sequence in $S_{\{\triangle\Box\}}$ on some model, then that sequence has to be equivalent to one of those in Proposition 1 on that model. To prove this, we need the following lemmas.

Lemma 1. *Let M be a model and $s \in S_D$. If s starts with a subsequence of the form $\triangle^n$ for some $n > 0$ and s is equivalent to $\bigcirc$ on M, then $\triangle^n$ is equivalent to $\triangle$ on M.*

Proof. Consider a sequence $s \in S_D$ such that s starts with a subsequence of the kind $\triangle^n$, for some $n > 0$, and s is equivalent to $\bigcirc$ on model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau\rangle$. Two cases: either $n = 1$, or $n > 1$. If $n = 1$, the claim is trivially true since $\triangle$ is equivalent to itself. Assume now that $n > 1$. Assume also, towards a contradiction, that $\triangle^n$ is not equivalent to $\triangle$ on M. By Def. 9, it follows that $M_{\triangle} \neq M_{\triangle^n}$. By Obs. 1, it follows that $M_{\triangle}$ and $M_{\triangle^n}$ must differ on whether they satisfy some atomic proposition. By Def. 3 and Def. 4, since features can never be abandoned, we know that, for all $a \in \mathscr{A}$ for all $f \in \mathscr{F}$, if $M \models \triangle f_a$, then $M \models \triangle s' f_a$ for all $s' \in S_D$; with a similar reasoning, from Def. 3 and Def. 4, since diffusion updates do not alter the network structure, we also know that for all $a, b \in \mathscr{A}$, $M \models \triangle N_{ab}$ iff $M \models \triangle^n N_{ab}$. These observations, together with the fact that $M_{\triangle^n}$ and the model $M_{\triangle}$ must differ on the satisfaction of some atomic proposition, imply that there are $a \in \mathscr{A}$ and $f \in \mathscr{F}$ such that $M \models \triangle^n f_a$ and $M \not\models \triangle f_a$. From Obs. 2, we know that, for all $a \in \mathscr{A}$ for all $f \in \mathscr{F}$, $M \models \triangle f_a$ iff $M \models \bigcirc f_a$. Since $M \not\models \triangle f_a$, we can infer that $M \not\models \bigcirc f_a$. At the same time, from the fact that features can never be abandoned, and the fact that $M \models \triangle^n f_a$, it follows that $M \models s f_a$, since $\triangle^n$ is the initial subsequence of s. Therefore, it must be the case that both $M \not\models \bigcirc f_a$ and that $M \models s f_a$. This contradicts the initial assumption that s and $\bigcirc$ are equivalent on M. Therefore, for all $n \geq 1$, $\triangle^n$ must be equivalent to $\triangle$ on M. □

Lemma 2. *Let M be a a model and $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$. If s starts with $\triangle$ and is equivalent to $\bigcirc$ on M, then s is equivalent to $\triangle\Box$ on M.*

Proof. Consider any $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$, such that s starts with $\triangle$ and is equivalent to $\bigcirc$ on some model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau\rangle$. Assume, towards contradiction, that s is not equivalent to $\triangle\Box$ on M. By Lemma 1, we know that if s starts with a sequence of the kind $\triangle^n$, then $\triangle^n$ must be equivalent to $\triangle$ on M. For this reason, we can restrict ourselves to consider the case in which s starts with the subsequence $\triangle\Box$. Since s is not equivalent to $\triangle\Box$ on M, by Obs.1, it follows that M_s and $M_{\triangle\Box}$ must differ on whether they satisfy some atomic proposition. From Obs. 2, we know that, for all $a \in \mathscr{A}$ for all $f \in \mathscr{F}$, $M \models \triangle f_a$ iff $M \models \bigcirc f_a$. From this and the fact that the network update does not affect the features of the agent, we know that $M \models \triangle\Box f_a$ iff $M \models \bigcirc f_a$, for all $a \in \mathscr{A}$ for all $f \in \mathscr{F}$. This, combined with the fact that s is equivalent to $\bigcirc$, implies that $M \models \bigcirc f_a$ iff $M \models s f_a$, and hence, that $M \models \triangle\Box f_a$ iff $M \models s f_a$, for all $a \in \mathscr{A}$ and $f \in \mathscr{F}$. Since by Obs.1, we know that M_s and $M_{\triangle\Box}$ must differ on whether they satisfy some atomic proposition, it follows that there are a, b such that either: (i) $M \not\models sN_{ab}$ and $M \models \triangle\Box N_{ab}$ or (ii) $M \models sN_{ab}$ and $M \not\models \triangle\Box N_{ab}$. Assume that (i) is the case. If $M \models \triangle\Box N_{ab}$, then from the Def. 3 and Def. 4, it follows that $M \models \triangle\Box s' N_{ab}$ for all $s' \in S_D$, since links cannot be deleted. This fact, together with the fact that $\Box\triangle$ is a subsequence of s, implies in particular that $M \models sN_{ab}$. This contradicts the fact that $M \not\models sN_{ab}$. Therefore (ii) must be the case, i.e. $M \models sN_{ab}$ and $M \not\models \triangle\Box N_{ab}$. By the assumption that s is equivalent to $\bigcirc$ it follows that $M \models \bigcirc N_{ab}$. Now, by the reduction axioms and the fact that $M \not\models \triangle\Box N_{ab}$, we know that $M \not\models N_{ab}$ and $M \not\models \triangle sim^{\omega}_{ab}$. Since $M \not\models \triangle sim^{\omega}_{ab}$ but $M \models sN_{ab}$, there must $m \in \mathbb{N}$ smaller than the length n of the sequence s, such that $M \models s^{1:m} sim^{\omega}_{ab}$: in other words, it must be the case that at some point of the sequence s, the agents a, b have become similar. The fact that $M \models s^{1:m} sim^{\omega}_{ab}$ holds implies that there exist at least one $f \in \mathscr{F}$ and a $1 < j \leq m$ such that either $M \models s^{1:j} f_a$ and $M \not\models s^{1:i} f_a$ for all $i < j$, or $M \models s^{1:j} f_b$ and $M \not\models s^{1:i} f_b$ for all $i < j$: this simply means that in order to become similar, at least one among a or b must have acquired at least one new feature that makes them similar at some point in the update sequence expressed by s. W.l.o.g. consider the case in which it is a that has acquired a new feature, i.e. that there exist $f \in \mathscr{F}$ and a $1 < j \leq m$ such that $M \models s^{1:j} f_a$ and $M \not\models s^{1:i} f_a$ for all $i < j$. From the fact that features are never abandoned, $M \models sf_a$. Since $M \models sf_a$, and, by assumption s is equivalent to $\bigcirc$ on M, it follows that $M \models \bigcirc f_a$. Since, by assumption s starts with $\triangle$, and it is the case that $M \not\models s^{1:i} f_a$ for all $i < j$, we know that $M \not\models \triangle f_a$ (triangle is the first operator in s). It follows that both $M \models \bigcirc f_a$ and $M \not\models \triangle f_a$ are true. By Obs. 2, we know that for all $a \in \mathscr{A}$ for all $f \in \mathscr{F}$, $M \models \bigcirc f_a$ iff $M \models \triangle f_a$. Contradiction. Therefore, there is no sequence $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$, such that s starts with $\triangle$, is equivalent to $\bigcirc$ on M, and is not equivalent to $\triangle\Box$ on M.

□

Lemma 3. *Let M be a model and $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$. If s starts with $\Box$ and is equivalent to $\bigcirc$ on M, then s is equivalent to a sequence in the set $\{\Box\triangle^n : n > 0\}$.*

Proof. Let a model $M = \langle \mathcal{N}, \mathcal{V}, \omega, \tau \rangle$ be given. Assume that $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$ starts with $\Box$ and is equivalent to $\bigcirc$ on M, and that s is not equivalent to any sequence in the set $\{\Box\triangle^n : n > 0\}$ on M. From the fact that any sequence $s' \in S_{\{\Box\}}$ is equivalent to the sequence $\Box$, it follows that, if s starts with a subsequence of the kind $\Box^n$ before the first occurrence of a $\triangle$, then s is equivalent on M to a sequence that starts with a single $\Box$ followed by the subsequence of s starting at the first occurrence of a $\triangle$ and ending with the last operator of s. In other words, it is sufficient to consider the case in which s starts with a subsequence of the kind $\Box\triangle$. From this, and the fact that $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$ and $s \notin \{\Box\triangle^n : n > 0\}$, s must be such that at some point of the subsequence of s starting with the third operator of s, at least another $\Box$ occurs in it. Furthermore, at least one such $\Box$, must be such that s is not equivalent on M to the sequence s without that $\Box$. Otherwise, s would be equivalent to a sequence with no further elements of the kind $\Box$ after the initial subsequence $\Box\triangle$, and hence would be a sequence in the set $\{\Box\triangle^n : n > 0\}$. From this, it follows that there must be a subsequence $s^{1:m}$ of s, with $m \leq n$, with n the length of the sequence s, where the m-th element is a $\Box$, such that for some $a, b \in \mathscr{A}$, $M \models s^{1:m} N_{ab}$, and for all $j < m$, $M \not\models s^{1:j} N_{ab}$. Observe that $\Box$ is a subsequence of $s^{1:m}$. Therefore, $M \not\models \Box N_{ab}$. By the fact that $M \models s^{1:m} N_{ab}$, and since $s^{1:m}$ is a subsequence of s, and connections between agents cannot be abandoned by Def. 4, it follows that $M \models sN_{ab}$. By the assumption that s and $\bigcirc$ are equivalent, it follows that $M \models sN_{ab}$. By Obs. 2, we know that for all $a, b \in \mathscr{A}$, $M \models \bigcirc N_{ab}$ iff $M \models \Box N_{ab}$. This contradicts the previous claim that $M \not\models \Box N_{ab}$. Therefore there is no sequence $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\triangle\}} \cup S_{\{\Box\}})$ that starts with $\Box$, is equivalent to $\bigcirc$ on M, and is not equivalent to any sequence in the set $\{\Box\triangle^n : n > 0\}$ on M. □

We can now combine the above lemmas to prove the following theorem.

Theorem 3. *Let M be a model and $s \in S_{\{\Box,\triangle\}}$. If s is equivalent to $\bigcirc$ on M, then s is equivalent to a sequence in the set $\{\Box, \triangle, \triangle\Box\} \cup \{\Box\triangle^n : n > 0\}$ on M.*

Proof. Consider an arbitrary model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$, and an arbitrary sequence in $s \in S_{\{\Box,\triangle\}}$ equivalent to $\bigcirc$ on M. One of three cases must hold: $s \in S_{\{\Box\}}$, $s \in S_{\{\triangle\}}$, $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\Box\}} \cup S_{\{\triangle\}})$. [Case 1: $s \in S_{\{\Box\}}$.] Since the model update in Def. 4 is idempotent, any sequence $s' \in S_{\{\Box\}}$ is equivalent to the sequence $\Box$. This implies that s is equivalent to $\Box$ on M. [Case 2: $s \in S_{\{\triangle\}}$.] Then, trivially, s starts with a subsequence of the form $\triangle^n$ for some $n > 0$. From Lemma 1, we know that s is equivalent to $\triangle$ on M. [Case 3: $s \in S_{\{\triangle,\Box\}} \setminus (S_{\{\Box\}} \cup S_{\{\triangle\}})$.] If s starts with a $\triangle$, we know by Lemma 2 that s is equivalent to $\triangle\Box$ on M. If s starts with a $\Box$, by Lemma 3, we know that s is equivalent on M to a sequence in the set $\{\Box, \triangle, \triangle\Box\} \cup \{\Box\triangle^n : n > 0\}$. From this, it follows that if s is equivalent to $\bigcirc$ on some model M, then s is equivalent on M to a sequence in the set $\{\Box, \triangle, \triangle\Box\} \cup \{\Box\triangle^n : n > 0\}$.

□

Using Observation 3, Proposition 1 and Theorem 3, we can now characterise the class of models on which $\bigcirc$ can be replaced by $S_{\{\triangle,\Box\}}$.

Theorem 4. *$\bigcirc$ is replaceable by $S_{\{\triangle,\Box\}}$ on a model M iff $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$.*

Proof. Consider an arbitrary model M.

[⇒] Assume that $\bigcirc$ is replaceable with $S_{\{\triangle,\Box\}}$ on M: there exists a sequence $s \in S_{\{\triangle,\Box\}}$ equivalent to $\bigcirc$ on M. By Theorem 3, we know that s is equivalent to a sequence $s' \in \{\Box, \triangle, \triangle\Box\} \cup \{\Box\triangle^n : n > 0\}$. We distinguish four cases: (i) s is equivalent to $\triangle$ on M, (ii) s is equivalent to $\Box$ on M; (iii) s is equivalent to $\triangle\Box$ on M; (iv) s is equivalent to a sequence in $\{\Box\triangle^n : n > 0\}$ on M. Assume that (i). By the first point in Prop. 1, we know that $M \models \psi_\triangle$. From this, it follows that $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$ for any n. Assume that (ii) is the case; by the second point in Prop. 1, and the fact that s is equivalent to $\Box$ on M, we know that s is equivalent to $\bigcirc$ on M iff $M \models \psi_\Box$. From this, it follows that $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$ for any n. Assume that (iii) is the case; by the third point in Prop. 1, and the fact that s is equivalent to $\triangle\Box$ on M, we know that s is equivalent to $\bigcirc$ on M iff $M \models \psi_{\triangle\Box}$. From this, it follows that $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$ for any n. Assume that (iv) is the case, and assume that s is equivalent to a sequence of the kind $\Box\triangle^n$ on M, for arbitrary $n > 0$. By Obs. 3, we know that for all sequences $\Box\triangle^n$ with $n \geq |\mathscr{A}|$, there exist an equivalent sequence $\Box\triangle^m$ on M such that $m < |\mathscr{A}|$. We therefore consider the case in which $n < |\mathscr{A}|$: in this case, by the fourth point in Prop. 1, it then follows that s is equivalent to $\bigcirc$ on M iff $M \models \psi_{\Box\triangle^n}$. From this, it follows that $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$. Since $n > 0$ was arbitrary, this holds for all $n > 0$ in $\mathbb{N}$.

[⇐] Assume that $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$. Therefore, one of the following four cases must hold: (i) $M \models \psi_\triangle$; (ii) $M \models \psi_\Box$; (iii) $M \models \psi_{\triangle\Box}$; (iv) $M \models \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$. (i) Assume that $M \models \psi_\triangle$. By Prop. 1, we know that this holds iff $\triangle$ is equivalent to $\bigcirc$ on M. In this case, we take s to be $\triangle$. (ii) Assume that $M \models \psi_\Box$. By Prop. 1, we know that this holds iff $\Box$ is equivalent to $\bigcirc$ on M. In this case, we take s to be $\Box$. (iii) Assume that $M \models \psi_{\triangle\Box}$. By Prop. 1, we know that this holds iff $\triangle\Box$ is equivalent to $\bigcirc$ on M. In this case, we take s to be $\triangle\Box$. (iv) Assume that $M \models \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$. Therefore there is an $n < |\mathscr{A}|$, such that $M \models \psi_{\Box\triangle^n}$.By Prop. 1, we know that this holds iff $\Box\triangle^n$ is equivalent to $\bigcirc$ on M. In this case, we take s to be $\Box\triangle^n$.

Since in all cases (i)-(iv), we can find a sequence $s \in S_{\{\triangle,\Box\}}$ equivalent to $\bigcirc$ on M, we have proven that, if $M \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0 \leq n < |\mathscr{A}|} \psi_{\Box\triangle^n}$, $\bigcirc$ is replaceable with $S_{\{\triangle,\Box\}}$ on M.

□

Informally, Theorem 4 tells us that a sequence s without synchronous operators can replace a synchronous operator only under one of the following circumstances: no agent has social pressure to adopt new features (s is equivalent to $\Box$); no agent is similar to any disconnected agent (s is equivalent to $\triangle$); conforming to social pressure preserves similarity with disconnected agents (s is equivalent to $\triangle\Box$); creating new connections with similar agents does not forbid conforming to old social pressures (s is equivalent to a sequence in $\{\Box\triangle^n : n > 0\}$).

Proposition 2. *Let M be a model. If $M \models \bigwedge_{0\leq i\leq(m-1)} \bigcirc^i(\psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0\leq n<|\mathscr{A}|} \psi_{\Box\triangle^n})$, then $\bigcirc^m$ is replaceable by $S_{\{\triangle,\Box\}}$ on M.*

Proof. Assume $M \models \bigwedge_{0\leq i\leq(m-1)} \bigcirc^i(\psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0\leq n<|\mathscr{A}|} \psi_{\Box\triangle^n})$. Then, for all $i \geq 0 \leq (m-1)$: $M_{\bigcirc^i} \models \psi_\triangle \vee \psi_\Box \vee \psi_{\triangle\Box} \vee \bigvee_{0\leq n<|\mathscr{A}|} \psi_{\Box\triangle^n}$, and therefore, by Theorem 4, $\bigcirc$ is replaceable by $S_{\{\triangle,\Box\}}$ on $M_{\bigcirc^i}$. Let s_i be a sequence that replaces $\bigcirc$ on $M_{\bigcirc^i}$. Then the sequence $s_0 ... s_i ... s_{m-1}$ replaces $\bigcirc^m$ on M. □

4 Conclusion

We have introduced a logical framework containing dynamic operators to reason about asynchronous as well as synchronous threshold-induced monotonic changes in social networks. We showed that, in general, our synchronous operator cannot be replaced (Theorem 2), and that, on the models on which it can be replaced, only sequences of four specific types can replace it (Theorem 3). Finally, we characterised the class of models on which the synchronous operator can be replaced (Theorem 4).

The two most natural continuations of this work would be, first, to characterise the models on which sequences of (more than one) synchronous operators can be replaced, and, second, to study the replaceability of synchronous operators in the non-monotonic frameworks from [19, 1].

Furthermore, it would be interesting to study the replaceability of richer operators studied in epistemic/doxastic settings such as [2, 18, 16, 14, 10], for instance the network announcements in [16, 14, 10] or the message passing updates in [10]. In this direction, we could compare which models different updates can reach, as done in [3, 2]. In particular, it would be interesting to investigate what types of group knowledge are reachable by different social network dynamic updates.

Acknowledgments

Zoé Christoff acknowledges support from the project Social Networks and Democracy (VENI project number Vl.Veni.201F.032) financed by the Netherlands Organisation for Scientific Research (NWO).

References

[1] Edoardo Baccini, Zoé Christoff & Rineke Verbrugge (2022): *Opinion diffusion in similarity-driven networks*. In: *Logic and the Foundations of Game and Decision Theory (LOFT 14)*.

[2] Alexandru Baltag, Zoé Christoff, Rasmus Kræmmer Rendsvig & Sonja Smets (2018): *Dynamic epistemic logic of diffusion and prediction is social networks*. *Studia Logica* 107, doi:10.1007/s11225-018-9804-x.

[3] Alexandru Baltag & Sonja Smets (2013): *Protocols for belief merge: Reaching agreement via communication*. *Logic Journal of IGPL* 21(3), pp. 468–487, doi:10.1093/jigpal/jzs049.

[4] Yann Bramoullé, Sergio Currarini, Matthew O. Jackson, Paolo Pin & Brian W. Rogers (2012): *Homophily and long-run integration in social networks*. Journal of Economic Theory 147(5), pp. 1754–1786, doi:10.1016/j.jet.2012.05.007.

[5] Zoé Christoff & Jens Ulrik Hansen (2015): *A logic for diffusion in social networks*. Journal of Applied Logic 13(1), pp. 48–77, doi:10.1016/j.jal.2014.11.011.

[6] Zoé Christoff & Pavel Naumov (2019): *Diffusion in social networks with recalcitrant agents*. Journal of Logic and Computation 29(1), pp. 53–70, doi:10.1093/logcom/exy037.

[7] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. Synthese Library Series, Springer, doi:10.1007/978-1-4020-5839-4.

[8] Peter Dodds & Duncan J. Watts (2011): *Threshold models of social influence*. In: The Oxford Handbook of Analytical Sociology, Oxford University Press, doi:10.1093/oxfordhb/9780199215362.013.20.

[9] David Easley & Jon Kleinberg (2010): *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, USA, doi:10.1017/CBO9780511761942.

[10] Saúl Fernández González (2022): *Change in social networks: Some dynamic extensions of Social Epistemic Logic*. Journal of Logic and Computation 32(6), pp. 1212–1233, doi:10.1093/logcom/exac024.

[11] Patrick Girard, Jeremy Seligman & Fenrong Liu (2012): *General dynamic dynamic logic*. In Thomas Bolander, Torben Brauner, Silvio Ghilardi & Lawrence Moss, editors: Advances in Modal Logic, Volume 9, College Publication, pp. 239–260.

[12] Mark Granovetter (1978): *Threshold models of collective behavior*. American Journal of Sociology 83(6), pp. 1420–1443, doi:10.1086/226707.

[13] Barteld Kooi (2007): *Expressivity and completeness for public update logics via reduction axioms*. Journal of Applied Non-Classical Logics 17(2), pp. 231–253, doi:10.3166/jancl.17.231-253.

[14] Fenrong Liu, Jeremy Seligman & Patrick Girard (2014): *Logical dynamics of belief change in the community*. Synthese 191(11), pp. 2403–2431, doi:10.1007/s11229-014-0432-3.

[15] Truls Pedersen & Marija Slavkovik (2017): *Formal models of conflicting social influence*. In Bo An, Ana Bazzan, João Leite, Serena Villata & Leendert van der Torre, editors: PRIMA 2017: Principles and Practice of Multi-Agent Systems, Springer International Publishing, Cham, pp. 349–365, doi:10.1007/978-3-319-69131-2_21.

[16] Jeremy Seligman, Fenrong Liu & Patrick Girard (2011): *Logic in the community*. In Mohua Banerjee & Anil Seth, editors: Logic and Its Applications, Lecture Notes in Computer Science 6521, Springer, pp. 178–188, doi:10.1007/978-3-642-18026-2_15.

[17] Pramesh Singh, Sameet Sreenivason, Boleslaw K. Szymanski & Gyorgy Korniss (2013): *Threshold-limited spreading in social networks with multiple initiators*. Scientific Reports 3(2330), doi:10.1038/srep02330.

[18] Sonja Smets & Fernando R. Velázquez-Quesada (2017): *How to make friends: A logical approach to social group creation*. In Alexandru Baltag, Jeremy Seligman & Tomoyuki Yamada, editors: Logic, Rationality, and Interaction, Springer, Berlin, Heidelberg, pp. 377–390, doi:10.1007/978-3-662-55665-8_26.

[19] Sonja Smets & Fernando R. Velázquez-Quesada (2020): *A logical analysis of the interplay between social influence and friendship selection*. In Luís Soares Barbosa & Alexandru Baltag, editors: Dynamic Logic. New Trends and Applications, Springer International Publishing, Cham, pp. 71–87, doi:10.1007/978-3-030-38808-9_5.

[20] Sonja Smets & Fernando R. Velázquez-Quesada (2019): *A logical study of group-size based social network creation*. Journal of Logical and Algebraic Methods in Programming 106, pp. 117–140, doi:10.1016/j.jlamp.2019.05.003.

[21] Sonja Smets & Fernando R. Velázquez-Quesada (2020): *A closeness- and priority-based logical study of social network creation*. Journal of Logic, Language and Information 29(1), pp. 21–51, doi:10.1007/s10849-019-09311-5.

[22] Anthia Solaki, Zoi Terzopoulou & Bonan Zhao (2016): *Logic of closeness revision*. *ESSLLI 2016 Student Session*, p. 123.

[23] Szymon Talaga & Andrzej Nowak (2020): *Homophily as a process generating social networks: Insights from social distance attachment model*. Journal of Artificial Societies and Social Simulation 23(2), doi:10.18564/jasss.4252.

Appendix

Proof sketch of Theorem 1

Proof. The proof uses standard methods.

[*Soundness*] The soundness of the reduction axioms for the dynamic operators $\triangle, \Box, \bigcirc$ follows from the fact that they spell out the model updates in Def. 3, Def. 4 and Def. 5 respectively. The soundness of the axioms $\triangle N_{ab} \leftrightarrow N_{ab}$ and $\Box f_a \leftrightarrow f_a$ follows from the fact that the model update $M_\triangle$ does not alter the connections between agents (Def. 3), and the fact that, respectively, the model update $M_\Box$ does not alter the features of the agents (Def. 4). The soundness of the axioms $\triangle f_a \leftrightarrow f_a \vee f^\tau_{N(a)}$ can be shown as in [2]: by Def. 5, the soundness of the axiom $\bigcirc f_a \leftrightarrow f_a \vee f^\tau_{N(a)}$ is shown in the same way.

As an example, we prove the validity of $\Box N_{ab} \leftrightarrow N_{ab} \vee sim^\omega_{ab}$.

Consider a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$. $M \models \Box N_{ab}$ iff, by Def. 6, $M_\Box \models N_{ab}$ iff, by Def. 4 either $(a,b) \in \mathscr{N}$, or $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$.

This holds iff $M \models N_{ab}$ or $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$.

We now show that $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$ iff $M \models sim^\omega_{ab}$.

[$\Rightarrow$] If $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$, there exist a subset $E \subseteq \mathscr{F}$, namely the set $(\mathscr{V}(a) \cap \mathscr{V}(b)) \cup (\mathscr{F} \setminus (\mathscr{V}(a) \cup \mathscr{V}(b)))$ such that for all $f \in E$, $M \models f_a \leftrightarrow f_b$ (indeed the set $(\mathscr{V}(a) \cap \mathscr{V}(b)) \cup (\mathscr{F} \setminus (\mathscr{V}(a) \cup \mathscr{V}(b)))$ contains by definition all and only those features that either both agents have or that both do not have). This in turn implies that $M \models \bigvee_{\{E \subseteq \mathscr{F} : \frac{|E|}{|\mathscr{F}|} \geq \omega\}} \bigwedge_{f \in E}(f_a \leftrightarrow f_b)$. Thus, $M \models sim^\omega_{ab}$.

[$\Leftarrow$] Now, assume that $M \models sim^\omega_{ab}$. This holds iff $M \models \bigvee_{\{E \subseteq \mathscr{F} : \frac{|E|}{|\mathscr{F}|} \geq \omega\}} \bigwedge_{f \in E}(f_a \leftrightarrow f_b)$. This means that there exist a subset $E \subseteq \mathscr{F}$, such that $\frac{|E|}{|\mathscr{F}|} \geq \omega$, and such that for all $f \in E$, $M \models f_a \leftrightarrow f_b$, i.e. $f \in V(a)$ iff $f \in V(b)$. From this, it is clear that E must be a subset of $(\mathscr{V}(a) \cap \mathscr{V}(b)) \cup (\mathscr{F} \setminus (\mathscr{V}(a) \cup \mathscr{V}(b)))$. From this and the fact that $\frac{|E|}{|\mathscr{F}|} \geq \omega$, we can conclude that $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$.

We thus proved that $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$ iff $M \models sim^\omega_{ab}$. From above we know that $M \models \Box N_{ab}$ iff $M \models f_a$ or $\frac{|(\mathscr{V}(a)\cap\mathscr{V}(b))\cup(\mathscr{F}\setminus(\mathscr{V}(a)\cup\mathscr{V}(b)))|}{|\mathscr{F}|} \geq \omega$. We can therefore conclude that $M \models \Box N_{ab}$ iff $M \models f_a$ or $M \models sim^\omega_{ab}$.

The soundness of the axiom $\bigcirc N_{ab} \leftrightarrow N_{ab} \vee sim^\omega_{ab}$ is proven in the same way. As usual, the soundness of the distributivity of the dynamic operators over conjunction and the clauses for negation can be proven by induction on the length of formulas. Finally, validity preservation of the inference rule in Table 1 can be shown by induction on the structure of φ.

[*Completeness*] Completeness is proven in the standard way by defining a translation from the dynamic language into the static fragment of the language, see for instance [7, 13].

□

Proof of Proposition 1

Proof. [First point]

[$\Rightarrow$] Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction, that $\triangle$ is equivalent to $\bigcirc$ on M, and that $M \not\models \psi_{\triangle}$. By the definition of $\psi_{\triangle}$ in Def. 11, $M \not\models \bigwedge_{a,b \in \mathscr{A}}(N_{ab} \vee \neg sim^{\omega}_{ab})$. Since $M \not\models \bigwedge_{a,b \in \mathscr{A}}(N_{ab} \vee \neg sim^{\omega}_{ab})$, it follows that there are $a, b \in \mathscr{A}$ such that $M \models \neg N_{ab} \wedge sim^{\omega}_{ab}$. From this, and the reduction axioms for the dynamic modalities $\bigcirc$ and $\triangle$ in Table 1, respectively, it follows that $M \models \bigcirc N_{ab}$ and $M \not\models \triangle N_{ab}$. By Def. 9, this contradicts the initial assumption that $\triangle$ is equivalent to $\bigcirc$ on M.

[$\Leftarrow$] Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction, that $M \models \psi_{\triangle}$, and that it is not the case that $\triangle$ is equivalent to $\bigcirc$ on M. From the fact that $\triangle$ is not equivalent to $\bigcirc$ on M, by Obs. 1, it follows that it must be the case that $M_{\triangle}$ and $M_{\bigcirc}$ differ on whether they satisfy some atomic formula. By Obs. 2, we know that: (i) for all $a \in \mathscr{A}$, for all $f \in \mathscr{F}$, $M \models \bigcirc f_a$ iff $M \models \triangle f_a$ (a synchronous update and a diffusion update modifies in the same way the features of the agents); (ii) for all $a, b \in \mathscr{A}$, if $M \models \triangle N_{ab}$, then $M \models \bigcirc N_{ab}$. From (i) and (ii), and the fact that $M_{\triangle}$ and $M_{\bigcirc}$ differ on whether they satisfy some atomic formula, it must be the case that there are $a, b, \in \mathscr{A}$ such that $M \models \bigcirc N_{ab}$ and $M \not\models \triangle N_{ab}$. By the reduction axioms for the modality $\triangle$ in Table 1 and the fact that $M \not\models \triangle N_{ab}$, it follows that $M \not\models N_{ab}$. By the facts that $M \not\models N_{ab}$ and that $M \models \bigcirc N_{ab}$, by the reduction axioms for the modality $\bigcirc$, it must be the case that $M \models sim^{\omega}_{ab}$. We therefore know that it is both the case that $M \not\models N_{ab}$ and that $M \models sim^{\omega}_{ab}$. Therefore for some $a, b \in \mathscr{A}$, it is true that $M \models \neg N_{ab} \wedge sim^{\omega}_{ab}$, i.e. $M \models \bigvee_{a,b \in \mathscr{A}}(\neg N_{ab} \wedge sim^{\omega}_{ab})$, which implies that $M \not\models \bigwedge_{a,b \in \mathscr{A}}(N_{ab} \vee \neg sim^{\omega}_{ab})$. This means that $M \not\models \psi_{\triangle}$. This contradicts the initial assumption that $M \models \psi_{\triangle}$.

[Second point]

[$\Rightarrow$] Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction, that, $\Box$ is equivalent to $\bigcirc$ on M, and that $M \not\models \psi_{\Box}$. From the definition of $\psi_{\Box}$ in Def. 11, it follows that $M \not\models \bigwedge_{a \in \mathscr{A}} \bigwedge_{f \in F}(f_a \vee \neg f^{\tau}_{N(a)})$. Since, $M \not\models \bigwedge_{a \in \mathscr{A}} \bigwedge_{f \in F}(f_a \vee \neg f^{\tau}_{N(a)})$, it follows that there are $a \in \mathscr{A}$, and $f \in \mathscr{F}$ such that $M \models \neg f_a \wedge f^{\tau}_{N(a)}$. From this and the reduction axioms for the dynamic modalities $\bigcirc$ and $\Box$ in Table 1, it follows that $M \models \bigcirc f_a$ and $M \not\models \Box f_a$. By Def. 9, this contradicts the initial assumption that $\Box$ is equivalent to $\bigcirc$ on $\mathscr{M}$.

[$\Leftarrow$] Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction that, $M \models \psi_{\Box}$, and that it is not the case that $\Box$ and $\bigcirc$ are equivalent on M. From this and Def. 9, we know that $M_{\Box} \neq M_{\bigcirc}$. By Obs. 1, it follows that it must be the case that $M_{\Box}$ and $M_{\bigcirc}$ differ on whether they satisfy some atomic proposition. By Obs. 2, we know that: (i) for all $a, b \in \mathscr{A}$, $M \models \bigcirc N_{ab}$ iff $M \models \Box N_{ab}$ (a network update and a synchronous update modifies the network in exactly the same way); (ii) for all $a \in \mathscr{A}$, for all $f \in \mathscr{F}$, if $M \models \Box f_a$, then $M \models \bigcirc f_a$. From (i) and (ii) and the fact that $M_{\Box}$ and $M_{\bigcirc}$ differ on whether they satisfy some atomic proposition, it must the case that there are $a \in \mathscr{A}$ and $f \in \mathscr{F}$, such that $M \models \bigcirc f_a$ and $M \not\models \Box f_a$. By the fact that $M \not\models \Box f_a$ and the reduction axioms for the $\Box$ modality in Table 1, it follows that $M \not\models f_a$. By the fact that $M \not\models f_a$ and $M \models \bigcirc f_a$, by the reduction axioms for the $\bigcirc$ modality in Table 1, it follows that it must be the case that $M \models f^{\tau}_{N(a)}$. We therefore know that it is the case that: $M \not\models f_a$ and $M \models f^{\tau}_{N(a)}$. Therefore, there exist $a \in \mathscr{A}$ and $f \in \mathscr{F}$ such that $M \models \neg f_a \wedge f^{\tau}_{N(a)}$. From this, it follows that $M \models \bigvee_{a \in \mathscr{A}} \bigvee_{f \in \mathscr{F}}(\neg f_a \wedge f^{\tau}_{N(a)})$, and, therefore, $M \not\models \bigwedge_{a \in \mathscr{A}} \bigwedge_{f \in \mathscr{F}}(f_a \vee \neg f^{\tau}_{N(a)})$. This means that $M \not\models \psi_{\Box}$. We have therefore reached a contradiction with the initial assumption that $M \models \psi_{\Box}$.

[Third point]

[$\Rightarrow$] Let a model $M = \langle \mathscr{N}, \mathscr{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction, that, $\triangle\Box$ is equivalent to $\bigcirc$ on M, and that $M \not\models \psi_{\triangle\Box}$. By the definition of $\psi_{\triangle\Box}$ in Def. 11, we know that $M \not\models$

$\bigwedge_{a,b\in\mathscr{A}}(\neg N_{ab}\to(sim^{\omega}_{ab}\leftrightarrow\triangle sim^{\omega}_{ab}))$. From this, it follows that, for some $a,b\in\mathscr{A}$, $M\models\neg N_{ab}$ and $M\not\models(sim^{\omega}_{ab}\leftrightarrow\triangle sim^{\omega}_{ab})$. One of the following two must be the case: (i) $M\models sim^{\omega}_{ab}$ and $M\not\models\triangle sim^{\omega}_{ab}$; (ii) $M\not\models sim^{\omega}_{ab}$ and $M\models\triangle sim^{\omega}_{ab}$. Assume that (i) is the case: the facts that $M\models sim^{\omega}_{ab}$ and $M\not\models\triangle sim^{\omega}_{ab}$, together with the fact that $M\models\neg N_{ab}$ and the reduction axioms in Table 1, imply that $M\models\bigcirc N_{ab}$ and $M\not\models\triangle\Box N_{ab}$. This implies that $\bigcirc$ is not equivalent to $\triangle\Box$ on M, contrary to our initial assumption. It must therefore be the case that (ii) holds. Assume that (ii) is true, i.e. that $M\not\models sim^{\omega}_{ab}$ and $M\models\triangle sim^{\omega}_{ab}$. These assumptions, together with the fact that $M\not\models N_{ab}$ and the reduction axioms in Table 1, imply that $M\not\models\bigcirc N_{ab}$ and $M\models\triangle\Box N_{ab}$. By Def. 9, this contradicts the initial assumption that $\bigcirc$ and $\triangle\Box$ are equivalent on M. Since neither (i) nor (ii) are possible, we have established that if $\bigcirc$ is equivalent $\triangle\Box$ on M, then $M\models\psi_{\triangle\Box}$.

[$\Leftarrow$] Let a model $M=\langle\mathscr{N},\mathscr{V},\omega,\tau\rangle$ be given. Assume, towards a contradiction that $M\models\psi_{\triangle\Box}$, and that $\triangle\Box$ is not equivalent to $\bigcirc$ on M. From the fact that $\triangle\Box$ is not equivalent to $\bigcirc$ on M, by Def. 9, it follows that $M_{\triangle\Box}\neq M_{\bigcirc}$. By Obs. 1, we know that $M_{\triangle\Box}$ and $M_{\bigcirc}$ must differ on whether they satisfy some atomic proposition. By Obs. 2, and by Def. 4 and Def. 5, we know that for all $a\in\mathscr{A}$, and all $f\in\mathscr{F}$, $M\models\bigcirc f_a$ iff $M\models\triangle f_a$ iff $M\models\triangle\Box f_a$ (informally, this simply mean that, since a network update does not affect the agent's features, and a synchronous update and a diffusion update change the features in the same way, the features of the agents after one synchronous update are the same as those after one diffusion update followed by a subsequent network update). From this and the fact that $M_{\triangle\Box}$ and $M_{\bigcirc}$ must differ on whether they satisfy some atomic proposition, it follows that one of the following two cases must hold: (i) there are a,b such that $M\models\triangle\Box N_{ab}$, and $M\not\models\bigcirc N_{ab}$; (ii) there are a,b such that $M\not\models\triangle\Box N_{ab}$, and $M\models\bigcirc N_{ab}$.

Assume that (i) is the case, i.e. there are a,b such that $M\models\triangle\Box N_{ab}$, and $M\not\models\bigcirc N_{ab}$; from the fact that $M\not\models\bigcirc N_{ab}$ and the reduction axioms for $\bigcirc$ in Table 1, we know that $M\not\models N_{ab}\vee sim^{\omega}_{ab}$, which implies that $M\models\neg N_{ab}$ and $M\models\neg sim^{\omega}_{ab}$; furthermore, from the fact that $M\models\triangle\Box N_{ab}$, and that $M\not\models N_{ab}$, we know that $M\models\triangle sim^{\omega}_{ab}$. If we put these together, we know that $M\models\neg N_{ab}$, $M\models\neg sim^{\omega}_{ab}$, and $M\models\triangle sim^{\omega}_{ab}$ at the same time: this implies that $M\models\bigvee_{a,b\in\mathscr{A}}(\neg N_{ab}\wedge\neg sim^{\omega}_{ab}\wedge\triangle sim^{\omega}_{ab})$. From this, it follows that $M\not\models\bigwedge_{a,b\in\mathscr{A}}(\neg N_{ab}\to(sim^{\omega}_{ab}\leftrightarrow\triangle sim^{\omega}_{ab}))$: by Def. 11, it follows that $M\not\models\psi_{\triangle\Box}$, which contradicts our initial assumption that $M\models\psi_{\triangle\Box}$.

Since (i) is not possible, it must be the case that (ii) holds, i.e. there are a,b such that $M\not\models\triangle\Box N_{ab}$, and $M\models\bigcirc N_{ab}$. From the fact that $M\not\models\triangle\Box N_{ab}$ and the reduction axioms for $\triangle$ and $\Box$, we know that $M\not\models\triangle sim^{\omega}_{ab}$ and $M\not\models N_{ab}$. From the fact that $M\models\bigcirc N_{ab}$ and $M\not\models N_{ab}$, by the reduction axioms for $\bigcirc$, we know that $M\models sim^{\omega}_{ab}$. Summarising the facts above, we therefore know that $M\models\neg N_{ab}$, $M\models sim^{\omega}_{ab}$ and $M\models\neg\triangle sim^{\omega}_{ab}$: this means that $M\models\bigvee_{a,b\in\mathscr{A}}(\neg N_{ab}\wedge sim^{\omega}_{ab}\wedge\neg\triangle sim^{\omega}_{ab})$. This implies the fact that $M\not\models\bigwedge_{a,b\in\mathscr{A}}(\neg N_{ab}\to(sim^{\omega}_{ab}\leftrightarrow\triangle sim^{\omega}_{ab}))$. By Def. 11, it follows that $M\not\models\psi_{\triangle\Box}$, which contradicts our initial assumption that $M\models\psi_{\triangle\Box}$.

Since neither (i) nor (ii) are possible, we have established that if $M\models\psi_{\triangle\Box}$, then $\bigcirc$ must be equivalent to $\triangle\Box$ on M.

[Fourth point]

[$\Rightarrow$] Let a model $M=\langle\mathscr{N},\mathscr{V},\omega,\tau\rangle$ be given. Assume, towards a contradiction, that, for some arbitrary $n>0$, $\Box\triangle^{n}$ is equivalent to $\bigcirc$ on $\mathscr{M}$ and that $M\not\models\psi_{\Box\triangle^{n}}$. From the fact that $M\not\models\psi_{\Box\triangle^{n}}$, by Def. 11, it follows that $M\not\models\bigwedge_{a\in\mathscr{A}}\bigwedge_{f\in F}(\neg f_a\to(f^{\tau}_{N(a)}\leftrightarrow\bigvee_{0\leq i\leq n-1}\Box\triangle^{i}f^{\tau}_{N(a)}))$. From this, it follows that there exist $a\in\mathscr{A}$ and $f\in\mathscr{F}$ such that $M\models\neg f_a$ and $M\not\models(f^{\tau}_{N(a)}\leftrightarrow\bigvee_{0\leq i\leq n-1}\Box\triangle^{i}f^{\tau}_{N(a)})$. One of the following two must hold: (i) $M\models f^{\tau}_{N(a)}$ and $M\not\models\bigvee_{0\leq i\leq n-1}\Box\triangle^{i}f^{\tau}_{N(a)}$, or (ii) $M\not\models f^{\tau}_{N(a)}$ and $M\models\bigvee_{0\leq i\leq n-1}\Box\triangle^{i}f^{\tau}_{N(a)}$.

Assume that (i) is the case and thus that $M\models f^{\tau}_{N(a)}$ and $M\not\models\bigvee_{0\leq i\leq n-1}\Box\triangle^{i}f^{\tau}_{N(a)}$. Since $M\models f^{\tau}_{N(a)}$, by the reduction axiom for $\bigcirc$, we know that $M \models \bigcirc f_a$. At the same time, since we know that $M \not\models \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}$, we know that for all $0 \leq i \leq n-1$ $M \not\models \Box\triangle^i f^\tau_{N(a)}$: this simply means that after a network update, there is no sequence of diffusion update after which the agent a has social conformity pressure to adopt feature f. From this and the fact that $M \not\models f_a$, by the reduction axioms of for the dynamic modalities, we know that $M \not\models \Box\triangle f_a$. Therefore, it is both the case that $M \models \bigcirc f_a$, and $M \not\models \Box\triangle^n f_a$, which, by Def. 9, contradicts the initial assumption that $\bigcirc$ is equivalent to $\Box\triangle^n$ on M.

Since (i) is not possible, it must be the case that (ii) holds, i.e. it is the case that $M \not\models f^\tau_{N(a)}$ and $M \models \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}$. From the fact that $M \not\models f^\tau_{N(a)}$, and the fact that $M \not\models f_a$, by the reduction axioms for $\bigcirc$, it follows that $M \not\models \bigcirc f_a$. From the fact that $M \not\models f^\tau_{N(a)}$ and $M \models \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}$, it follows that there is an $i \leq (n-1)$ such that $M \models \Box\triangle^i f^\tau_{N(a)}$ (this means that, after a network update, at some point of a sequence of further diffusion update, agent a has pressure to adopt feature f). From this and the reduction axioms for the dynamic modality $\triangle$, it follows that $M \models \Box\triangle^{i+1} f_a$. Since $i \leq n-1$, and by the fact that features cannot be abandoned, it follows that $M \models \Box\triangle^n f_a$. Thus, it is both true that $M \not\models \bigcirc f_a$, and $M \models \Box\triangle^n f_a$. By Def. 9, this contradicts the initial assumption that $\Box\triangle^n$ is equivalent to $\bigcirc$ on M.

Since neither (i) nor (ii) are possible, we have established that if $\bigcirc$ is equivalent $\Box\triangle^n$ on M, then $M \models \psi_{\Box\triangle^n}$.

[$\Leftarrow$] Let a model $M = \langle \mathcal{N}, \mathcal{V}, \omega, \tau \rangle$ be given. Assume, towards a contradiction, that, for some $n > 0$, $M \models \psi_{\Box\triangle^n}$ and that $\bigcirc$ is not equivalent to $\Box\triangle^n$ in M. From the fact that $\bigcirc$ is not equivalent to $\Box\triangle^n$ in M, by Def. 9, it follows that $M_{\Box\triangle^n} \neq M_{\bigcirc}$. Thus, by Obs. 1, we know that $M_{\Box\triangle^n}$ and $M_{\bigcirc}$ must differ on whether they satisfy some atomic proposition. By Obs. 2 and by Def. 4 and Def. 5, we know that, for all $a, b \in \mathcal{A}$, $M \models \bigcirc N_{ab}$ iff $M \models \Box N_{ab}$ iff $M \models \Box\triangle^n N_{ab}$. Informally, this follows from the fact that, since a diffusion update does not affect the agent's features, and a synchronous update and a network update change the network structure in the same way, the connections between the agents after one synchronous update are the same as those obtained after one network update followed by multiple subsequent diffusion update. From this and the fact that $M_{\Box\triangle^n}$ and $M_{\bigcirc}$ must differ on whether they satisfy some atomic proposition, there must exist $a \in \mathcal{A}$ and $f \in \mathcal{F}$, s.t. it is not the case that $M \models \Box\triangle^n f_a$ iff $M \models \bigcirc f_a$. Therefore, either one of the following cases must hold: (i) $M \models \Box\triangle^n f_a$ and $M \not\models \bigcirc f_a$, or (ii) $M \not\models \Box\triangle^n f_a$ and $M \models \bigcirc f_a$.

Assume that (i): $M \models \Box\triangle^n f_a$ and $M \not\models \bigcirc f_a$. By the fact $M \not\models \bigcirc f_a$ and the reduction axioms for $\bigcirc$, we know that $M \not\models f^\tau_{N(a)}$ and $M \not\models f_a$. From the fact that $M \models \Box\triangle^n f_a$, we know that there exist an $0 \leq i \leq n-1$ such that $M \models \Box\triangle^i f^\tau_{N(a)}$: this simply means that, at some point, after a network-update and potentially after subsequent diffusion updates, a has pressure to adopt f. From this, it follows that $M \models \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)})$. Therefore, from the above we know that $M \models \neg f_a$, $M \models \neg f^\tau_{N(a)}$ and $M \models \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)})$. This imply that $M \models \bigvee_{a \in \mathcal{A}} \bigvee_{f \in \mathcal{F}} (\neg f_a \wedge \neg f^\tau_{N(a)} \wedge \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}))$. Therefore $M \not\models \bigwedge_{a \in \mathcal{A}} \bigwedge_{f \in F} (\neg f_a \rightarrow (f^\tau_{N(a)} \leftrightarrow \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}))$. Thus, by Def. 11, we know that $M \not\models \psi_{\Box\triangle^n}$, which contradicts our initial assumption that $M \models \psi_{\Box\triangle^n}$.

Since (i) cannot be the case, it must be the case that (ii): $M \not\models \Box\triangle^n f_a$ and $M \models \bigcirc f_a$.

From the fact $M \not\models \Box\triangle^n f_a$, by the reduction axioms for the dynamic modalities, we know two things: $M \not\models f_a$, and there does not exist $0 \leq i \leq n-1$, such that $M \models \Box\triangle^i f^\tau_{N(a)}$ (this simply means that at no point after a network update and subsequent diffusion updates a has pressure to adopt f; indeed, if this was the case then a would at some point adopt f, and will never abandon it). From the fact that $M \not\models f_a$ and that $M \models \bigcirc f_a$, by the reduction axiom for $\bigcirc$, we know that $M \models f^\tau_{N(a)}$. Summarising the above we know that: $M \models \neg f_a$, $M \models f^\tau_{N(a)}$ and $M \models \neg \bigvee_{0 \leq i \leq n-1} \Box\triangle^i f^\tau_{N(a)}$. Thus, $M \models \bigvee_{a \in \mathcal{A}} \bigvee_{f \in \mathcal{F}} (\neg f_a \wedge f^\tau_{N(a)} \wedge$

$\neg\bigvee_{0\leq i\leq n-1}\Box\triangle^i f^{\tau}_{N(a)})$. Thus, $M \not\models \bigwedge_{a\in\mathscr{A}}\bigwedge_{f\in F}(\neg f_a \to (f^{\tau}_{N(a)} \leftrightarrow \bigvee_{0\leq i\leq n-1}\Box\triangle^i f^{\tau}_{N(a)}))$. By Def. 11, it follows that $M \not\models \psi_{\Box\triangle^n}$, contrary to the initial assumption that $M \models \psi_{\Box\triangle^n}$.

Since neither (i) nor (ii) are possible, we have established that if $M \models \psi_{\Box\triangle^n}$, then $\bigcirc$ is equivalent to $\Box\triangle^n$ on M.

□

Simple Axioms for Local Properties

Philippe Balbiani
Institut de Recherche en Informatique de Toulouse
philippe.balbiani@irit.fr

Wiebe van der Hoek
University of Liverpool
wiebe@liverpool.ac.uk

Louwe B. Kuijer
University of Liverpool
lbkuijer@liverpool.ac.uk

Correspondence theory allows us to create sound and complete axiomatizations for modal logic on frames with certain properties. For example, if we restrict ourselves to transitive frames we should add the axiom $\Box\varphi \to \Box\Box\varphi$ which, among other things, can be interpreted as positive introspection. One limitation of this technique is that the frame property and the axiom are assumed to hold globally, i.e., the relation is transitive throughout the frame, and the agent's knowledge satisfies positive introspection in every world.

In a modal logic with local properties, we can reason about properties that are not global. So, for example, transitivity might hold only in certain parts of the model and, as a result, the agent's knowledge might satisfy positive introspection in some worlds but not in others. Van Ditmarsch et al. [9] introduced sound and complete axiomatizations for modal logics with certain local properties. Unfortunately, those axiomatizations are rather complex. Here, we introduce far simpler axiomatizations for a wide range of local properties.

1 Introduction

Modal logic is a formalism used throughout computer science, artificial intelligence and philosophy to represent various concepts including knowledge or belief (*epistemic/doxastic logic*, e.g., [13, 10, 6]), time (*temporal logic*, e.g., [14, 3]), necessity (*alethic logic*, e.g., [5, 4]), obligation (*deontic logic*, e.g., [1, 12]) and more. The main operator of modal logic is usually denoted $\Box$, or, in a multi-agent setting, $\Box_a$ where a is an agent index. A formula $\Box\varphi$ can then be read, depending on the context, as "φ is known", "φ is believed", "φ is true in every possible future", "φ is necessarily true" or "φ is obligatory".

The most commonly used kind of semantics for modal logic uses *relational models*, also known as *Kripke models*. Such a model M consists of three parts, $M = (W,R,V)$, where W is a set of worlds or states, R is an accessibility relation on those worlds (or, in the multi-agent case, a set of relations), and V is a valuation that determines on which worlds an atomic formula is true. A formula $\Box\varphi$ then holds on a world w_1 if φ is true on every w_2 such that $(w_1,w_2) \in R$.

Depending on the specific concept being represented, some additional assumptions are typically made, however. Let us consider two such assumptions as examples. Firstly, in epistemic logic knowledge is generally assumed to be truthful, represented by the axiom $\Box_a\varphi \to \varphi$, which can be read as "if agent a knows φ, then φ is true". Secondly, in alethic logic it is usually assumed that everything that is necessary is necessarily necessary, represented by the axiom $\Box\varphi \to \Box\Box\varphi$.

Each such axiom *corresponds*, in the precise technical sense of correspondence theory [15, 2], to a constraint on models. The axiom $\Box_a\varphi \to \varphi$ corresponds to a's accessibility relation being reflexive, $\neg\Box\bot$ corresponds to the accessibility relation being serial and $\Box\varphi \to \Box\Box\varphi$ corresponds to the relation being transitive.

Some non-standard assumptions can be handled in the same way. Suppose, for example, that agent b is smarter and more well-informed than agent a, and that, as a result, everything known by a is also

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 82–92, doi:10.4204/EPTCS.379.9

known by b. This can be represented by an axiom $\Box_a\varphi \to \Box_b\varphi$ and the corresponding property that the relation $R(b)$ for agent b is a subset of the relation $R(a)$ for agent a.

One limitation of this approach, to use axioms and their corresponding model conditions, is that the properties they represent are inherently global. If we assume the axiom $\Box_a\varphi \to \varphi$, then a's accessibility relation should be reflexive throughout the model. As such, not only is a's knowledge truthful, that truthfulness must be common knowledge. Similarly, if we take $\Box_a\varphi \to \Box_b\varphi$ as an axiom then b knows at least as much as a in every world.

One solution to this issue was introduced in [7, 8, 9] as *local properties*. A local property is, as the name implies, a "local" variant of a property such as reflexivity, transitivity, or one relation being a subset of another. A world w is locally reflexive if w is a successor of itself, it is locally transitive if whenever w_1 is a successor of w and w_2 is a successor of w_1, w_2 is also a successor of w, and b locally knows more than a in w if $\{w' \mid (w,w') \in R(b)\} \subseteq \{w' \mid (w,w') \in R(a)\}$. On the syntax side, for a given local property a special symbol ϑ is then introduced, that holds on a world if and only if that world satisfies the local property.

The main outcome of [8, 9] was a method to create sound and complete axiomatizations for logic with any local properties that satisfy a technical condition known as r-persistence. Unfortunately, while these axiomatizations are indeed sound and complete, they use a relatively complex introduction rule for local properties (see Section 3 for details). Here, we introduce axiomatizations for many local properties that instead use an introduction axiom, which is also considerably simpler than the rule from [9]. The downside of our approach is that it is less general than the one from [9], although it does include all commonly discussed local properties and we expect that it can be generalized.

The structure of this paper is as follows. First, in Section 2 we introduce the necessary definitions and some background results. Then, in Section 3, we describe the axiomatizations from [9]. Following that, in Section 4, we introduce a sound and complete axiomatization for one local property using a case study. Finally, in Section 5, we discuss axiomatizations for other local properties.

2 Preliminaries

2.1 Modal logic

Before we define the specifics of local properties, it is useful to first define the usual semantics for modal logic. Let $\mathscr{A}$ be a finite set of agents and $\mathscr{P}$ a countable set of propositional atoms.

Definition 1. A *model* M is a triple $M = (W,R,V)$ where W is a set of worlds, $R : \mathscr{A} \to 2^{W\times W}$ assigns to each agent an accessibility relation and $V : \mathscr{P} \to 2^W$ is a valuation that assigns to each atom a subset of W.

We also write wR_aw' for $(w,w') \in R(a)$ and denote the set $\{w' \mid wR_aw'\}$ by $R_a(w)$.

A pointed model is a pair M,w where $M = (W,R,V)$ and $w \in W$.

A frame is a pair $F = (W,R)$, where $R : \mathscr{A} \to 2^{W\times W}$.

Definition 2. The modal formulas are given by the following normal form:

$$\varphi ::= p \mid \top \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box_a\varphi,$$

where $p \in \mathscr{P}$ and $a \in \mathscr{A}$.

The operators $\wedge, \to$ and $\leftrightarrow$ are defined as abbreviations in the usual way. Furthermore, we use $\Diamond_a$ as an abbreviation for $\neg\Box_a\neg$. When $|\mathscr{A}| = 1$ we omit the index a, writing $\Box$ and $\Diamond$ for $\Box_a$ and $\Diamond_a$.

The semantics are as usual.

Definition 3. Let M,w be a pointed model. The satisfaction relation $\models$ is given recursively by

$$\begin{array}{lll} M,w \models p & \Leftrightarrow & w \in V(p) \\ M,w \models \neg\varphi & \Leftrightarrow & M,w \not\models \varphi \\ M,w \models \varphi_1 \vee \varphi_2 & \Leftrightarrow & M,w \models \varphi_1 \text{ or } M,w \models \varphi_2 \\ M,w \models \Box_a\varphi & \Leftrightarrow & \forall w' \in R_a(w) : M,w' \models \varphi \end{array}$$

If $M,w \models \varphi$ for every $w \in W$, we write $M \models \varphi$. If $M \models \varphi$ for every M we say that φ is *valid* and write $\models \varphi$.

If $F = (W,R)$ is a frame, we say that $F,w \models \varphi$ if $(W,R,V),w \models \varphi$ for every V.

The usual proof system **K** is as follows.

Definition 4. The proof system **K** is given by the following axioms and rules.

T	any substitution instance of a validity of propositional logic
K	$\Box_a(\varphi \to \psi) \to (\Box_a\varphi \to \Box_a\psi)$
Nec	From ψ, infer $\Box_a\psi$
MP	From $\varphi \to \psi$ and φ, infer ψ.

It is well known that **K** is sound and strongly complete for modal logic. All axiomatizations that we discuss in this paper extend **K**.

Finally, we should define bisimulations, as these will become important later.

Definition 5. Let $M_1 = (W_1,R_1,V_1)$ and $M_2 = (W_2,R_2,V_2)$ be models. A *bisimulation* is a relation $\sim \subseteq W_1 \times W_2$ such that the following three properties hold.

Atomic agreement If $w_1 \sim w_2$ then for all $p \in \mathscr{P}$, $w_1 \in V(p)$ iff $w_2 \in V(p)$.

Forth If $w_1 \sim w_2$ and $(w_1,w_1') \in R_1(a)$, then there is a $w_2' \in W_2$ such that $(w_2,w_2') \in R_2(a)$ and $w_1' \sim w_2'$.

Back If $w_1 \sim w_2$ and $(w_2,w_2') \in R_2(a)$, then there is a $w_1' \in W_1$ such that $(w_1,w_1') \in R_1(a)$ and $w_1' \sim w_2'$.

Famously, modal logic is the bisimulation-invariant fragment of first-order logic [2]. In particular, modal logic is invariant under bisimulation, so if $w_1 \sim w_2$ then the two worlds satisfy the same modal formulas.

2.2 Local properties

Now we can define local properties, and their interaction with modal logic.

Definition 6. Let X be a set of first order variables. A *local property* is a formula with one free variable in the first order language given by

$$\Theta ::= (x,y) \in R(a) \mid x = y \mid \neg\Theta \mid (\Theta \vee \Theta) \mid \forall x \Theta,$$

where $a \in \mathscr{A}$ and $x,y \in X$.

We follow [9] in this definition, by allowing only relational predicates, $(x,y) \in R(a)$, and not valuation predicates $x \in V(p)$. This restriction is not necessary for our analysis, but it seems harmless, given that we do not know of any interesting local properties that depend on such valuation predicates. A local property is a first order formula, so it can be evaluated on models, where we take the set W of worlds to be the domain of quantification. Furthermore, because we restrict ourselves to relational predicates, the valuation does not affect the value of any local property, so we can also evaluate a local property on the frame underlying a model instead of on the model itself.

We will focus primarily on local properties that are not invariant under bisimulation. This is not because formulas that are invariant under bisimulation are necessarily uninteresting, but because any such property is equivalent to a formula of modal logic. Consider, for example, the property of local seriality, $\Theta_{ser} = \exists x(w,x) \in R$. We have $\Theta_{ser}(w)$ if and only if $M,w \models \Diamond\top$, so if we wish to reason about local seriality we need not introduce a special symbol ϑ_{ser} but can instead reason about $\Diamond\top$.

If we wish to reason about a given local property Θ in modal logic, we add an extra symbol ϑ to the language of modal logic. On a technical level, ϑ can be seen as either a nullary modality or a designated propositional atom. In order to emphasize its special role, we denote its extension separately in our notation for models. Specifically, we write $M = (W,R,\Delta,V)$, where $\Delta \subseteq W$ and $M,w \models \vartheta$ iff $w \in \Delta$. If we look at multiple local properties at a time, each is represented by a different symbol, i.e., $M = (W,R,\Delta_1,\cdots,\Delta_k,V)$, with $M,w \models \vartheta_i$ iff $w \in \Delta_i$. Because Δ is simply a subset of W, there is no inherent guarantee that the atom ϑ is in any way related to the property Θ. In order to force this connection, we look at models that are in *harmony*.

Definition 7. A model $M = (W,R,\Delta,V)$ is in *Θ-ϑ-harmony* if for every $w \in W$, we have $w \in \Delta$ if and only if $\Theta(w)$.

When Θ and ϑ are clear from context we omit reference to them, and simply say that a model is in harmony. Our goal, and the goal of [8, 9], is to find axiomatizations that are sound and complete for the class of models that are in harmony.

Finally, we need the notion of a modal formula locally defining a first-order property.

Definition 8. Let φ be a schema of modal logic and Θ a local property. The formula φ *locally defines* Θ if for every frame $F = (W,R)$ and every $w \in W$, we have $F,w \models \varphi$ if and only if $\Theta(w)$.

Generally, the schema that we will use to locally define a property is the same one that globally corresponds to that property. So, for example, the property of (local) transitivity is locally defined by $\Box\psi \to \Box\Box\psi$, and (local) reflexivity is locally defined by $\Box\psi \to \psi$. If $\varphi(p_1,\dots,p_n)$ is a modal formula constructed from the propositional atoms $p_1,\dots,p_n$ then for all modal formulas $\chi_1,\dots,\chi_n$, $\varphi(\chi_1,\dots,\chi_n)$ will denote the modal formula obtained from $\varphi(p_1,\dots,p_n)$ by respectively replacing the occurrences of $p_1,\dots,p_n$ by $\chi_1,\dots,\chi_n$.

3 Axiomatizations using an inference rule

In order for the method from [9] to apply for a local property Θ, we require two things. Firstly, we must have a schema φ that locally defines Θ. Secondly, φ should be locally r-persistent. Defining this property requires a lot of further definitions, so for practical reasons we will not give such a definition here and instead refer the reader to [11] for details. We will note, however, that the axioms corresponding to the usual properties (transitivity, reflexivity, etc.) are locally r-persistent.

If these conditions are satisfied, the axiomatizations from [9] work by adding an *elimination axiom* and an *introduction rule* for ϑ to the proof system **K**. The elimination axiom Eϑ is quite simple.

$$\text{E}\vartheta \qquad \vartheta \to \varphi,$$

where φ is the schema that locally defines Θ. Consider, for example, local Euclidicity, formally given by $\Theta_{Euc} = \forall x,y(((w,x) \in R(a) \wedge (w,y) \in R(a)) \to (x,y) \in R(a))$. This property is locally defined by the schema $\Diamond_a\psi \to \Box_a\Diamond_a\psi$, i.e., negative introspection. Hence the elimination axiom Eϑ_{Euc} is given by

$\vartheta_{Euc} \to (\Diamond_a \psi \to \Box_a \Diamond_a \psi)$. This, of course, is exactly as desired, since it means that in every world where ϑ_{Euc} holds, the agent a is capable of negative introspection.

The introduction rule is more complex, and before we can formally state it we first need to introduce *pseudo-modalities*. Let $s = x_1, \cdots, x_n$ be a (possibly empty) finite sequence of pairs $x_i = (a_i, \chi_i)$, where $a_i \in \mathscr{A}$ and χ_i is a formula of modal logic. The pseudo-modality $[s]$ is then an abbreviation, where $[s]\psi$ stands for $\chi_1 \to \Box_{a_1}(\chi_2 \to \Box_{a_2}(\chi_3 \to \Box_{a_3}(\cdots \Box_{a_{n-1}}(\chi_n \to \Box_{a_n}\psi))$.

Let k be the number of different schematic variables in φ. Then the introduction rule Iϑ is as follows.

$$\text{I}\vartheta \qquad \text{from } [s]\varphi(p_1, \cdots, p_k), \text{ infer } [s]\vartheta,$$

where $[s]$ is a pseudo-modality and $p_1, \cdots, p_k$ are fresh atoms.

Consider again the property Θ_{Euc}. By taking s to be the empty sequence, the rule Iϑ_{Euc} allows us to infer ϑ_{Euc} from $\Diamond_a p \to \Box_a \Diamond_a p$. By taking a non-empty sequence s, for example $s = (\psi_1, b), (\psi_2, a), (\psi_3, c)$, it also allows us to infer

$$\psi_1 \to \Box_b(\psi_2 \to \Box_a(\psi_3 \to \Box_c \vartheta_{Euc}))$$

from

$$\psi_1 \to \Box_b(\psi_2 \to \Box_a(\psi_3 \to \Box_c(\Diamond_a p \to \Box_a \Diamond_a p))),$$

as long as p does not occur in ψ_1, ψ_2 and ψ_3.

The main result from [9] is that if $\Theta_1, \cdots, \Theta_n$ are local properties that satisfy the required conditions, then the proof system $\mathbf{K} + \text{E}\vartheta_1 + \text{I}\vartheta_1 + \cdots + \text{E}\vartheta_n + \text{I}\vartheta_n$ is sound and complete for the class of models that are in Θ_i-ϑ_i-harmony for all $1 \leq i \leq n$.

4 Case study: transitivity

Here, we introduce alternative axiomatizations for local properties. We use the same elimination axiom, but instead of an introduction rule we use an introduction axiom. Furthermore, our introduction axiom is syntactically simpler than the rules from [9]. Our axiomatization is based on the observation that most of the local properties that seem interesting (including, but not limited to, the examples from [7, 8, 9]) are not merely *not preserved* under bisimilarity, but *anti-preserved*, in the sense that apart from some trivial exceptions,[1] for every M and w such that $\Theta(w)$, there are M', w' such that $w \sim w'$ and $\neg\Theta(w')$.

In this section we use the property of local transitivity to show how we can leverage this anti-preservation in order to obtain a sound and strongly complete axiomatization. So take

$$\Theta_{tr} = \forall x, y(((w,x) \in R \land (x,y) \in R) \to (w,y) \in R).$$

Our elimination axiom for ϑ_{tr} is the same as that of [9].

$$\text{E}\vartheta_{tr} \qquad \vartheta_{tr} \to (\Box\psi \to \Box\Box\psi).$$

Furthermore, it is clear that when the antecedent of the implication in Θ_{tr} is not satisfied, Θ_{tr} trivially holds. So if $\forall x, y \neg((w,x) \in R \land (x,y) \in R)$, then $\Theta_{tr}(w)$. Unlike Θ_{tr} itself, this latter property is invariant under bisimulation. In fact, it is equivalent to the modal formula $\Box\Box\bot$. The following *trivial introduction axiom* is therefore sound.

$$\text{TI}\vartheta_{tr} \qquad \Box\Box\bot \to \vartheta_{tr}$$

[1] These exceptions apply when a world has no successors. For example, if $\forall x \neg(w,x) \in R$, then $w \sim w'$ implies that w' is locally Euclidean, as well as locally transitive, locally dense, et cetera.

We will show that $\mathrm{E}\vartheta_{tr}$ and $\mathrm{TI}\vartheta_{tr}$ suffice, so $\mathbf{K}+\mathrm{E}\vartheta_{tr}+\mathrm{TI}\vartheta_{tr}$ is sound and complete for the class of models that are in Θ_{tr}-ϑ_{tr}-harmony. For an informal overview of why we do not need any further introduction axioms, note that for every pointed model M,w, if $M,w \not\models \Box\Box\bot$ and $\Theta_{tr}(w)$ then there is a bisimilar model M',w' such that $\neg\Theta_{tr}(w')$. In particular, the tree unravelling of M will have that property.

We will show that, as a consequence, no introduction axiom $\chi \to \vartheta_{tr}$ can be sound unless χ contains ϑ_{tr} (which would make it a rather useless introduction axiom) or χ implies $\Box\Box\bot$ (which would render the axiom superfluous in the presence of $\mathrm{TI}\vartheta_{tr}$). So suppose towards a contradiction that there is some modal formula χ such that (i) ϑ_{tr} does not occur in χ, (ii) χ does not imply $\Box\Box\bot$ and (iii) $\chi \to \vartheta_{tr}$ is valid on models that are in harmony. Then there is some M,w such that $M,w \not\models \Box\Box\bot$ and $M,w \models \chi$. Take M',w' to be the bisimilar model such that $\neg\Theta_{tr}(w')$. Because χ is a modal formula, it is invariant under bisimulation. Hence $M',w' \models \chi$. Furthermore, because χ does not contain ϑ_{tr}, we can choose M' to be in harmony. Since $\neg\Theta_{tr}(w')$ holds this implies that $M',w' \not\models \vartheta_{tr}$, which contradicts the soundness of $\chi \to \vartheta_{tr}$ on models in harmony.

From this contradiction, it follows that no axiom $\chi \to \vartheta_{tr}$ can be sound unless χ contains the symbol ϑ_{tr} or χ implies $\Box\Box\bot$. This does not fully suffice to prove that $\mathbf{K}+\mathrm{E}\vartheta_{tr}+\mathrm{TI}\vartheta_{tr}$ is complete for models in harmony, but it does show why we should not expect to need any further introduction axioms.

In order to turn this informal proof sketch into a full proof, let us first introduce one further auxiliary definition.

Definition 9. A model M is Θ_{tr}-ϑ_{tr}-*nice* if

1. for every w, if $M,w \models \vartheta_{tr}$ then $\Theta_{tr}(w)$ and
2. if $M,w \models \Box\Box\bot$ then $M,w \models \vartheta_{tr}$.

Every model that is in harmony is also nice, but not necessarily vice versa. It is also quite easy to see that $\mathbf{K}+\mathrm{E}\vartheta_{tr}+\mathrm{TI}\vartheta_{tr}$ is sound and strongly complete for the class of nice models, since $\mathrm{E}\vartheta_{tr}$ directly corresponds to the first condition of niceness and $\mathrm{TI}\vartheta_{tr}$ corresponds to the second condition. This completeness can be proven using the standard canonical model construction. We will show that for every nice model, there is a bisimilar harmonious model.

Definition 10. Let $M=(W,R,\Delta,V)$ be a nice model. We define a sequence of models as follows:

- $M_0=(W',R_0,\Delta',V')$ is the tree unravelling of M, i.e.,
  - W' is the set of finite sequences $w'=(w_1,\cdots,w_n)$ such that (i) $w_j \in W$ for all $1\leq j\leq n$ and (ii) $(w_j,w_j+1)\in R$ for all $1\leq j<n$,
  - $(w'_1,w'_2)\in R_0$ if and only if $w'_1=(w_1,\cdots,w_n)$ and $w'_2=(w_1,\cdots,w_n,w_{n+1})$ for some $w_1,\cdots,$ $w_{n+1}\in W$,
  - $(w_1,\cdots,w_n)\in\Delta'$ if and only if $w_n\in\Delta$,
  - $(w_1,\cdots,w_n)\in V'(p)$ if and only if $w_n\in V(p)$.
- $M_{i+i}=(W',R_{i+1},\Delta',V')$, where $(w'_1,w'_2)\in R_{i+1}$ if and only if
  - $(w'_1,w'_2)\in R_i$ or
  - $w'_1\in\Delta'$ and there is a w'_3 such that $(w'_1,w'_3)\in R_i$ and $(w'_3,w'_2)\in R_i$.
- $M_\infty=(W',R_\infty,\Delta',V')$, where $R_\infty=\bigcup_{i\in\mathbb{N}}R_i$.

Because M_0 is a tree model, no world is locally transitive (unless none of its successors have a successor). We then add additional edges to the relation, but only where needed to make every $w'\in\Delta'$ locally transitive. As a consequence, we will be able to show that there is a total bisimulation between M and M_∞, and that M_∞ is in harmony.

Note that all models M_i, for $i \in \mathbb{N} \cup \{\infty\}$ use the same set W' of worlds, and that this world was obtained by taking the tree unravelling of M. As such, every $w' = (w_1, \cdots, w_n) \in W'$ has a unique original, namely w_n, in W.

Lemma 1. *Take any $x', y' \in W'$ and let $x, y \in W$ be the originals of x' and y', respectively. If $(x', y') \in R_i$ for some $i \in \mathbb{N} \cup \{\infty\}$, then $(x, y) \in R$.*

Proof. By induction on i. As base case, suppose that $i = 0$. Then the lemma follows immediately from the fact that M_0 is the tree unravelling of M. Suppose then, as induction hypothesis, that $i \in \mathbb{N}_{>0}$ and that the lemma holds for all $i' < i$.

Take any $(x', y') \in R_i$. If already $(x', y') \in R_{i-1}$, then by the induction hypothesis we have $(x, y) \in R$, as was to be shown.

If $(x', y') \in R_i \setminus R_{i-1}$, then the arrow must have been added in stage i, meaning that $x' \in \Delta'$ and there is some z' such that $(x', z') \in R_{i-1}$ and $(z', y') \in R_{i-1}$. Then, by the induction hypothesis, $(x, z) \in R$ and $(z, y) \in R$, where z is the original of z'. Furthermore, because M_0 is the tree unravelling of M, and all M_i use the same set Δ', it follows from $x' \in \Delta'$ that $x \in \Delta$.

From the fact that M is nice, it then follows that $\Theta_{tr}(w)$, so w is locally transitive. Because $(x, z) \in R$ and $(z, y) \in R$, we then have $(x, y) \in R$, as was to be shown. This completes the induction step for $i \in \mathbb{N}_{>0}$. Finally, if $(x', y') \in R_\infty$, then $(x', y') \in R_j$ for some $j \in \mathbb{N}$, and therefore $(x, y) \in R$. □

Now, we can take the important step of showing bisimilarity.

Lemma 2. *Let $\sim \subseteq W \times W'$ be the relation such that $x \sim y'$ if and only if x is the original of y'. Then $\sim$ is a bisimulation between M and M_i, for every $i \in \mathbb{N} \cup \{\infty\}$.*

Proof. We show that $\sim$ satisfies the three conditions of bisimulation. Take any $w_1 \in W$ and $w_1' \in W'$ such that $w_1 \sim w_1'$.

Atoms Because M_0 is the tree unraveling of M, and M_i and M' only differ from M_0 by the addition of edges, we have $w_1 \in V(p)$ iff $w_1' \in V'(p)$.

Forth Take any w_2 such that $(w_1, w_2) \in R$. As M_0 is the tree unraveling of M, there is some $w_2' \in W'$ such that $(w_1', w_2') \in R_0$ and w_2 is the original of w_2' (and therefore $w_2 \sim w_2'$). Furthermore, $(w_1', w_2') \in R_0$ implies $(w_1', w_2') \in R_i$ for every $i \in \mathbb{N} \cup \{\infty\}$.

Back Take any w_2' such that $(w_1', w_2') \in R_i$ for some $i \in \mathbb{N} \cup \{\infty\}$. Then by the preceding lemma we have $(w_1, w_2) \in R$, where w_2 is the original of w_2'. Hence we also have $w_2 \sim w_2'$.

□

Left to show is that M_∞ is Θ_{tr}-ϑ_{tr}-harmonious.

Lemma 3. *If $x' \in \Delta'$, $(x', y') \in R_\infty$ and $(y', z') \in R_\infty$ then $(x', z') \in R_\infty$.*

Proof. If $(x', y') \in R_\infty$ and $(y', z') \in R_\infty$, then there is some $i \in \mathbb{N}$ such that $(x', y') \in R_i$ and $(y', z') \in R_i$. As $x' \in \Delta'$, this implies that $(x', z') \in R_{i+1}$, and therefore $(x', z') \in R_\infty$. □

Lemma 4. *If $x' \notin \Delta'$, then there are y', z' such that $(x', y') \in R_\infty$ and $(y', z') \in R_\infty$ while $(x', z') \notin R_\infty$.*

Proof. From $x' \notin \Delta'$ it follows that $x \notin \Delta$. Because M is nice, this implies that there are y, z such that $(x, y) \in R$ and $(y, z) \in R$. Then there are $y', z' \in W'$ such that $(x', y') \in R_0$ and $(y', z') \in R_0$. As M_0 is a tree model, we have $(x', z') \notin R_0$. Furthermore, because $x' \notin \Delta'$, no edge from x' to z' is added in any model M_i. Hence $(x', z') \notin R_\infty$. □

Corollary 1. *The model $M' = M_\infty$ is Θ_{tr}-ϑ_{tr}-harmonious.*

It now follows quite easily that $\mathbf{K} + \mathrm{E}\vartheta_{tr} + \mathrm{TI}\vartheta_{tr}$ is sound and complete.

Corollary 2. *The axiomatization $\boldsymbol{K} + E\vartheta_{tr} + TI\vartheta_{tr}$ is sound and strongly complete for the class of models that are Θ_{tr}-ϑ_{tr}-harmonious.*

Proof. The axiomatization is sound and strongly complete for Θ_{tr}-ϑ_{tr}-nice models. Furthermore, every nice model M can be transformed into a bisimilar model M_∞ that is in harmony. It follows that the axiomatization is sound and strongly complete for harmonious models. □

5 Axioms for local properties

Many other local properties can be given a simple axiomatization in a way similar to local transitivity. Here, we consider local reflexivity, Euclidicity, symmetry, superset, density and functionality. This includes all of the examples from [9]. Note that each of these properties requires the existence of one or more edges. In the case of reflexivity this requirement is unconditional; in order for w to be locally reflexive there must be an edge $(w,w) \in R$. For the other properties, the requirement for the edge to exist is conditional on the existence of one or more other edges. For example, local Euclidicity requires that if $(w_1,w_2) \in R$ and $(w_1,w_3) \in R$ then $(w_2,w_3) \in R$, while local symmetry requires that if $(w_1,w_2) \in R$ then $(w_2,w_1) \in R$. It is important to note that for each of these properties it is an edge between two specific worlds that needs to exist. In the example of local Euclidicity, there must be an edge from the world w_2 to the world w_3. Any other edge, even if it is from w_2 to a world that is bisimilar to w_3, does not suffice.

An introduction axiom takes the form $\chi \to \vartheta$, where χ is a modal formula that does not contain ϑ. In order for this axiom to be sound for the models where Θ and ϑ are in harmony, it therefore has to be the case that whenever χ is true, either (i) the condition under which the local property Θ requires an edge to exist is false or (ii) the specific edge(s) required by Θ do exist.

There is no modal formula that implies the existence of any specific edge, so χ cannot guarantee that option (ii) is the case.[2] There are modal formulas that imply the *non*existence of a particular edge, however. For example, if $\Box\bot$ is true in w_1 then there are no edges that start in w_1 and therefore, in particular, it is not the case that $(w_1,w_2) \in R$ and $(w_1,w_3) \in R$. For most local properties, it is therefore possible for χ to guarantee that condition (i) holds, and therefore for the introduction axiom to be sound, if we take $\chi = \Box\bot$ or $\chi = \Box\Box\bot$. The unconditional nature of reflexivity makes it an exception in this regard, $\chi \to \vartheta_{ref}$ is sound only if χ is unsatisfiable.

Of course the above reasoning only tells us when an introduction axiom of this form is sound, completeness remains to be shown. As in the preceding section, showing completeness is done by proving that every nice model can be transformed into a bisimilar harmonious one.

Let us now consider each of the aforementioned local properties in some more detail. Local reflexivity, given by $\Theta_{ref} = (w,w) \in R$, has an even simpler axiomatization than local transitivity. The elimination axiom is as one would expect, $\mathrm{E}\vartheta_{ref}$ is $\vartheta_{ref} \to (\Box\varphi \to \varphi)$. But unlike ϑ_{tr}, we do not require any introduction axiom for ϑ_{ref}. This is because, unlike any of the other properties that we consider here, Θ_{ref} does not contain an implication of which the antecedent can be false. As a result, there is no modal formula φ such that $M,w \models \varphi$ trivially implies that $\Theta_{ref}(w)$.

Nice models with respect to Θ_{ref} and ϑ_{ref} are therefore simply the ones where $\Theta_{ref}(w)$ holds for all $w \in \Delta$. As in the case of transitivity, we can then unravel any nice model M into a tree model M_0, and

[2] There are modal formulas, such as $\Diamond\top$, that guarantee the existence of an edge, but these formulas don't guarantee the existence of a *specific* edge.

then modify that tree model into a bisimilar model that is in harmony. It follows that $\mathbf{K} + \mathrm{E}\vartheta_{ref}$ is sound and strongly complete for the class of models that are in Θ_{ref}-ϑ_{ref}-harmony.

Local Euclidity, symmetry and density are given by

$$\Theta_{Euc} = \forall x, y(((w,x) \in R \wedge (w,y) \in R) \rightarrow (x,y) \in R),$$

$$\Theta_{sym} = \forall x((w,x) \in R \rightarrow (x,w) \in R)$$

and

$$\Theta_{dense} = \forall x \exists y((w,x) \in R \rightarrow ((w,y) \in R) \wedge (y,x) \in R),$$

respectively. The elimination axioms for these properties are as one would expect:

$\mathrm{E}\vartheta_{Euc}$	$\vartheta_{Euc} \rightarrow (\Diamond\varphi \rightarrow \Box\Diamond\varphi)$
$\mathrm{E}\vartheta_{sym}$	$\vartheta_{sym} \rightarrow (\varphi \rightarrow \Box\Diamond\varphi)$
$\mathrm{E}\vartheta_{dense}$	$\vartheta_{dense} \rightarrow (\Diamond\varphi \rightarrow \Diamond\Diamond\varphi)$

For each of these properties, the antecedent of the implication is trivially false when there is no x such that $(w,x) \in R$. Hence the trivial introduction axioms are simply $\Box\bot \rightarrow \vartheta_{Euc}$, $\Box\bot \rightarrow \vartheta_{sym}$ and $\Box\bot \rightarrow \vartheta_{dense}$. Completeness is shown as before, by turning nice models into tree models and then those tree models into harmonious models. It follows that $\mathbf{K} + \mathrm{E}\vartheta_{Euc} + \mathrm{TI}\vartheta_{Euc}$ is sound and strongly complete for models that are in Θ_{Euc}-ϑ_{Euc}-harmony, $\mathbf{K} + \mathrm{E}\vartheta_{sym} + \mathrm{TI}\vartheta_{sym}$ is sound and complete for models in Θ_{sym}-ϑ_{sym}-harmony and $\mathbf{K} + \mathrm{E}\vartheta_{dense} + \mathrm{TI}\vartheta_{dense}$ is sound and complete for models in Θ_{dense}-ϑ_{dense}-harmony.

The local property of $R(a)$ being a superset of $R(b)$ is given by $\Theta_{sup(a,b)} = \forall x((w,x) \in R(b) \rightarrow (w,x) \in R(a))$. Recall that this property can be read as b knowing at least as much as a. The elimination axiom $\mathrm{E}\vartheta_{sup(a,b)}$ is therefore, unsurprisingly, $\vartheta_{sup(a,b)} \rightarrow (\Box_a\varphi \rightarrow \Box_b\varphi)$. The antecedent of $\Theta_{sup(a,b)}$ is false when there is no x such that $(w,x) \in R(b)$, so when $M, w \models \Box_b\bot$. The introduction axiom $\mathrm{TI}\vartheta_{sup(a,b)}$ is therefore $\Box_b\bot \rightarrow \vartheta_{sup(a,b)}$. The presence of the different agents a and b does not interfere in our procedure of unraveling M and turning that unraveling into a harmonious model, so $\mathbf{K} + \mathrm{E}\vartheta_{sup(a,b)} + \mathrm{TI}\vartheta_{sup(a,b)}$ is sound and complete for models in $\Theta_{sup(a,b)}$-$\vartheta_{sup(a,b)}$-harmony.

Finally, let us consider local functionality, $\Theta_{func} = \forall x, y((R(w,x) \wedge R(w,y)) \rightarrow x = y)$. The elimination axiom $\mathrm{E}\vartheta_{func}$ is $\vartheta_{func} \rightarrow ((\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi))$, and the trivial elimination axiom $\mathrm{TI}\vartheta_{func}$ by $\Box\bot \rightarrow \vartheta_{func}$. As with the other properties we can then unravel any nice model M into a tree model M_0, and turn M_0 into a harmonious model M_∞. The only difference is that in this case it does not suffice to merely add edges in order to obtain M_{i+1} from M_i. Instead, for every world $w' \notin \Delta'$, if there is only one x' such that $(w', x') \in R_i$ then we need to create a copy of the sub-tree rooted in x', the root of which we will call x''. Then we add this tree to M_{i+1}, as well as an edge $(w, x'') \in R_{i+1}$. The result of this procedure will be a harmonious model M_∞ such that M_∞, w' is bisimilar to M, w. It follows that $\mathbf{K} + \mathrm{E}\vartheta_{func} + \mathrm{TI}\vartheta_{func}$ is sound and complete for models in Θ_{func}-ϑ_{func}-harmony.

In summary, we have the following elimination and introduction axioms.

Property	Elimination axiom	Introduction axiom
Θ_{tr}	$\vartheta_{tr} \rightarrow (\Box\varphi \rightarrow \Box\Box\varphi)$	$\Box\Box\bot \rightarrow \vartheta_{tr}$
Θ_{ref}	$\vartheta_{ref} \rightarrow (\Box\varphi \rightarrow \varphi)$	-
Θ_{Euc}	$\vartheta_{Euc} \rightarrow (\Diamond\varphi \rightarrow \Box\Diamond\varphi)$	$\Box\bot \rightarrow \vartheta_{Euc}$
Θ_{sym}	$\vartheta_{sym} \rightarrow (\varphi \rightarrow \Box\Diamond\varphi)$	$\Box\bot \rightarrow \vartheta_{sym}$
$\Theta_{sup(a,b)}$	$\vartheta_{sup(a,b)} \rightarrow (\Box_a\varphi \rightarrow \Box_b\varphi)$	$\Box_b\bot \rightarrow \vartheta_{sup(a,b)}$
Θ_{dense}	$\vartheta_{dense} \rightarrow (\Diamond\varphi \rightarrow \Diamond\Diamond\varphi)$	$\Box\bot \rightarrow \vartheta_{dense}$
Θ_{func}	$\vartheta_{func} \rightarrow ((\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi))$	$\Box\bot \rightarrow \vartheta_{func}$

Unfortunately, while this method does yield simple and elegant axiomatizations for these local properties, it is not as general as the method from [9]. For one thing, we do not have a general technique that allows for the automatic generation of axiomatizations for large classes of local properties. Furthermore, unlike [9] adding the axioms for multiple local properties does not necessarily yield a sound and complete axiomatization for the class of models that are in harmony for each property.

For example, $\mathbf{K} + \mathrm{E}\vartheta_{Euc} + \mathrm{TI}\vartheta_{Euc} + \mathrm{E}\vartheta_{ref}$ is not sound and complete for the models that are in both Θ_{Euc}-ϑ_{Euc} and Θ_{ref}-ϑ_{ref}-harmony. This is because if $\Theta_{Euc}(w_1)$ and $(w_1,w_2) \in R$ then we also have $\Theta_{ref}(w_2)$. Hence we would need the additional introduction axiom $\mathrm{I}\vartheta_{Euc}\vartheta_{ref}$, namely $\vartheta_{Euc} \to \Box\vartheta_{ref}$. With that additional axiom, we can once again use the same construction to turn every nice model into a harmonious model, so $\mathbf{K} + \mathrm{E}\vartheta_{Euc} + \mathrm{TI}\vartheta_{Euc} + \mathrm{E}\vartheta_{ref} + \mathrm{I}\vartheta_{Euc}\vartheta_{ref}$ is sound and complete for the models in Θ_{Euc}-ϑ_{Euc} and Θ_{ref}-ϑ_{ref}-harmony.

Such additional axioms are not always needed. For example, $\mathbf{K} + \mathrm{E}\vartheta_{tr} + \mathrm{TI}\vartheta_{tr} + \mathrm{E}\vartheta_{ref}$ is sound and complete for the models in Θ_{tr}-ϑ_{tr} and Θ_{ref}-ϑ_{ref}-harmony. Furthermore, where needed the extra axioms seems relatively easy to find. Yet we do not currently have a systematic way of determining whether an additional axiom is required and, if so, which axiom.

6 Conclusion

We have presented an alternative way to create axiomatizations for local properties. In contrast to the existing axiomatizations from [9], our approach uses introduction *axioms*, as opposed to introduction *rules*. Furthermore, our axioms are simpler and, in our opinion, more elegant than the rules from [9].

The price we pay for this simplicity is that we do not, as of yet, have a general way to create axiomatizations for further local properties, or for combinations of multiple local properties. Given the extremely strong similarities between the completeness proofs for the axiomatizations that we considered, we expect that some kind of generalization is possible, but we have not found it yet.

As such, the main direction for further work would be to find such generalizations.

References

[1] Lennart Åqvist (1987): *Introduction to Deontic Logic and the Theory of Normative Systems*. Biblopolis.

[2] Johan van Benthem (1976): *Modal correspondence theory*. Ph.D. thesis, University of Amsterdam.

[3] Johan van Benthem (1995): *Temporal Logic*. In Dov M. Gabbay, C. J. Hogger & J. A. Robinson, editors: *Handbook of logic in artificial intelligence and logic programming*, pp. 241–350.

[4] Patrick Blackburn, Johan van Benthem & Frank Wolter, editors (2007): *Handbook of Modal Logic*. Elsevier.

[5] Rudolf Carnap (1947): *Meaning and Necessity*. University of Chicago Press.

[6] Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek & Barteld Kooi, editors (2015): *Handbook of epistemic logic*. College Publications.

[7] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2009): *Knowing more – From Global to Local Correspondence*. In: *Proceedings of IJCAI-09*, pp. 955–960.

[8] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2011): *Reasoning about local properties in modal logic*. In: *Proceedings of AAMAS-11*, pp. 711–718.

[9] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2012): *Local properties in modal logic*. *Artificial Intelligence* 187–188, pp. 133–155, doi:10.1016/j.artint.2012.04.007.

[10] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Vardi (1995): *Reasoning about knowledge*. MIT press.

[11] Valentin Goranko (1998): *Axiomatizations with context rules of inference in modal logic*. Studia Logica 61, pp. 179–197, doi:10.1023/A:1005021313747.

[12] John Horty (2001): *Agency and Deontic Logic*. Oxford University Press, doi:10.1093/0195134613.001.0001.

[13] John-Jules Ch. Meyer & Wiebe van der Hoek (1995): *Epistemic logic for AI and computer science*. Cambridge University Press, doi:10.1017/CBO9780511569852.

[14] Amir Pnueli (1977): *The temporal logic of programs*. In: *18th annual symposium on foundations of computer science*, pp. 44–67.

[15] Henrik Sahlqvist (1975): *Completeness and Correspondence in First and Second Order Semantics for Modal Logics*. In: *Proceedings of the Third Scandinavian Logic Symposium*, pp. 110–143, doi:10.1016/S0049-237X(08)70728-6.

Implicit Knowledge in Unawareness Structures
- Extended Abstract -*

Gaia Belardinelli
Center for Information and Bubble Studies
University of Copenhagen
belardinelli@hum.ku.dk

Burkhard C. Schipper
Department of Economics
University of California, Davis
bcschipper@ucdavis.edu

Awareness structures by Fagin and Halpern (1988) (FH) feature a syntactic awareness correspondence and accessibility relations modeling implicit knowledge. They are a flexible model of unawareness, and best interpreted from a outside modeler's perspective. Unawareness structures by Heifetz, Meier, and Schipper (2006, 2008) (HMS) model awareness by a lattice of state-spaces and explicit knowledge via a possibility correspondence. They can be interpreted as providing the subjective views of agents. Open questions include (1) how implicit knowledge can be defined in HMS structures, and (2) in which way FH structures can be extended to model the agents' subjective views. In this paper, we address (1) by showing how to derive implicit knowledge from explicit knowledge in HMS models. We also introduce a variant of HMS models that instead of explicit knowledge, takes implicit knowledge and awareness as primitives. Further, we address (2) by introducing a category of FH models that are modally equivalent relative to sublanguages and can be interpreted as agents' subjective views depending on their awareness. These constructions allow us to show an equivalence between HMS and FH models. As a corollary, we obtain soundness and completeness of HMS models with respect to the Logic of Propositional Awareness, based on a language featuring *both* implicit and explicit knowledge.

Keywords: Unawareness, awareness, implicit knowledge, explicit knowledge.

JEL-Classifications: D83, C70.

1 Introduction

Models of unawareness are of interest in various disciplines, most notably in computer science, economics, game theory, decision theory, and philosophy. The seminal contribution in computer science and philosophy are awareness structures by Fagin and Halpern (1988) (henceforth, *FH models*) who extended Kripke structures with a syntactic awareness correspondence in order to feature notions of implicit knowledge, explicit knowledge, and awareness. In economics, Heifetz, Meier, and Schipper (2006, 2008) introduced unawareness structures (henceforth, *HMS models*) that consist of a lattice of state spaces featuring a notion of explicit knowledge and awareness. Like Kripke structures, HMS models can be constructed canonically and three different sound and complete axiomatizations have been presented (Halpern and Rêgo, 2008, Heifetz, Meier, and Schipper, 2008).[1] There have already been a

[*] Dedicated to Joe Halpern on the occasion of his 70th birthday. A complete version of the paper with proofs of the results is available from https://faculty.econ.ucdavis.edu/faculty/schipper/implicit.pdf. We thank three anonymous reviewers, Hans van Ditmarsch, participants in the CSLI Workshop 2022, Stanford University, and the 2022 Zoom Mini-Workshop on Epistemic Logic with Unawareness for helpful comments. Gaia gratefully acknowledges funding from the Carlsberg Foundation. Burkhard gratefully acknowledges financial support via ARO Contract W911NF2210282.

[1] For other approaches and an overview, see Schipper (2015).

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 93–112, doi:10.4204/EPTCS.379.10

number of applications to game theory, decision theory, mechanism design and contracting, financial markets, electoral campaigning, conflict resolution, social network formation, business strategy and entrepreneurship etc.

The different modeling approaches may be seen as reflecting the different foci of the fields. HMS models in economics are very much motivated by game theory and its applications. The main underlying idea is that *explicit* rational reasoning of players is what drives their behavior. Hence, the model features only explicit knowledge (without the detour via implicit knowledge) and awareness, and it can be interpreted as encompassing the subjective views of players. Moreover, the syntax-free frame lends itself seamlessly to the existing body of work in decision theory and game theory. The focus on behavioral implications also explains why the model is built on strong properties of knowledge such as (positive) introspection and factivity: this allows for the identification of the behavioral implications of unawareness per se without confounding it with mistakes in information processing. Somewhat differently, FH models were motivated more generally by the study of the logical non-omniscience problem in computer science and philosophy (see e.g., Hintikka, 1975, Levesque, 1984, Lakemeyer, 1986, Stalnaker, 1991). They represent awareness via syntactic awareness correspondences, which for each agent assigns a set of formulas to each state. This approach to awareness modeling offers a great deal of flexibility, because the set of formulas an agent is aware of may be arbitrary, thereby allowing potentially for the representation of different notions of awareness.[2] However, because their semantics is not syntax-free, their applications to decision or game theory require more effort. This is because in decision theory and game theory and applications thereof, the primitives are typically not described syntactically. Moreover, FH models are best interpreted as a tool used by an outside modeler (like a systems designer of a multi-agent distributed system) for two reasons: First, the primitive notion of knowledge is *implicit* knowledge while explicit knowledge is derived from implicit knowledge and awareness. Implicit knowledge is not necessarily something that the agent herself can consciously reason about. Second, we cannot think of FH models as models that the agents themselves use for analyzing their epistemic universe unless they are aware of everything.[3] This becomes relevant in interactive settings when we are interested in the players' interactive perceptions of the epistemic universe. Despite the differences in modeling approaches, Halpern and Rêgo (2008) and Belardinelli and Rendsvig (2022) formalize in which ways HMS models are equivalent to FH models in terms of explicit knowledge and awareness. However, as the discussion above makes clear, there remain open questions: First, can implicit knowledge be captured also in HMS models and how would this notion of implicit knowledge be related to implicit knowledge in FH models? Second, can we extend FH models so as to interpret them from the agents' subjective point of views? These questions will be answered in this paper.

By showing how to derive implicit knowledge from explicit knowledge in HMS models, we provide a way to understand implicit knowledge in terms of explicit knowledge. We are aware of only a few other approaches deriving implicit knowledge from explicit knowledge. Using neighborhood models without a notion of awareness, Velázquez-Quesada (2013) takes explicit knowledge as the primitive and then derives implicit knowledge as closure of logical consequences of explicit knowledge. Implicit knowledge is then understood as knowledge that the agent ideally could deduce from her explicit knowledge. Lorini (2020) takes an agent's belief base as explicit knowledge and derives implicit knowledge as what is deducible from an agent's belief base and common background information. While we find these two notions of implicit knowledge easy to interpret, it is not the notion of implicit knowledge that is captured

[2] See Fagin and Halpern (1988, pp. 54-55) and Fagin et al. (1995, Chapter 9.5) for discussions.

[3] The view of the systems designer is expressed eloquently by Fagin, Halpern, and Vardi (1986): "The notion of knowledge is *external*. A process cannot answer questions based on its knowledge with respect to this notion of knowledge. Rather, this is a notion meant to be used by the system designer reasoning about the system. ... (I)t does seem to capture the type of intuitive reasoning that goes on by system designers."

by propositionally determined FH models, namely models where an agent is aware of a formula if and only if she is aware of all atomic formula that appear in it. We also introduce a variant of HMS models in which we take the notion of implicit knowledge and a semantic awareness function as the primitive and then derive explicit knowledge. This shows that in HMS models, implicit and explicit knowledge are "interdefinable", at least in the sense that taking any of the two as primitive is sufficient to recover the other, so one may choose either one as primitive.

We are also interested in an extension of FH models that allows us to interpret them as subjective views of agents. Starting from an FH model, we show how to form a category of FH models with FH models as objects and surjective bounded morphisms as morphisms. Each category of FH models is literally a category of FH models that are modally equivalent relative to sublanguages formed by taking subsets of atomic formulae. The category of FH models forms a complete lattice ordered by subset inclusion on sets of atomic formulae or ordered by surjective bounded morphisms or ordered by modal equivalence relative to sublanguages. Each FH model in the lattice can be interpreted as the subjective model of an agent with that awareness level given by the subset of atomic formulae for which the FH model is defined. The construction now suggests transformations between FH and HMS models. The transformation from FH to HMS models relies on a transformation of each FH category into an implicit knowledge-based HMS model mentioned above. This implicit knowledge-based HMS model can be complemented with explicit knowledge and thus yields a HMS model. The transformation from HMS to FH models simply relies on pruning away the subjective spaces, only maintaining the upmost space in the lattice, as well as deriving the syntactic awareness correspondences from possibility correspondences and the lattice of spaces in HMS models. For each model class, its transformation into a model of the other class satisfy the same formulas from a language for explicit, implicit knowledge, and awareness. It shows how the model classes and implicit knowledge notions relate to each other. As a corollary of soundness and completeness of the Logic of Propositional Awareness w.r.t. the class of FH models, the results allow us to derive soundness and completeness for the class of HMS models with implicit knowledge, complementing earlier axiomatizations of HMS models that made use of explicit knowledge (and awareness) only (Halpern and Rêgo, 2008, Heifetz, Meier, and Schipper, 2008).

2 HMS Models

HMS models are multi-agent models for awareness originally proposed by Heifetz, Meier, and Schipper (2006, 2008). For lack of space, we refer the reader for explanations and intuitions to that papers. Throughout the paper, we let At be a non-empty set of atomic formulas.

Definition 1 *A* HMS model $\mathsf{M} = \langle I, \{S_\Phi\}_{\Phi\subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi\subseteq\Phi\subseteq\mathsf{At}}, (\Pi_i)_{i\in I}, v\rangle$ *for* At *consists of*

- *a non-empty set of individuals I,*
- *a non-empty collection of non-empty disjoint state spaces $\{S_\Phi\}_{\Phi\subseteq\mathsf{At}}$ indexed by subsets of atomic formulae $\Phi \subseteq \mathsf{At}$. Note that $\{S_\Phi\}_{\Phi\subseteq\mathsf{At}}$ forms a complete lattice by subset inclusion on the set of atomic formulae $\Phi\subseteq\mathsf{At}$. Denote the set of all states in spaces of the lattice by $\Omega := \bigcup_{\Phi\subseteq\mathsf{At}} S_\Phi$.*
- *Projections $(r^\Phi_\Psi)_{\Psi\subseteq\Phi\subseteq\mathsf{At}}$ such that, for any $\Phi,\Psi\subseteq\mathsf{At}$ with $\Psi\subseteq\Phi$, $r^\Phi_\Psi : S_\Phi \longrightarrow S_\Psi$ is surjective, for any $\Phi\subseteq\mathsf{At}$, r^Φ_Φ is the identity on S_Φ, and for any $\Phi,\Psi,\Upsilon\subseteq\mathsf{At}$, $\Upsilon\subseteq\Psi\subseteq\Phi$, $r^\Phi_\Upsilon = r^\Psi_\Upsilon \circ r^\Phi_\Psi$. For any $\Phi\subseteq\mathsf{At}$ and $D\subseteq S_\Phi$, denote by $D^\uparrow := \bigcup_{\Phi\subseteq\Psi\subseteq\mathsf{At}} (r^\Psi_\Phi)^{-1}(D)$. An* event *$E\subseteq\Omega$ is defined by a $\Phi\subseteq\mathsf{At}$ and a subset $D\subseteq S_\Phi$ such that $E := D^\uparrow$. We call S_Φ the base-space of the event E and D the base of the event E. We denote by Σ the set of events.*

- *A possibility correspondence* $\Pi_i : \Omega \longrightarrow 2^{\Omega} \setminus \{\emptyset\}$ *for each individual* $i \in I$.
- *A valuation function* $v : \mathsf{At} \longrightarrow \Sigma$.

Not every subset of the union of spaces is an event. Intuitively, $D^{\uparrow}$ collects all the "extensions of descriptions in D to at least as expressive vocabularies" (Heifetz, Meier, and Schipper, 2006). Events are well defined by the above definition except for the case of vacuous events. Since the empty set is a subset of any space, we have as many vacuous events as there are state-spaces. These vacuous events are distinguished by their base-space, so we denote them by $\emptyset^{S_\Phi}$ for $\Phi \subseteq \mathsf{At}$. At a first glance, the existence of many vacuous events may be puzzling. Note that vacuous events essentially represent contradictions, i.e., propositions that cannot hold at any state. Contradictions are formed with atomic formulae. Thus, they can be more or less complicated depending on the expressiveness of the underlying language describing states and hence are represented by different vacuous events.

We define Boolean operations on events. Negation of events is defined as follows: Let E be an event with base D and base-space S_Φ. Then $\neg E := (S_\Phi \setminus D)^{\uparrow}$. Conjunction of events is defined by intersection of events. Disjunction of events is defined by the DeMorgan Law using negation and conjunction as just defined. Note that in HMS models we typically have $E \cup \neg E \subsetneqq \Omega$ unless the base-space of E is $S_\emptyset$, the meet of the lattice of spaces. Also, disjunction of two events is typically a proper subset of the union of these events unless both events have the same base-space, since it is just the union of the events in spaces in which both events are "expressible".

The following notation will be convenient: Sometimes we denote by S_ω the state-space that contains state ω. For any $D \subseteq S_\Phi$, we denote by D_{S_Ψ} the projection of D to S_Ψ for $\Psi \subseteq \Phi \subseteq \mathsf{At}$. We simplify notation further and let for any $D \subseteq S_\Phi$ and $\Psi \subseteq \Phi \subseteq \mathsf{At}$, D_Ψ be the projection of D to S_Ψ. Similarly, for any $D \subseteq S_\Psi$, we denote by D^Φ the "elaboration" of D in the space S_Φ with $\Psi \subseteq \Phi$, i.e., $D^\Phi := (r^\Phi_\Psi)^{-1}(D)$. The same applies to states, i.e., ω_Ψ is the projection of $\omega \in S_\Phi$ to S_Ψ with $\Psi \subseteq \Phi$. Finally, for any event $E \in \Sigma$, we denote by $S(E)$ the base-space of E. We say that an event E is expressible in S_Φ if $S(E) \preceq S_\Phi$.

As usual in epistemic structures used in game theory and economics, information is modeled by a possibility correspondence instead of an accessibility relation. In HMS models, having mappings rather than relations adds extra convenience in that we can easily compose projections with possibility correspondences and vice versa. It is precisely the projective structure that makes HMS models tractable in applications and lets us analyze phenomena across "awareness levels" $\{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$. Since the motivation for HMS models in game theory and economics is to isolate the behavioral implications of unawareness from other factors like mistakes in information processing etc., we require the possibility correspondences satisfy strong properties analogous to S5.[4]

Assumption 1 *For any individual* $i \in I$, *we require that the possibility correspondence* Π_i *satisfies*

Confinement: *If* $\omega \in S_\Phi$, *then* $\Pi_i(\omega) \subseteq S_\Psi$ *for some* $\Psi \subseteq \Phi$.

Generalized Reflexivity: $\omega \in \Pi_i^{\uparrow}(\omega)$ *for every* $\omega \in \Omega$.[5]

Stationarity: $\omega' \in \Pi_i(\omega)$ *implies* $\Pi_i(\omega') = \Pi_i(\omega)$.

Projections Preserve Ignorance: *If* $\omega \in S_\Phi$ *and* $\Psi \subseteq \Phi$, *then* $\Pi_i^{\uparrow}(\omega) \subseteq \Pi_i^{\uparrow}(\omega_\Psi)$.

Projections Preserve Knowledge: *If* $\Upsilon \subseteq \Psi \subseteq \Phi$, $\omega \in S_\Phi$ *and* $\Pi_i(\omega) \subseteq S_\Psi$ *then* $(\Pi_i(\omega))_\Upsilon = \Pi_i(\omega_\Upsilon)$.

[4] Again, for lack of space we refer to Heifetz, Meier, and Schipper (2006, 2008) for discussions of these properties. Generalizations are considered by Heifetz, Meier, and Schipper (2013a), Halpern and Rêgo (2008), Board, Chung, and Schipper (2011), and Galanis (2011, 2013).

[5] Here and in what follows, we abuse notation slightly and write $\Pi_i^{\uparrow}(\omega)$ for $(\Pi_i(\omega))^{\uparrow}$.

Sometimes we denote by $S_{\Pi_i(\omega)}$ the state-space S for which $\Pi_i(\omega) \subseteq S$. We refer to Heifetz, Meier, and Schipper (2006, 2008) for discussions of these properties.

Given the possibility correspondence, the knowledge operator is defined as usual

Definition 2 *For every individual $i \in I$, the* knowledge operator *on events is defined by, for every event $E \in \Sigma$, $K_i(E) := \{\omega \in \Omega : \Pi_i(\omega) \subseteq E\}$ if there exists a state $\omega \in \Omega$ such that $\Pi_i(\omega) \subseteq E$, and by $K_i(E) = \emptyset^{S(E)}$ otherwise.*

Definition 3 *For every individual $i \in I$, the* awareness operator *on events is defined by, for every event $E \in \Sigma$, $A_i(E) := \{\omega \in \Omega : S_{\Pi_i(\omega)} \succeq S(E)\}$ if there exists a state $\omega \in \Omega$ such that $S_{\Pi_i(\omega)} \succeq S(E)$, and by $A_i(E) = \emptyset^{S(E)}$ otherwise. The unawareness operator is defined by $U_i(E) := \neg A_i(E)$.*

We read $K_i(E)$ as "individual i knows the event E" and $A_i(E)$ as "individual i is aware of event E".

Lemma 1 (Heifetz, Meier, and Schipper, 2006) *For every individual $i \in I$ and event $E \in \Sigma$, both $K_i(E)$ and $A_i(E)$ are $S(E)$-based events.*

Proposition 1 (Heifetz, Meier, and Schipper, 2006) *For every individual $i \in I$, the knowledge operator K_i satisfies the following properties: For every $E, F \in \Sigma$ and $\{E_n\}_n \subseteq \Sigma$,*

(i) Necessitation: $K_i(\Omega) = \Omega$,

(ii) Conjunction: $K_i\left(\bigcap_n E_n\right) = \bigcap_n K_i(E_n)$,

(iii) Truth: $K_i(E) \subseteq E$,

(iv) Positive Introspection: $K_i(E) \subseteq K_i K_i(E)$,

(v) Monotonicity: $E \subseteq F$ implies $K_i(E) \subseteq K_i(F)$.

(vi) Weak Negative Introspection I: $\neg K_i(E) \cap \neg K_i \neg K_i(E) \subseteq \neg K_i \neg K_i \neg K_i(E)$.

Proposition 2 (Heifetz, Meier, and Schipper, 2006) *For every individual $i \in I$, the following properties of knowledge and awareness obtain: For every $E \in \Sigma$ and $\{E_n\}_n \subseteq \Sigma$,*

1. *KU Introspection: $K_i U_i(E) = \emptyset^{S(E)}$,*
2. *AU Introspection: $U_i(E) = U_i U_i(E)$*
3. *Weak Necessitation: $A_i(E) = K_i(S(E)^{\uparrow})$,*
4. *Plausibility: $A_i(E) = K_i(E) \cup K_i \neg K_i(E)$,*
5. *Strong Plausibility: $U_i(E) = \bigcap_{n=1}^{\infty} (\neg K_i)^n (E)$,*
6. *Weak Negative Introspection II: $\neg K_i(E) \cap A_i \neg K_i(E) = K_i \neg K_i(E)$,*
7. *Symmetry: $A_i(E) = A_i(\neg E)$,*
8. *A-Conjunction: $\bigcap_n A_i(E_n) = A_i\left(\bigcap_n E_n\right)$,*
9. *AK-Self Reflection: $A_i(E) = A_i K_i(E)$,*
10. *AA-Self Reflection: $A_i(E) = A_i A_i(E)$,*
11. *A-Introspection: $A_i(E) = K_i A_i(E)$.*

The following lemma turns out to be very useful but has not been proved in the literature. (For the proof, see the full version of the paper.)

Lemma 2 *For every individual $i \in I$ and any $\Upsilon \subseteq \Psi \subseteq \Phi \subseteq \mathsf{At}$, if $\omega \in S_\Phi$ and $\Pi_i(\omega) \subseteq S_\Upsilon$, then $\Pi_i(\omega_\Psi) = \Pi_i(\omega)$.*

3 From Explicit to Implicit

In this section, we introduce the implicit possibility correspondence Λ_i as derived from Π_i. We then define implicit knowledge as based on Λ_i and show that it satisfies standard S5 properties as well as properties of Fagin and Halpern (1988) that are jointly satisfied by implicit knowledge, explicit knowledge, and awareness.

From now on we call for any individual $i \in I$, Π_i the *explicit* possibility correspondence, $\Pi_i(\omega)$ *explicit* possibility set at ω, and $K_i(E)$ the event that i *explicitly* knows E.

Definition 4 *Given the explicit possibility correspondence* Π_i *of individual* $i \in I$*, let the* implicit possibility correspondence $\Lambda_i : \Omega \longrightarrow 2^{\Omega}$ *satisfy*

Reflexivity: *For any* $\omega \in \Omega$*,* $\omega \in \Lambda_i(\omega)$*.*

Stationarity*:* $\omega' \in \Lambda_i(\omega)$ *implies* $\Lambda_i(\omega') = \Lambda_i(\omega)$*.*

Projections Preserve Implicit Knowledge: *For any* $\Phi \subseteq \mathsf{At}$*, if* $\omega \in S_\Phi$*, then* $\Lambda_i(\omega)_\Psi = \Lambda_i(\omega_\Psi)$ *for all* $\Psi \subseteq \Phi$*.*

Explicit Measurability: $\omega' \in \Lambda_i(\omega)$ *implies* $\Pi_i(\omega') = \Pi_i(\omega)$*.*

Implicit Measurability: $\omega' \in \Pi_i(\omega)$ *implies* $\Lambda_i(\omega') = \Lambda_i(\omega)_{S_{\Pi_i(\omega)}}$*.*

Given an HMS model $M = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Pi_i)_{i \in I}, v \rangle$ *and a collection of implicit possibility correspondences* $(\Lambda_i)_{i \in I}$ *satisfying the above properties, we call* $\overline{\mathsf{M}} = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Pi_i)_{i \in I}, (\Lambda_i)_{i \in I}, v \rangle$ *a* complemented HMS model.

A complemented HMS model is a HMS model complemented with implicit possibility correspondences for each individual. In the following, we discuss and derive some properties of the implicit possibility correspondence. It also demonstrates ways in which the implicit possibility correspondence is consistent with the explicit possibility correspondence.

Reflexivity and Stationarity are standard and imply that $\{\Lambda_i(\omega)\}_{\omega \in S_\Phi}$ forms a partition of S_Φ for every $\Phi \subseteq \mathsf{At}$. It is straightforward to see that they also imply a strengthening of Confinement (Assumption 1): The implicit possibility set at a state must be a subset of the state's space. That is, both the state and the implicit possibility set are described using the *same* language. More formally:

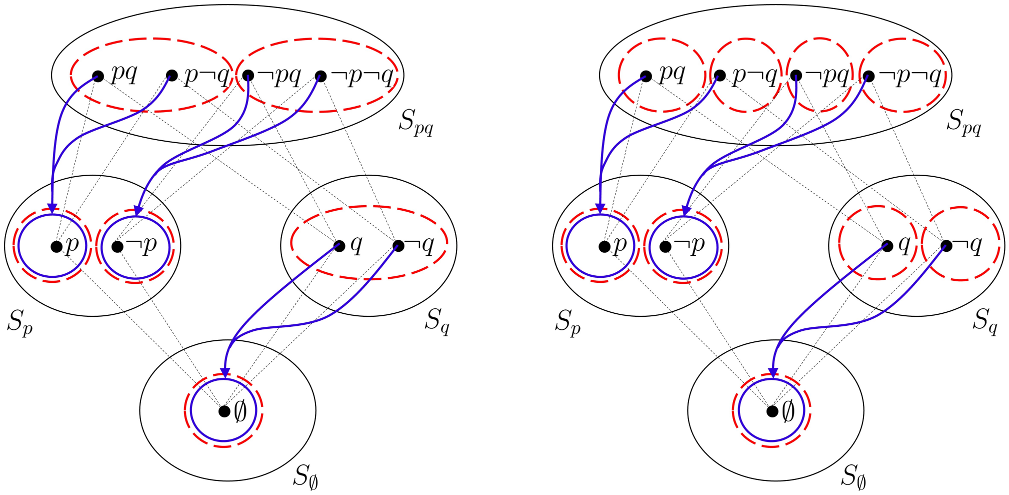
Lemma 3 (Strong Confinement) *For any individual* $i \in I$*,* $\Phi \subseteq \mathsf{At}$*, and* $\omega \in S_\Phi$*,* $\Lambda_i(\omega) \subseteq S_\Phi$*.*

Projections Preserve Implicit Knowledge is analogous to Projections Preserve Knowledge satisfied by Π_i. The absence of Projections Preserve (Implicit) Ignorance from the above list of imposed properties may look puzzling at the first glance. Yet, as we show below it is implied by Strong Confinement and Projections Preserve Implicit Knowledge.

Lemma 4 (Projections Preserve Implicit Ignorance) *For any individual* $i \in I$*, if* Λ_i *satisfies Strong Confinement and Projections Preserve Implicit Knowledge, then* Λ_i *satisfies Projections Preserve Implicit Ignorance. That is, for all* $\Phi \subseteq \mathsf{At}$*, if* $\omega \in S_\Phi$*, then* $\Lambda_i^{\uparrow}(\omega) \subseteq \Lambda_i^{\uparrow}(\omega_\Psi)$ *for all* $\Psi \subseteq \Phi$*.*

Explicit Measurability says that explicit knowledge is measurable with respect to implicit knowledge. That is, the agent always implicitly knows her explicit knowledge. The converse, Implicit Measurability, is more subtle because of awareness. An individual may not explicitly know her implicit knowledge because she might be unaware of some events. However, the individual always explicitly knows her implicit knowledge modulo awareness. That is, she might implicitly know more at a higher awareness

Figure 1: Examples of Implicit Knowledge in Unawareness Structures



level than what she knows at her awareness level (like in the structure to the right in Figure 1) but at her awareness level, her implicit knowledge equals her explicit knowledge. The following lemma formalizes the last conclusion. The proof uses all properties of Π_i and Λ_i except Projections Preserve Knowledge of both Λ_i and Π_i and Projections Preserve Ignorance of Π_i.

Lemma 5 *For any individual $i \in I$, if $\omega' \in \Pi_i(\omega)$, then $\Lambda_i(\omega') = \Pi_i(\omega')$.*

Lemma 6 (Coherence) *For any individual $i \in I$, $\omega \in \Omega$, $\Lambda_i(\omega)_{S_{\Pi_i(\omega)}} = \Pi_i(\omega)$.*

Figure 1 illustrates with two examples of how implicit knowledge can be "fitted" to explicit knowledge. Consider first the HMS model to the left. There are four spaces indexed by subsets of atomic formulae. Anticipating the semantics of HMS models introduced later, we describe and call states by their atomic formulae. The explicit possibility correspondence of the individual is indicated by the solid blue ovals and arrows. For instance, at state pq she considers possible state p. That is, she is unaware of q and knows p. Similarly, at state $\neg pq$ she is unaware of q and knows $\neg p$. Her implicit possibility correspondence is given by the red dashed ovals. Note that in this complemented HMS model she does not implicitly know more than she does explicitly. Contrast this with the HMS model to the right. There, she implicitly knows q for instance at state pq (because her implicit possibility set at pq is $\{pq\}$) although she is not aware of q (because her explicit possibility set at pq is on $S_{\{p\}}$). and hence does not explicitly know q. The figures demonstrate that the implicit possibility correspondence may be consistent with the explicit possibility correspondence in two different ways. It may model implicit knowledge that is finer than the explicit knowledge (like in the figure to the right) or implicit knowledge that is as coarse as the explicit knowledge but not coarser (like in the figure to the left). Note that a version of the models in Figure 1 in which only $\{pq, p\neg q\}$ is in a red dashed oval while $\neg pq$ and $\neg p\neg q$ are in distinct circles in S_{pq} is ruled out by Projections Preserve Implicit Knowledge.

Given implicit possibility correspondences, we proceed with the definition of the implicit knowledge operators.

Definition 5 *For any individual $i \in I$, the* implicit knowledge operator *on events $E \in \Sigma$ is*

$$L_i(E) := \{\omega \in \Omega : \Lambda_i(\omega) \subseteq E\}$$

if there exists a state $\omega \in \Omega$ such that $\Lambda_i(\omega) \subseteq E$ and by $L_i(E) = \emptyset^{S(E)}$ otherwise.

The next observation follows immediately from the properties of the implicit possibility correspondence and the proof of Lemma 1 in Heifetz, Meier, and Schipper (2006).

Lemma 7 *For any individual $i \in I$ and event $E \in \Sigma$, $L_i(E)$ is an $S(E)$-based event.*

Implicit knowledge satisfies all properties of "partitional" knowledge.

Proposition 3 *For any individual $i \in I$, L_i satisfies for any $E, F \in \Sigma$ and $\{E_n\}_n \subseteq \Sigma$,*

(i) Necessitation: For $\Phi \subseteq \mathsf{At}$, $L_i(S_\Phi^\uparrow) = S_\Phi^\uparrow$,

(ii) Conjunction: $L_i\left(\bigcap_n E_n\right) = \bigcap_n L_i(E_n)$,

(iii) Monotonicity: $E \subseteq F$ implies $L_i(E) \subseteq L_i(F)$,

(iv) Truth: $L_i(E) \subseteq E$,

(v) Positive Introspection: $L_i(E) \subseteq L_i L_i(E)$,

(vi) Negative Introspection: $\neg L_i(E) \subseteq L_i \neg L_i(E)$.

We observe that as in Fagin and Halpern (1988), explicit knowledge of an event equals implicit knowledge and awareness of that event.

Proposition 4 *For any $i \in I$ and event $E \in \Sigma$,*

1. $K_i(E) = L_i(E) \cap A_i(E)$,

2. $U_i(E) = L_i(U_i(E))$,

3. $A_i(E) = L_i(A_i(E))$,

4. $A_i L_i(E) = A_i(E)$.

Properties 2. and 3. above mean that the individual implicitly knows her unawareness. This is in contrast to explicit knowledge since by KU introspection an individual can never explicitly know that she is unaware of an event. Property 4 says that an individual is aware of her implicit knowledge of an event if and only she is aware of the event. That is, the moment she can reason about an event, she can also reason about her implicit knowledge of the event. This is analogous to AK-Self-Reflection of explicit knowledge.

4 From Implicit to Explicit

In the previous section, we showed how implicit knowledge can be derived from explicit knowledge. In this section, we go the other direction. We can devise a version of HMS model that features possibility correspondences capturing implicit knowledge and (non-syntactic) awareness functions as primitives, and then derive the possibility correspondence capturing explicit knowledge.

Definition 6 *An* implicit knowledge-based HMS model $\mathsf{M}^* = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r_\Psi^\Phi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda_i^*)_{i \in I}, (\alpha_i)_{i \in I}, v\rangle$ *consists of*

- *a non-empty set of individuals I,*

- *a nonempty collection of nonempty disjoint state spaces* $\{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$ *(as in Definition 1),*
- *projections* $(r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}$ *(as in Definition 1),*
- *an implicit possibility correspondence* $\Lambda^*_i : \Omega \longrightarrow 2^\Omega \setminus \{\emptyset\}$*, for all* $i \in I$*,*
- *an awareness function* $\alpha_i : \Omega \longrightarrow \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$*, for all* $i \in I$*,*
- *a valuation function* $v : \mathsf{At} \longrightarrow \Sigma$*.*

Like HMS models, implicit knowledge-based HMS models feature a projective lattice of state-spaces. However, instead of the explicit possibility correspondence, we now take the implicit possibility correspondences as a primitive. As before, we are interested in strong properties of knowledge associated with S5 because (1) these properties have been used for explicit knowledge in applications, and (2) we will require explicit knowledge to be consistent with implicit knowledge. As such, we are interested how the rich structure of S5 translates into properties of a derived explicit possibility correspondence. To that end, we require:

Assumption 2 *For each individual* $i \in I$*, the implicit possibility correspondence* Λ^*_i *satisfies Reflexivity, Stationarity, and Projections Preserve Implicit Knowledge.*

These properties were also satisfied by implicit possibility correspondences in the previous section.[6]

The second primitive of implicit knowledge-based HMS models is the awareness function α_i for every individual $i \in I$. We impose the following properties on α_i:

Assumption 3 *For each individual* $i \in I$*, the awareness function* $\alpha_i : \Omega \longrightarrow \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$ *satisfies*

O. Lack of Conception: If $\omega \in S_\Phi$*, then* $\alpha_i(\omega) \preceq S_\Phi$*.*

I. Awareness Measurability: If $\omega' \in \Lambda^*_i(\omega)$*, then* $\alpha_i(\omega') = \alpha_i(\omega)$*.*

II. If $\omega \in S_\Phi$ *and* $S_\Psi \preceq \alpha_i(\omega)$*, then* $\alpha_i(\omega_\Psi) = S_\Psi$*.*

III. If $\omega \in S_\Phi$ *and* $\alpha_i(\omega) \preceq S_\Psi \preceq S_\Phi$*, then* $\alpha_i(\omega_\Psi) = \alpha_i(\omega)$*.*

IV. If $\omega \in S_\Phi$ *and* $\Psi \subseteq \Phi$*, then* $\alpha_i(\omega) \succeq \alpha_i(\omega_\Psi)$*.*

When $\alpha_i(\omega) \in S$ *for some* $S \in \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$*, we call* S *the* awareness level of i at ω.

Property O. models one feature of Confinement of HMS models (see Assumption 1). Note that Confinement in HMS models has two features: First, it requires that the possibility set at a state is a subset of exactly one space. Second, it says that this space must be weakly less expressive than the space containing the state. Only this second last feature is captured by property O. The idea is that an individual may have lack of conception. Property I. is a measurability condition. Awareness is measurable with respect to implicit knowledge. The implication is that an agent implicitly knows her own awareness. Properties II. to IV. are consistency properties of awareness across the lattice. Projections preserve awareness as long as it is still expressible in the spaces. While property II. preserves awareness for corresponding states in spaces less expressive than the awareness level at a state, property III. preserves awareness for corresponding states in spaces more expressive than the awareness level at that state.

Definition 7 *Given an implicit knowledge-based HMS model* $\mathsf{M}^* = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I},$ $(\alpha_i)_{i \in I}, v\rangle$*, define the explicit possibility correspondence* $\Pi^*_i : \Omega \longrightarrow 2^\Omega$ *by, for all* $\omega \in \Omega$ *and* $\Phi \subseteq \mathsf{At}$*,* $\Pi^*_i(\omega_\Phi) := \Lambda^*_i(\omega)_{\alpha_i(\omega_\Phi)}$*. We call* $\overline{\mathsf{M}}^* = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\Pi^*_i)_{i \in I}, (\alpha_i)_{i \in I}, v\rangle$ *the* complemented implicit knowledge-based HMS model.

[6] Note again that Reflexivity and Stationarity implies Strong Confinement. In more general settings without Reflexivity or Stationarity, at least Strong Confinement would have to be imposed in Λ^*_i for every $i \in I$.

The defining condition for the explicit possibility correspondence in implicit knowledge-based HMS models is a slight strengthening of Coherence derived from the explicit and implicit measurability in Lemma 6. Here we take it as the primitive to connect explicit knowledge to implicit knowledge.

The following observations are immediate:

Lemma 8 *For all $\omega \in \Omega$,*

A. $\Pi_i^*(\omega) = \Lambda_i^*(\omega)_{\alpha_i(\omega)}$,

B. $\Pi_i^*(\omega_\Phi) = \Lambda_i^*(\omega)_\Phi$ *for all* $\Phi \subseteq \mathsf{At}$ *with* $S_\Phi \preceq \alpha_i(\omega)$,

C. $\Pi_i^*(\omega_\Phi) = \Lambda_i^*(\omega)_{\alpha_i(\omega)}$ *for all* $\Phi \subseteq \mathsf{At}$ *with* $S_\omega \succeq S_\Phi \succeq \alpha_i(\omega)$.

The following lemma records properties of the derived explicit possibility correspondence. It shows that it satisfies the properties of the explicit possibility correspondence of HMS models.

Lemma 9 *For any individual $i \in I$, if α_i satisfies O., I., II., III., and IV., then Π_i^* satisfies Confinement, Generalised Reflexivity, Stationarity, Projections Preserve Ignorance, Projections Preserve Knowledge.*

We conclude that the derived explicit possibility correspondence Π_i^* is a possibility correspondence as in Heifetz, Meier, and Schipper (2006, 2008), i.e., satisfies Assumption 1. To show that the connection between the derived explicit possibility correspondence and the implicit possibility correspondence is as in the complemented HMS model of the prior section, we note the following lemma.

Lemma 10 *For any individual $i \in I$, Λ_i^* and Π_i^* jointly satisfy explicit and implicit measurability.*

The above lemmata imply the following:

Corollary 1 *For any implicit knowledge-based HMS model* $\mathsf{M}^* = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r_\Psi^\Phi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda_i^*)_{i \in I}, (\alpha_i)_{i \in I}, v\rangle$ *with derived explicit possibility correspondences* $(\Pi_i^*)_{i \in I}$ *we have that* $\overline{\mathsf{M}} = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r_\Psi^\Phi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda_i^*)_{i \in I}, (\Pi_i^*)_{i \in I}, v\rangle$ *is a complemented HMS model and* $\mathsf{M} = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r_\Psi^\Phi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Pi_i^*)_{i \in I}, v\rangle$ *is a HMS model.*

The awareness function can be directly used to define an awareness operator on events.

Definition 8 *For each individual $i \in I$, define an awareness operator on events by for all $E \in \Sigma$, $A_i^*(E) := \{\omega \in \Omega : \alpha_i(\omega) \succeq S(E)\}$ if there is a state $\omega \in \Omega$ such that $\alpha_i(\omega) \succeq S(E)$ and by $A_i^*(E) = \emptyset^{S(E)}$ otherwise.*

Similarly, for each individual $i \in I$, we can use the possibility correspondence Λ_i^* to define an implicit knowledge operator L_i^* as in Definition 5. Finally, let K_i be the explicit knowledge operator and A_i be the awareness operator defined from the derived explicit possibility correspondence Π_i^* as in Definitions 2 and 3, respectively.

The following proposition shows that awareness defined with the awareness function is equivalent to awareness defined with the derived explicit possibility correspondence. It also shows that explicit knowledge defined from the derived explicit possibility correspondence is equivalent to implicit knowledge and awareness.

Proposition 5 *For every $i \in I$ and any event $E \in \Sigma$,*

1. $A_i^*(E) = A_i(E)$

2. $K_i(E) = L_i^*(E) \cap A_i^*(E)$

The last two sections show an interdefinability of explicit and implicit knowledge in HMS models. Implicit knowledge can be defined in terms of explicit knowledge and vice versa. We can use either the explicit possibility correspondences as primitive or the implicit possibility correspondence together with the awareness function. Implicit knowledge-based HMS models are arguably closer to FH models than HMS models. We will use them to build a bridge between HMS and FH models.

5 Category of FH Models

In this section, we introduce FH models and bounded morphisms, a notion of structure preserving maps between FH models, and use these to form a category with FH models as objects and bounded morphisms as morphisms.

The semantics of FH models is not syntax-free since each agent's awareness function assigns to each state a set of formulae. Thus, we first introduce the formal language featuring implicit knowledge, awareness, and explicit knowledge. With $i \in I$ and $p \in \mathsf{At}$, define the language $\mathscr{L}_{\mathsf{At}}$ by

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \ell_i\varphi \mid a_i\varphi \mid k_i\varphi$$

Let $\mathsf{At}(\varphi) := \{p \in \mathsf{At}\colon p \text{ is a subformula of } \varphi\}$, for any $\varphi \in \mathscr{L}_{\mathsf{At}}$, be the set of atomic formulae that are contained in φ, and let $\mathscr{L}_\Phi := \{\varphi \in \mathscr{L}_{\mathsf{At}} : \mathsf{At}(\varphi) \subseteq \Phi\}$ be the sublanguage of $\mathscr{L}_{\mathsf{At}}$ built on propositions p in $\Phi \subseteq \mathsf{At}$.

The formula $\ell_i\varphi$ reads "agent i implicitly knows φ", $a_i\varphi$ reads "i is aware of φ", and $k_i\varphi$ reads "i explicitly knows φ". Fagin and Halpern (1988) define explicit knowledge as the conjunction of implicit knowledge and awareness, namely $k_i\varphi = a_i\varphi \wedge \ell_i\varphi$, for $\varphi \in \mathscr{L}_{\mathsf{At}}$.

Definition 9 *For any $\Phi \subseteq \mathsf{At}$, a* FH model $\mathsf{K}_\Phi = \langle I, W_\Phi, (R_{\Phi,i})_{i\in I}, (\mathscr{A}_{\Phi,i})_{i\in I}, V_\Phi\rangle$ *for Φ consists of*

- *a non-empty set of individuals I,*
- *a non-empty set of states W_Φ,*
- *an accessibility relation $R_{\Phi,i} \subseteq W_\Phi \times W_\Phi$, for all $i \in I$,*
- *an awareness function $\mathscr{A}_{\Phi,i} : W_\Phi \longrightarrow 2^{\mathscr{L}_\Phi}$, for all $i \in I$, assigning to each state $w \in W_\Phi$ a set of formulas $\mathscr{A}_{\Phi,i}(w) \subseteq \mathscr{L}_\Phi$. The set $\mathscr{A}_{\Phi,i}(w)$ is called the* awareness set of i at w.
- *a valuation function $V_\Phi : \Phi \longrightarrow 2^{W_\Phi}$.*

Assumption 4 *We require that the FH model K_Φ is* propositionally determined, *i.e., for every $i \in I$, the awareness functions satisfy*

> Awareness is Generated by Primitive Propositions: *For all $\varphi \in \mathscr{L}_\Phi$, $\varphi \in \mathscr{A}_{\Phi,i}(w)$ if and only if for all $p \in \mathsf{At}(\varphi)$, $p \in \mathscr{A}_{\Phi,i}(w)$.*
>
> Agents Know What They Are Aware of: *$(w,t) \in R_{\Phi,i}$ implies $\mathscr{A}_{\Phi,i}(w) = \mathscr{A}_{\Phi,i}(t)$.*

We also require that the FH model K_Φ is partitional, *that is, $R_{\Phi,i}$ is an equivalence relation, i.e., satisfies reflexivity, transitivity, and Euclideaness.*

Throughout the paper, we focus on partitional and propositionally determined FH models because these models capture the notion of awareness and knowledge used in most applications so far and it is also the notion of awareness used in HMS models. We are interested in how this rich structure maps between FH models as well as between FH and HMS models.

Definition 10 *For any $\Psi \subseteq \Phi \subseteq \mathsf{At}$ and FH models $\mathsf{K}_\Phi = \langle I, W_\Phi, (R_{\Phi,i})_{i\in I}, (\mathscr{A}_{\Phi,i})_{i\in I}, V_\Phi\rangle$ and $\mathsf{K}_\Psi = \langle I, W_\Psi, (R_{\Psi,i})_{i\in I}, (\mathscr{A}_{\Psi,i})_{i\in I}, V_\Psi\rangle$, the mapping $f^\Phi_\Psi : \mathsf{K}_\Phi \longrightarrow \mathsf{K}_\Psi$ is a* surjective bounded morphism *if for every $i \in I$ and $w \in W_\Phi$*

- Surjectivity: *$f^\Phi_\Psi : W_\Phi \longrightarrow W_\Psi$ is a surjection,*
- Atomic harmony: *for every $p \in \Psi$, $w \in V_\Phi(p)$ if and only if $f^\Phi_\Psi(w) \in V_\Psi(p)$,*
- Awareness consistency: $\mathscr{A}_{\Phi,i}(w) \cap \mathscr{L}_\Psi = \mathscr{A}_{\Psi,i}(f^\Phi_\Psi(w))$

- Homomorphism: *f^{Φ}_{Ψ} is a homomorphism w.r.t. $R_{\Phi,i}$, i.e., if $(w,t) \in R_{\Phi,i}$, then $(f^{\Phi}_{\Psi}(w), f^{\Phi}_{\Psi}(t)) \in R_{\Psi,i}$,*
- Back: *if $(f^{\Phi}_{\Psi}(w), t') \in R_{\Psi,i}$, then there is a state $t \in W_{\Phi}$ such that $f^{\Phi}_{\Psi}(t) = t'$ and $(w,t) \in R_{\Phi,i}$.*

This is the standard notion of bounded morphism (also called p-morphism) (see for instance, Blackburn, de Rijke, and Venema, 2001, pp. 59–62) except for the additional property of Awareness Consistency. In our context, the bounded morphism literally bounds the language over which FH models are defined. We can now consider collections of FH models and bounded morphisms between them:

Definition 11 *Given the FH model K_{At}, the* category of FH models $\mathscr{C}(\mathsf{K}_{\mathsf{At}}) = \langle (\mathsf{K}_{\Phi})_{\Phi \subseteq \mathsf{At}}, (f^{\Phi}_{\Psi})_{\Psi \subseteq \Phi \subseteq \mathsf{At}} \rangle$ *consists of*

- *a collection of FH models K_{Φ}, one for each $\Phi \subseteq \mathsf{At}$,*
- *for any $\Phi, \Psi \subseteq \mathsf{At}$ with $\Psi \subseteq \Phi$, there is one surjective bounded morphism f^{Φ}_{Ψ}, such that*
  - *for any $\Phi \subseteq \mathsf{At}$, f^{Φ}_{Φ} is the identity,*
  - *for any $\Upsilon, \Phi, \Psi \subseteq \mathsf{At}$ with $\Upsilon \subseteq \Psi \subseteq \Phi$, $f^{\Phi}_{\Upsilon} = f^{\Psi}_{\Upsilon} \circ f^{\Phi}_{\Psi}$.*

Our terminology is not arbitrary. The category of FH models is indeed a category in the sense of category theory. It has an initial object, the most expressive FH model K_{At}, as well as a terminal object, the FH model $\mathsf{K}_{\emptyset}$.

Since the category of FH models is defined with bounded morphisms, it suggests that all FH models in the category are in some sense epistemically equivalent. Indeed, we interpret each category of FH models literally as the category of FH models that vary with the language but are otherwise modally equivalent. That is, for any $\Psi \subseteq \Phi \subseteq \mathsf{At}$, modal satisfaction for K_{Ψ} is as for K_{Φ} w.r.t. formulae in $\mathscr{L}_{\Psi}$ (see Lemma 11 below). We interpret this as follows: When a modeler represents a context with a FH model K_{At}, an agent i at state $w \in W_{\mathsf{At}}$ can be thought of representing it with the FH model $\mathsf{K}_{\mathsf{At}(\mathscr{A}_{\mathsf{At},i}(w))}$. And this agent i considers it possible at w that at t with $(f^{\mathsf{At}}_{\mathsf{At}(\mathscr{A}_{\mathsf{At},i}(w))}(w), t) \in R_{\mathsf{At}(\mathscr{A}_{\mathsf{At},i}(w)),i}$ agent j represents the situation with the FH model $\mathsf{K}_{\mathsf{At}(\mathscr{A}_{\mathsf{At},j}(t))}$, etc. These models can all be seen as equivalent except for the language of which they are defined. With this construction, we do not just endow agents with a formal language to reason about their context but we also allow them to analyze their context with semantic devices like logicians do. This is relevant because in many multi-agent contexts of game theory, the analysis proceeds using semantic devices like state spaces etc. rather than at the level of syntax. For instance, in a principal-agent problem, the principal may want to use a FH model augmented by actions and utility functions to analyze optimal contract design realizing that the (unaware) agent may also use a less expressive but otherwise equivalent FH model to analyze how to optimally interact with the principal.

To make the equivalence between models in the category of FH models precise, we need to introduce the semantics of FH models.

Definition 12 *For any $\Phi \subseteq \mathsf{At}$, FH model $\mathsf{K}_{\Phi} = \langle I, W_{\Phi}, (R_{\Phi,i})_{i \in I}, (\mathscr{A}_{\Phi,i})_{i \in I}, V_{\Phi} \rangle$, and $\omega \in W_{\Phi}$, satisfaction of formulae in $\mathscr{L}_{\Phi}$ is given by the following clauses:*

$$\begin{array}{llllll}
\mathsf{K}_{\Phi}, w \Vdash \top & & \textit{for all } w \in W_{\Phi}; & \mathsf{K}_{\Phi}, w \Vdash \varphi \wedge \psi & \textit{iff} & \mathsf{K}_{\Phi}, w \Vdash \varphi \textit{ and } \mathsf{K}_{\Phi}, w \Vdash \psi; \\
\mathsf{K}_{\Phi}, w \Vdash p & \textit{iff} & w \in V_{\Phi}(p); & \mathsf{K}_{\Phi}, w \Vdash a_i\varphi & \textit{iff} & \varphi \in \mathscr{A}_{\Phi,i}(w); \\
\mathsf{K}_{\Phi}, w \Vdash \neg\varphi & \textit{iff} & \mathsf{K}_{\Phi}, w \nVdash \varphi; & \mathsf{K}_{\Phi}, w \Vdash \ell_i\varphi & \textit{iff} & \mathsf{K}_{\Phi}, t \Vdash \varphi \textit{ for all } (w,t) \in R_{\Phi,i}.
\end{array}$$

From this semantics and the syntactic definition $k_i\varphi := \ell_i\varphi \wedge a_i\varphi$, it follows that $\mathsf{K}_{\Phi}, w \Vdash k_i\varphi$ if and only if for all t s.t. $(w,t) \in R_{\Phi,i}$, $\mathsf{K}_{\Phi}, t \Vdash \varphi$ and $\varphi \in \mathscr{A}_{\Phi,i}(w)$.

The category of FH models forms a complete lattice induced by set inclusion on sets of atomic formulae with the initial object being the join of the lattice and the terminal object being the meet of the lattice. We now show that it gives rise to a complete lattice when ordered using the (directed) bounded morphism or, epistemically more relevant, when ordered by modal equivalence relative to sublanguages.

Proposition 6 *Given a category of FH models, $\langle (\mathsf{K}_\Phi)_{\Phi\subseteq\mathsf{At}}, (f^\Phi_\Psi)_{\Psi\subseteq\Phi\subseteq\mathsf{At}}\rangle$, modal equivalence relative to sublanguages forms a complete lattice of FH models in the category as follows: For any nonempty set of subsets of atomic formulae $\mathscr{F} \subseteq 2^{\mathsf{At}}$,*

(i) $\mathsf{K}_{\bigcup_{\Phi\in\mathscr{F}}\Phi}$ *is modally equivalent to* K_Ψ *w.r.t.* $\mathscr{L}_\Psi$ *for every* $\Psi \in \mathscr{F}$*, i.e., for any* $w \in W_{\bigcup_{\Phi\in\mathscr{F}}\Phi}$, $\varphi \in \mathscr{L}_\Psi$, $\mathsf{K}_{\bigcup_{\Phi\in\mathscr{F}}\Phi}, w \Vdash \varphi$ *iff* $\mathsf{K}_\Psi, f_\Psi^{\bigcup_{\Phi\in\mathscr{F}}\Phi}(w) \Vdash \varphi$*, and*

(ii) $\mathsf{K}_{\bigcap_{\Phi\in\mathscr{F}}\Phi}$ *is modally equivalent to* K_Ψ *w.r.t.* $\mathscr{L}_{\bigcap_{\Phi\in\mathscr{F}}}$ *for every* $\Psi \in \mathscr{F}$*, i.e., for any* $w \in W_\Psi$, $\varphi \in \mathscr{L}_{\bigcap_{\Phi\in\mathscr{F}}\Phi}$, $\mathsf{K}_\Psi, w \Vdash \varphi$ *iff* $\mathsf{K}_{\bigcap_{\Phi\in\mathscr{F}}\Phi}, f^\Psi_{\bigcap_{\Phi\in\mathscr{F}}\Phi}(w) \Vdash \varphi$.

The proof of the proposition uses of the following lemma:

Lemma 11 *For any* $\Psi, \Phi \subseteq \mathsf{At}$ *with* $\Psi \subseteq \Phi$*, all* $w \in W_\Phi$*, and all* $\varphi \in \mathscr{L}_\Psi$*,* $\mathsf{K}_\Phi, w \Vdash \varphi$ *if and only if* $\mathsf{K}_\Psi, f^\Phi_\Psi(w) \Vdash \varphi$.

Note that for a collection of FH models $\{\mathsf{K}_\Psi\}_{\Psi\in\mathscr{F}}$, the “join” and “meet” FH models are $\mathsf{K}_{\bigcup_{\Phi\in\mathscr{F}}\Phi}$ and $\mathsf{K}_{\bigcap_{\Phi\in\mathscr{F}}\Phi}$, respectively. So for any collection of FH models, Proposition 6 shows modal equivalence between any FH model in the collection and its join and meet models, respectively.

Our notion of bounded morphism is inspired by bisimulation of FH models introduced by van Ditmarsch et al. (2018). Clearly, the surjective bounded morphism is a bisimulation. Here we discuss some differences and similarities. While bisimulation more generally is a relation between models without a particular direction, the bounded morphism has a natural direction from the more expressive FH model to the less expressive FH model. Further, it is a function on W_Φ. That is, it maps *every* state in W_Φ to a state in W_Ψ with $\Psi \subseteq \Phi$. Moreover, surjectivity is a property that is straightforward to define for functions. Finally, bounded morphisms easily compose and almost naturally lead to the notion of category of FH models although we do not really make much use here of the machinery of category theory. For all these reasons, we use the notion of bounded morphism. Van Ditmarsch et al. (2018) introduced two notions of bisimulation for FH models, standard bisimulation and awareness bisimulation. Like our bounded morphism, both of their notions of bisimulation also depend on a subset of atomic formulae for FH models. The clauses Atomic harmony, Awareness consistency, Homomorphism, and Back have counterparts in their notions of bisimulations. Our notion of bounded morphism is closer to what they call standard bisimulation because our Homomorphism and Back clauses do not involve the awareness function. Although van Ditmarsch et al. (2018, p. 63) mention the projective lattice structure of Heifetz, Meier, and Schipper (2006, 2008) as a motivation for their notion of awareness bisimulation, we believe it is particularly useful for their notion of speculative knowledge. Their notions of bisimulations do not require surjectivity although when considering maximal bisimulations, they must be surjective since they yield quotient models. Compositions of maximal bisimulation commute like we require our bounded morphism to do in the category of FH models but bisimulations that are not maximal do not necessarily commute. Moreover, maximal bisimulations yield necessarily contractions that eliminate redundancies. We are unsure whether it is necessarily a natural property when we interpret categories of FH models as collections of subjective views of agents. An agent may not realize or may not be bothered by redundancies and use an FH model with redundancies to analyze her situation. That is, differences in awareness and redundancies are orthogonal to each other and reduction in awareness does not necessitate elimination of redundancies.

6 Transformations

6.1 From FH Models to HMS Models

We can use the tools of the prior sections to define a transformation of a FH model into a HMS model. The transformation works as follows: For any FH model K_{At} for At, consider the category of FM models $\langle (\mathsf{K}_\Phi)_{\Phi \subseteq \mathsf{At}}, (f^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}} \rangle$. This category is transformed into an implicit knowledge-based HMS model. We then derive the explicit possibility correspondences and add them to the implicit knowledge-based HMS model, obtaining a complemented implicit knowledge-based HMS model. In the next step, we erase the awareness functions and get a complemented HMS model. The core step is to transform a category of FH models into an implicit knowledge-based HMS model. This is defined next.

Definition 13 *For any category of FH models $\mathscr{C}(\mathsf{K}_{\mathsf{At}}) = \langle (\mathsf{K}_\Phi)_{\Phi \subseteq \mathsf{At}}, (f^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}} \rangle$, the T-*transform *of $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$ is the implicit knowledge-based HMS model $T(\mathscr{C}(\mathsf{K}_{\mathsf{At}})) = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\alpha_i)_{i \in I}, v \rangle$ defined by:*

- *$S_\Phi = W_\Phi$ for all $\Phi \subseteq \mathsf{At}$, where W_Φ is the state space of the FH model K_Φ of the category $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$. Denote $\Omega = \bigcup_{\Phi \subseteq \mathsf{A}} S_\Phi$.*
- *$r^\Phi_\Psi = f^\Phi_\Psi$ for any $\Phi, \Psi \subseteq \mathsf{At}$ with $\Psi \subseteq \Phi$, where f^Φ_Ψ is the surjective bounded morphism of the category $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$.*
- *$\Lambda^*_i : \Omega \longrightarrow 2^\Omega$ such that $\Phi \subseteq \mathsf{At}$ and $w \in S_\Phi$, $w' \in \Lambda^*_i(w)$ if and only if $(w, w') \in R_{\Phi,i}$, for any $i \in I$,*
- *$\alpha_i : \Omega \longrightarrow \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}$ such that for all $\Psi \subseteq \mathsf{At}$ and $w \in S_\Psi$, $\alpha_i(w) = S_\Upsilon$ if and only if $\mathsf{At}(\mathscr{A}_{\Psi,i}(w)) = \Upsilon$, for any $i \in I$,*
- *$v(p) = \bigcup_{\Phi \subseteq \mathsf{At}} V_\Phi(p)$, for any $p \in \mathsf{At}$.*

The T-transform indeed transforms any category of FH models into an implicit knowledge-based HMS model.

Proposition 7 *For any category of FH models $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$, the T-transform $T(\mathscr{C}(\mathsf{K}_{\mathsf{At}}))$ is an implicit knowledge-based HMS model.*

We have all ingredients to define the transformation of FH models into complemented HMS models.

Definition 14 *For any FH model K_{At}, the* HMS transform *$HMS(\mathsf{K}_{\mathsf{At}}) = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\Pi^*_i)_{i \in I}, v \rangle$ is defined by the following steps:*

1. *Form the category of FH models $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$ (Definition 11).*
2. *Apply the T-transform to $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$ to obtain the implicit knowledge-based HMS model $T(\mathscr{C}(\mathsf{K}_{\mathsf{At}})) = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\alpha_i)_{i \in I}, v \rangle$ (Defin. 13).*
3. *Form the complemented implicit knowledge-based HMS model $\overline{T(\mathscr{C}(\mathsf{K}_{\mathsf{At}}))} = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\Pi^*_i)_{i \in I}, (\alpha_i)_{i \in I}, v \rangle$ by deriving the explicit possibility correspondences $(\Pi^*_i)_{i \in I}$ (Definition 7).*
4. *Erase the awareness functions $(\alpha_i)_{i \in I}$ from the complemented implicit knowledge-based HMS model $\overline{T(\mathscr{C}(\mathsf{K}_{\mathsf{At}}))}$ to obtain the complemented HMS model $\langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\Pi^*_i)_{i \in I}, v \rangle$.*

Corollary 2 *For any FH model K_{At}, its $HMS(\mathsf{K}_{\mathsf{At}})$ is a complemented HMS model.*

6.2 From HMS Models to FH Models

To transform a complemented HMS model into a FH model we simply need to consider the upmost space of the lattice of spaces of the HMS model, copy the domain, define accessibility relations from implicit possibility correspondences, and the valuation function, and, for every state $\omega \in S_{At}$, construct the awareness set at ω by collecting all the formulas that contain the atoms defined in the space where $\Pi_i(\omega)$ lies.

Definition 15 *For any complemented HMS model* $\overline{\mathsf{M}} = \langle I, \{S_\Phi\}_{\Phi\subseteq\mathsf{At}}, (r^{\Phi}_{\Psi})_{\Psi\subseteq\Phi\subseteq\mathsf{At}}, (\Lambda_i)_{i\in I}, (\Pi_i)_{i\in I}, v\rangle$, *the FH-*transform $FH(\overline{\mathsf{M}}) = \langle I, W_{\mathsf{At}}, (R_{\mathsf{At},i})_{i\in I}, (\mathscr{A}_{\mathsf{At},i})_{i\in I}, V_{\mathsf{At}}\rangle$ *is defined by*

- $W_{\mathsf{At}} = S_{\mathsf{At}}$,
- $R_{\mathsf{At},i} \subseteq W_{\mathsf{At}} \times W_{\mathsf{At}}$ *is such that* $(\omega,\omega') \in R_{\mathsf{At},i}$ *if and only if* $\omega' \in \Lambda_i(\omega)$, *for all* $i \in I$,
- $\mathscr{A}_{\mathsf{At},i} : W_{\mathsf{At}} \longrightarrow 2^{\mathscr{L}_{\mathsf{At}}}$ *is such that* $\mathscr{A}_{\mathsf{At},i}(\omega) = \mathscr{L}_\Phi$ *for* $\Phi \subseteq \mathsf{At}$ *with* $\Pi_i(\omega) \subseteq S_\Phi$, *for all* $i \in I$,
- $V_{\mathsf{At}} : \mathsf{At} \longrightarrow 2^{W_{\mathsf{At}}}$ *is such that* $V_{\mathsf{At}}(p) = v(p) \cap S_{\mathsf{At}}$, *for every* $p \in \mathsf{At}$.

The FH transform indeed transforms any complemented HMS model into a FH model.

Proposition 8 *For every complemented HMS model* $\overline{\mathsf{M}}$, $FH(\overline{\mathsf{M}})$ *is a FH model for* At.

6.3 Equivalence of HMS and FH Models

Before we can prove an equivalence of HMS and FH models, we need to introduce the semantics of complemented HMS models.

Definition 16 *Let* $\overline{\mathsf{M}} = \langle I, \{S_\Phi\}_{\Phi\subseteq\mathsf{At}}, (r^{\Phi}_{\Psi})_{\Psi\subseteq\Phi\subseteq\mathsf{At}}, (\Lambda_i)_{i\in I}, (\Pi_i)_{i\in I}, v\rangle$ *be a complemented HMS model and let* $\omega \in \Omega$. *Satisfaction of* $\mathscr{L}_{\mathsf{At}}$ *formulas in* $\overline{\mathsf{M}}$ *is given by* $\overline{\mathsf{M}}, \omega \vDash \top$ *for all* $\omega \in \Omega$;

$$\begin{array}{llllll}
\overline{\mathsf{M}}, \omega \vDash p & \textit{iff} & \omega \in v(p); & \overline{\mathsf{M}}, \omega \vDash a_i\varphi & \textit{iff} & S_{\Pi_i(\omega)} \succeq S([\varphi]); \\
\overline{\mathsf{M}}, \omega \vDash \neg\varphi & \textit{iff} & \omega \in \neg[\varphi]; & \overline{\mathsf{M}}, \omega \vDash \ell_i\varphi & \textit{iff} & \Lambda_i(\omega) \subseteq [\varphi]; \\
\overline{\mathsf{M}}, \omega \vDash \varphi \wedge \psi & \textit{iff} & \omega \in [\varphi] \cap [\psi]; & \overline{\mathsf{M}}, \omega \vDash k_i\varphi & \textit{iff} & \Pi_i(\omega) \subseteq [\varphi];
\end{array}$$

where $[\varphi] := \{\omega' \in \Omega : \overline{\mathsf{M}}, \omega' \vDash \varphi\}$ *for all* $\varphi \in \mathscr{L}_{\mathsf{At}}$.

In HMS models, formulae may have undefined truth value since formulae may not be even defined in every state. The same happens in FH models of a category of FH models. For instance, the truth value of p is not defined for all FH models K_Φ with $\Phi \not\ni p$. We will return to this issue later when we prove soundness and completeness.

Recall that for all $p \in At, v(p)$ is an event, so $[p]$ is an event in Σ. Negation and intersection of events are events. By Lemmata 1 and 7, explicit knowledge, awareness, and implicit knowledge of events are also events, respectively. Thus, for every $\varphi \in \mathscr{L}_{\mathsf{At}}$, $[\varphi]$ is an event.

Proposition 4 shows that in complemented HMS models, $K_i(E) = L_i(E) \cap A_i(E)$, for any event $E \in \Sigma$, so the semantics of $\mathscr{L}_{\mathsf{At}}$ provided above immediately implies that:

Proposition 9 *For any complemented HMS model* $\overline{\mathsf{M}}$, $\omega \in \Omega$, $\varphi \in \mathscr{L}_{\mathsf{At}}$, *and* $\Psi \subseteq \mathsf{At}$ *with* $\mathsf{At}(\varphi) \subseteq \Psi$, $\overline{\mathsf{M}}, \omega_\Psi \vDash k_i\varphi \leftrightarrow (\ell_i\varphi \wedge a_i\varphi)$.

An FH model and its HMS transform satisfy the same formulas in the language $\mathscr{L}_{\mathsf{At}}$ with implicit knowledge, explicit knowledge, and awareness as long as these formulas are defined at the corresponding states of the HMS transform.

Proposition 10 *For any FH model K_{At} and its HMS transform $HMS(\mathsf{K}_{\mathsf{At}})$, for all $w \in W_{\mathsf{At}}$, $\varphi \in \mathscr{L}_{\mathsf{At}}$, and $\Phi \subseteq \mathsf{At}$ with $\mathsf{At}(\varphi) \subseteq \Phi$, $\mathsf{K}_{\mathsf{At}}, w \Vdash \varphi$ if and only if $HMS(\mathsf{K}_{\mathsf{At}}), w_\Phi \vDash \varphi$.*

We now show that any complemented HMS model and its FH transform satisfy the same formulas from the language $\mathscr{L}_{\mathsf{At}}$ with implicit knowledge, explicit knowledge, and awareness.

Proposition 11 *For any complemented HMS model $\overline{\mathsf{M}}$ and its FH transform $FH(\overline{\mathsf{M}})$, for all $\varphi \in \mathscr{L}_{\mathsf{At}}$ and all $\omega \in S_{At}$, $\overline{\mathsf{M}}, \omega \vDash \varphi$ if and only if $FH(\overline{\mathsf{M}}), \omega \Vdash \varphi$.*

7 Implicit Knowledge-based HMS Models and FH Models

In this section, we focus on the relationship between implicit knowledge-based HMS and FH models. This relationship is even simpler than between HMS and FH models since implicit knowledge-based HMS models are arguably already closer to FH models. This is due to taking implicit knowledge and the awareness functions as primitives.

7.1 From FH Models to Implicit Knowledge-based HMS Models

We can define a version of HMS transformation that is "truncated" after the T-transformation. It just keeps the first two steps of the HMS transformation.

Definition 17 *For any FH model K_{At}, the* truncated HMS transform *$HMS^*(\mathsf{K}_{\mathsf{At}}) = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda^*_i)_{i \in I}, (\alpha^*_i)_{i \in I}, v\rangle$ is defined the first two steps of the HMS transform.*

From Proposition 7 follows now immediately:

Corollary 3 *For any FH model K_{At}, the truncated HMS-transform $HMS^*(\mathsf{K}_{\mathsf{At}})$ is an implicit knowledge-based HMS model.*

7.2 From Implicit Knowledge-based HMS Models to FH Models

Definition 18 *For any implicit knowledge-based HMS model $\overline{\mathsf{M}}^* = \langle I, \{S_\Phi\}_{\Phi \subseteq \mathsf{At}}, (r^\Phi_\Psi)_{\Psi \subseteq \Phi \subseteq \mathsf{At}}, (\Lambda_i)_{i \in I}, (\alpha_i)_{i \in I}, v\rangle$, the* FH*-transform *$FH^*(\overline{\mathsf{M}}^*) = \langle I, W_{\mathsf{At}}, (R_{\mathsf{At},i})_{i \in I}, (\mathscr{A}_{\mathsf{At},i})_{i \in I}, V_{\mathsf{At}}\rangle$ is defined like the FH transform except that the clause for the awareness correspondence is replaced by for any $i \in I$,*

- *$\mathscr{A}_{\mathsf{At},i} : W_{\mathsf{At}} \longrightarrow 2^{\mathscr{L}_{\mathsf{At}}}$ is such that $\mathscr{A}_{\mathsf{At},i}(\omega) = \mathscr{L}_\Phi$ for $\Phi \subseteq \mathsf{At}$ with $\alpha_i(\omega) = S_\Phi$.*

The FH* transform indeed transforms any implicit knowledge-based HMS model into a FH model.

Proposition 12 *For every implicit knowledge-based HMS model $\overline{\mathsf{M}}^*$, $FH^*(\overline{\mathsf{M}}^*)$ is a FH model for At.*

7.3 Equivalence of Implicit Knowledge-based HMS and FH Models

Definition 19 *Satisfaction of $\mathscr{L}_{\mathsf{At}}$ formulas in an implicit knowledge-based HMS model M^* is given like for complemented HMS models except that we have $\mathsf{M}^*, \omega \vDash a_i\varphi$ if and only if $\alpha_i(\omega) \succeq S([\varphi])$.*

An FH model and its truncated HMS transform satisfy the same formulas in the language $\mathscr{L}_{\mathsf{At}}$ with implicit knowledge, explicit knowledge, and awareness with the provision that these formulas are defined at the corresponding states of the implicit knowledge-based HMS transform. This follows directly from the proof of Proposition 10.

Corollary 4 *For any FH model* K_{At} *and its HMS transform* $HMS(\mathsf{K}_{\mathsf{At}})$, *for all* $w \in W_{\mathsf{At}}$, $\varphi \in \mathscr{L}_{\mathsf{At}}$, *and* $\Phi \subseteq \mathsf{At}$ *with* $\mathsf{At}(\varphi) \subseteq \Phi$, $\mathsf{K}_{\mathsf{At}}, w \Vdash \varphi$ *if and only if* $HMS(\mathsf{K}_{\mathsf{At}}), w_{\Phi} \models \varphi$.

Any implicit knowledge-based HMS model and its FH* transform satisfy the same formulas from the language $\mathscr{L}_{\mathsf{At}}$ with implicit knowledge, explicit knowledge, and awareness.

Proposition 13 *For any implicit knowledge-based HMS model* M^* *and its FH** *transform* $FH^*(\mathsf{M}^*)$, *for all* $\varphi \in \mathscr{L}_{\mathsf{At}}$ *and all* $\omega \in S_{At}$, $\mathsf{M}^*, \omega \models \varphi$ *if and only if* $FH^*(\mathsf{M}^*), \omega \Vdash \varphi$.

8 Logic of Propositional Awareness

In the penultimate section, we explore the implications of the prior sections for axiomatizations of both the category of FH models and the complemented HMS models. In particular, we show that the Logic of Propositional Awareness is sound and complete with respect to the class of complemented HMS models. This is the first axiomatization of HMS models that feature also the notion of implicit knowledge. Previous axiomatizations of HMS models (Heifetz, Meier, and Schipper, 2008, Halpern and Rêgo, 2008) were confined to explicit knowledge and awareness only. We also show that the Logic of Propositional Awareness is sound and complete with respect to the class of implicit knowledge-based HMS models. This is the first axiomatization of implicit knowledge-based HMS models. Finally, it is also sound and complete with respect to the class of *categories* of FH models.

Definition 20 *The logic LPA is the smallest set of* $\mathscr{L}_{\mathsf{At}}$ *formulas that contains the axioms and is closed under the inference rules as follows: All substitution instances of propositional logic, including* $\top$

$(\ell_i\varphi \wedge (\ell_i\varphi \rightarrow \ell_i\psi)) \rightarrow \ell_i\psi$
$k_i\varphi \leftrightarrow (\ell_i\varphi \wedge a_i\varphi)$
$a_i(\varphi \wedge \psi) \leftrightarrow (a_i\varphi \wedge a_i\psi)$
$a_i\neg\varphi \leftrightarrow a_i\varphi$
$a_ik_j\varphi \leftrightarrow a_i\varphi$
$a_ia_j\varphi \leftrightarrow a_i\varphi$
$a_i\ell_j\varphi \leftrightarrow a_i\varphi$
$a_i\varphi \rightarrow \ell_ia_i\varphi$
$\neg a_i\varphi \rightarrow \ell_i\neg a_i\varphi$
From φ *and* $\varphi \rightarrow \psi$, *infer* ψ
From φ *infer* $\ell_i\varphi$
$\ell_i\varphi \rightarrow \varphi$
$\ell_i\varphi \rightarrow \ell_i\ell_i\varphi$
$\neg\ell_i\varphi \rightarrow \ell_i\neg\ell_i\varphi$

Recall that in a Kripke model or FH model, a formula is valid if it is true in every state. However, a formula $\varphi \in \mathscr{L}_{\mathsf{At}}$ may not even be defined at states of the FH model K_{Ψ} with $\mathsf{At}(\varphi) \nsubseteq \Psi$. Similarly, as we remarked earlier when introducing the semantics for HMS models, a formula may not be defined in every state of a HMS model. We say that φ is defined in state ω in the complemented HMS model $\overline{\mathsf{M}}$ if $\omega \in \bigcap_{p \in \mathsf{At}(\varphi)}(v(p) \cup \neg v(p))$ (and analogously for implicit knowledge-based HMS models). Similarly, we say that φ is defined in the FH model K_{Ψ} if $\mathsf{At}(\varphi) \subseteq \Psi$.

Now we say that a formula φ is valid in the complemented HMS model $\overline{M}$ if $\overline{M}, \omega \models \varphi$ for all ω in which φ is defined (and analogously for the implicit knowledge-based HMS model). Similarly, we say that φ is valid in the category of FH models $\mathscr{C}(\mathsf{K}_{\Psi})$ if $\mathsf{K}_{\Psi}, w \Vdash \varphi$ for all $w \in W_{\Psi}$ for all K_{Ψ} in $\mathscr{C}(\mathsf{K}_{\mathsf{At}})$ for which φ is defined. A formula is valid in a class of complemented HMS models $\mathscr{M}$ if it is valid in every complemented HMS model of the class (and analogously for the class of implicit knowledge-based HMS models). A formula is valid in a class of categories of FH models $\mathfrak{C}$ if it is valid in every category of FH models of the class.

A proof in an axiom system consists of a sequence of formulae, where each formula in the sequence is either an axiom in the axiom system or follows from the prior formula in the sequence by an application of an inference rule of the axiom system. A proof of a formula φ is a proof where the last formula of the sequence is φ. A formula φ is provable in an axiom system, if there is a proof of φ in the axiom system. An axiom system is sound for the language $\mathscr{L}_{\mathsf{At}}$ w.r.t. a class of complemented HMS models $\mathscr{M}$ if every formula in $\mathscr{L}_{\mathsf{At}}$ that is provable in the axiom system is valid in every complemented HMS model of the class $\mathscr{M}$ (and analogously for the class of implicit knowledge-based HMS models). Similarly, an axiom system is sound for the language $\mathscr{L}_{\mathsf{At}}$ w.r.t. a class of categories of FH models $\mathfrak{C}$ if every formula in $\mathscr{L}_{\mathsf{At}}$ that is provable in the axiom system is valid in every category of FH models of the class $\mathfrak{C}$. An axiom system is complete for the language $\mathscr{L}_{\mathsf{At}}$ w.r.t. a class of complemented HMS models $\mathscr{M}$ if every formula in $\mathscr{L}_{\mathsf{At}}$ that is valid in $\mathscr{M}$ is provable in the axiom system (and analogously for the class of implicit knowledge-based HMS models). Similarly, an axiom system is complete for the language $\mathscr{L}_{\mathsf{At}}$ w.r.t. a class of categories of FH models $\mathfrak{C}$ if every formula in $\mathscr{L}_{\mathsf{At}}$ that is valid in $\mathfrak{C}$ is provable in the axiom system.

Corollary 5 *LPA is sound and complete w.r.t.*

1. *the class of categories of FH models,*
2. *the class of complemented HMS models,*
3. *the class of implicit knowledge-based HMS models.*

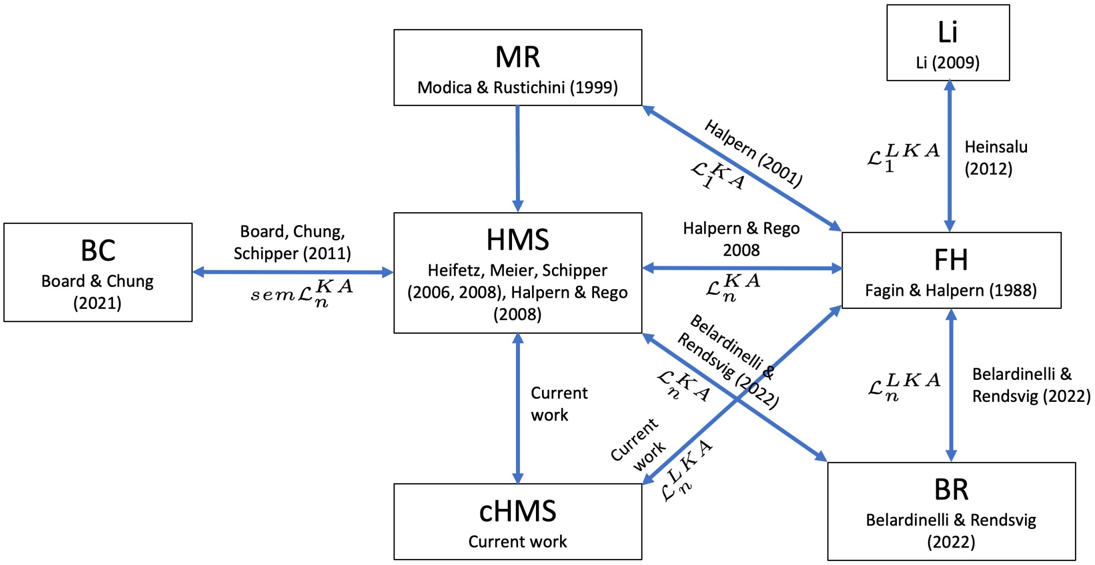
Fagin and Halpern (1988), Halpern (2001), and Halpern and Rêgo (2008) claim that LPA is sound and complete w.r.t. the class of FH models. The proof of 1. now follows from invariance of modal satisfaction relative to sublanguages between FH models in each category of FH models, i.e., Proposition 6. The proof of 2. follows from Propositions 10 and 11. The proof of 3. follows from Corollary 4 and Proposition 13.

9 Discussion

The constructions also allowed us to consider the relation between FH models and HMS models, not just with respect to explicit knowledge and awareness as in the prior literature but also with respect to implicit knowledge. We show modal equivalence between FH and HMS models by transforming one model into another. Each model and its transform satisfy the same formulae from a language of implicit, explicit knowledge and awareness. This equivalence is used to show that the Logic of Propositional Awareness is sound and complete for the class of HMS models. Compared to the prior literature, this axiomatization is now for a language that also features implicit knowledge.

The relations between various models of awareness in the literature are depicted in Figure 2. Beside FH models of Fagin and Halpern (1988) and HMS models of Heifetz, Meier, and Schipper (2006, 2008), we consider generalized standard models by Modica and Rustichini (1999), information structures with unawareness by Li (2009), object-based unawareness models by Board and Chung (2021), and Kripke lattices by Belardinelli and Rendsvig (2022). Equivalences hold for various languages also shown in the figure. We indicate the implicit, explicit, and awareness modality by superscripts L, K, and A, respectively. Some structures like Modica and Rustichini (1999) and Li (2009) feature just a single agent. We indicate this with the subscript "1" for single agent and "n" for multiple agents. For instance, $\mathscr{L}_n^{L,K,A}$ is the language featuring multiple agents, implicit knowledge, explicit knowledge, and awareness. The equivalence between Board and Chung (2021) is shown only at the level of semantics, i.e., at the level of events. The relation between Modica and Rustichini (1999) and Heifetz, Meier, and Schipper (2008) indicates that latter axiomatization can be seen as a multi-agent version of former. All shown relations pertain to rich structures featuring partitional knowledge and awareness generated by primitive propositions.

Figure 2: Relations between Approaches to Awareness



Recently, Schipper (2022) extended HMS models to awareness of unawareness by introducing quantified events. It would be straightforward to complement his structure with implicit knowledge as defined in the current work. Agents could then reason about the existence of their implicit knowledge that they are not aware of. Such reasoning bears some similarity to the notion of speculative knowledge in van Ditmarsch et al. (2018). Awareness-of-unawareness structures with implicit knowledge would also allow for a better comparison to awareness structures with quantification of formulae for modeling reasoning about knowledge of unawareness (Halpern and Rêgo, 2009, 2012), object-based unawareness (Board and Chung, 2021), and quantified neighborhood structures with awareness (Sillari, 2008).

References

[1] G. Belardinelli & K.R. Rendsvig (2022): *Awareness logic: Kripke lattices as a middle ground between syntactic and semantic models*. *Journal of Logic and Computation*:exac009, doi:10.1093/logcom/exac009.

[2] P. Blackburn, M., de Rijke & Y. Venema (2001): *Modal logic*. Cambridge University Press.

[3] O. Board & K.S. Chung (2021): *Object-based unawareness: Axioms*. *Journal of Mechanism and Institutional Design* 6, pp. 1–36, doi:10.22574/jmid.2021.12.001.

[4] O. Board, K.S. Chung & B.C. Schipper (2011): *Two models of unawareness: Comparing the object-based and subjective-state-space approaches*. *Synthese* 179, pp. 13–34, doi:10.1007/s11229-010-9850-z.

[5] R. Fagin & J. Halpern (1988): *Belief, awareness, and limited reasoning*. *Artificial Intelligence* 34, pp. 39–76, doi:10.1016/0004-3702(87)90003-8.

[6] R. Fagin, J. Halpern & M. Vardi (1986): *What can machines know? On the epistemic properties of machines*. *AAAI'86: Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, pp. 428–434, `https://dl.acm.org/doi/abs/10.5555/2887770.2887842`.

[7] R. Fagin, J. Halpern, Y. Moses & M. Vardi (1995): *Reasoning about knowledge*, MIT Press.

[8] S. Galanis (2013): *Unawareness of theorems*. *Economic Theory* 52, pp. 41–73, doi:10.1007/s00199-011-0683-x.

[9] S. Galanis (2011): *Syntactic foundations of unawareness of theorems*. Theory and Decision 71, pp. 593–614, doi:10.1007/s11238-010-9218-3.

[10] J. Halpern (2001): *Alternative Semantics for Unawareness*. Games and Economic Behavior 37, pp. 321–339, doi:10.1006/game.2000.0832.

[11] J. Halpern & L.C. Rêgo (2008): *Interactive unawareness revisited*. Games and Economic Behavior 62, pp. 232-–262, doi:10.1016/j.geb.2007.01.012.

[12] J. Halpern & L.C. Rêgo (2009): *Reasoning about knowledge of unawareness*. Games and Economic Behavior 67, pp. 503–525, doi:10.1016/j.geb.2009.02.001.

[13] J. Halpern & L.C. Rêgo (2012): *Reasoning about knowledge of unawareness revisited*. Mathematical Social Sciences 65, pp. 73–84, doi:10.1016/j.mathsocsci.2012.08.003.

[14] A. Heifetz, M. Meier & B.C. Schipper (2006): *Interactive unawareness*. Journal of Economic Theory 130, pp. 78–94, doi:10.1016/j.jet.2005.02.007.

[15] A. Heifetz, M. Meier, & B.C. Schipper (2008): *A canonical model for interactive unawareness*. Games and Economic Behavior 62, pp. 305–324, doi:10.1016/j.geb.2007.07.003.

[16] S. Heinsalu (2012): *Equivalence of the information structure with unawareness to the logic of awareness*. Journal of Economic Theory 147, pp. 2453–2468, doi:10.1016/j.jet.2012.05.010.

[17] J. Hintikka (1975): *Impossible possible worlds vindicated*. Journal of Philosophical Logic 4, 475–484, `https://www.jstor.org/stable/30226996`.

[18] G. Lakemeyer (1986): *Steps towards a first-order logic of explicit and implicit belief*. Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference, pp. 325–340, doi:10.1016/B978-0-934613-04-0.50027-2.

[19] H.J. Levesque (1984): *A logic of implicit and explicit belief*. AAAI'84: Proceedings of the Fourth AAAI Conference on Artificial Intelligence, pp. 198–202, `https://dl.acm.org/doi/proceedings/10.5555/2886937`.

[20] J. Li (2009): *Information structures with unawareness*. Journal of Economic Theory 144, pp. 977–993, doi:10.1016/j.jet.2008.10.001.

[21] E. Lorini (2020): *Rethinking epistemic logic with belief bases*. Artificial Intelligence 282:203233, doi:10.1016/j.artint.2020.103233.

[22] S. Modica & A. Rustichini (1999): *Unawareness and partitional information structures*. Games and Economic Behavior 27, pp. 265–298, doi:10.1006/game.1998.0666.

[23] R. Stalnaker (1991): *The problem of logical omniscience. I*. Synthese 89, pp. 425–440, `https://www.jstor.org/stable/20116982`.

[24] B.C. Schipper (2022): *Interactive awareness of unawareness*. University of California, Davis, `https://faculty.econ.ucdavis.edu/faculty/schipper/awunaw.pdf`.

[25] B.C. Schipper (2015): *Awareness*. In H. van Ditmarsch, J. Halpern, W. van der Hoek & B. Kooi (editors): Handbook of epistemic logic, College Publications, London, pp. 77–146, `https://faculty.econ.ucdavis.edu/faculty/schipper/unawhb.pdf`.

[26] G. Sillari (2008): *Quantified logic of awareness and impossible possible worlds*. Review of Symbolic Logic 1, pp. 514–529, doi:10.1017/S1755020308090072.

[27] H. van Ditmarsch, T. French, F. Velázquez-Quesada & Y.N. Wáng (2018): *Implicit, explicit and speculative knowledge*. Artificial Intelligence 256, pp. 35–67, doi:10.1016/j.artint.2017.11.004.

[28] F. Velázquez-Quesada (2013): *Explicit and implicit knowledge in neighbourhood models*. In D. Grossi, O. Roy & H. Huang (editors): Logic, rationality, and interaction: LORI 2013, Springer, Berlin, pp. 239–252, doi:10.1007/978-3-642-40948-6_19.

Epistemic Logics of Structured Intensional Groups

Marta Bílková   Igor Sedlár[*]
Institute of Computer Science of the Czech Academy of Sciences
Prague, Czech Republic
{bilkova, sedlar}@cs.cas.cz

Epistemic logics of intensional groups lift the assumption that membership in a group of agents is common knowledge. Instead of being represented directly as a set of agents, intensional groups are represented by a property that may change its extension from world to world. Several authors have considered versions of the intensional group framework where group-specifying properties are articulated using structured terms of a language, such as the language of Boolean algebras or of description logic. In this paper we formulate a general semantic framework for epistemic logics of structured intensional groups, develop the basic theory leading to completeness-via-canonicity results, and show that several frameworks presented in the literature correspond to special cases of the general framework.

1 Introduction

One of the usual assumptions of multi-agent epistemic logic is that groups of agents are given *extensionally* as sets of agents. Membership in an extensional group is common knowledge among all agents and change in membership implies change of identity of an extensional group. This is not how we usually think of groups, however. We are commonly reasoning about groups in various contexts without knowing their extensions—we might routinely refer to groups such as "bot accounts", "democrats", or "correct processes"—and we do not settle for reducing such groups to their extensions either, as clearly they can change across the state space of a system, or possible states of the world. To reason about groups in a more realistic way is made possible by groups being given to us *intensionally* by a common property.

In their seminal work [5, 6], Grove and Halpern introduced an elegant generalization of multi-agent epistemic logic where labels denoting (sets of) agents are replaced by abstract *names* whose extensions can vary from world to world. Their language contains two types of modalities, equipped with a relational Kripke-style semantics: $E_n\varphi$ means that every agent in the current extension of n knows that φ ("everyone named n knows"), and $S_n\varphi$ means that some agent in the current extension of n knows that φ ("someone named n knows"). In the intensional setting, S_n is in general not definable using disjunction and other epistemic operators. Grove and Halpern also consider a natural extension of their basic framework where names are replaced by formulas expressing *structured* group-defining concepts.

Motivated mainly by applications such as dynamic networks of processes, a framework where the agent set can vary not only across models, but also from state to state, have been developed in a form of term-modal logic TML. Introduced by [4], TML builds upon first order logic, indexing modalities by terms that can be quantified over. TML is conveniently expressive but undecidable in general, and the attention therefore turns to identify some decidable fragments ([16, 17] (see [19] for more references relevant for epistemic logic). Epistemic logic with names of [6] was seminal in some sense to the development of TML, and can be seen as its simple decidable fragment. Another closely related language is that of implicitly quantified modal logic, studied in [15].

[*] Work of both authors on this paper was supported by the grant no. 22-23022L CELIA of the Czech Science Foundation.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 113–130, doi:10.4204/EPTCS.379.11

To model non-rigididity of group names, Kooi [10] introduces dynamic term-modal logic with assignment modalities. Wang and Seligman [20] adopt a minimalist approach of using the basic assignment modalities with a quantifier-free term modal logic to obtain an easy-to-handle fragment of the logic in [10], expressing various de re/de dicto distinctions in reading higher-order knowledge[1].

Grove and Halpern's work is enjoying a recent resurgence of interest in the epistemic logic community. [2] identifies a monotone neighborhood-style semantics for Grove and Halpern's language and, building directly on the S_n and E_n modalities, considers expansions with non-rigid versions of common and distributed knowledge. Distributed or common knowledge for intensional group names has also been studied by [12, 13]. A monotone neighborhood perspective has recently been adopted by [3] and applied to a logic containing the somebody-knows modality of [1]. Humml and Schröder [9] generalize Grove and Halpern's approach to structured names represented by formulas defining group membership, including e.g. formulas of the description logic ALC. Their abstract-group epistemic logic (AGEL) contains a common knowledge modality as the only modality and, unlike in [2, 6], their group names are rigid.

In this paper, following [2, 6, 15] to various extent, we adopt the perspective that both "everyone labeled *a* knows" and "someone labeled *a* knows" form a minimal epistemic language for group knowledge, that groups are understood intensionally, and that labels reflect their structured nature. We use languages built on top of classical propositional language containing modalities $[a]$ and $\langle a]$ indexed by elements of an algebra of a given signature of interest. As our main contribution, we set up a general framework for epistemic logics for structured groups in terms of relational semantics involving an algebra of group labels that index (sets of) relations in each world (Section 2), we show how some existing versions of frame semantics of closely related logics can be modelled in such a way, and then generalize relational frames in terms of two-sorted algebras involving propositions and groups, develop an algebraic duality and prove completeness of the minimal logic (Subsection 2.1). We show that the semantics can be seen as an interesting version of monotone neighborhood frame semantics (Subsection 2.2). In the remaining part of the paper we discuss several examples of algebraic signatures giving rise to interesting and useful variants of group structure (Section 3).

2 Frame semantics for structured groups

Definition 1 (**Relational frame).** *Let Σ be an algebraic similarity type. A Σ-algebra is any structure of the form $\mathbf{X} = (X, \{o^{\mathbf{X}} \mid o \in \Sigma\})$, where each $o^{\mathbf{X}}$ is an n-ary operator on X for some n. A* relational Σ-frame *is $\mathfrak{F} = (W, R, \mathbf{G})$, where $W \neq \emptyset$ ("worlds"); $R \subseteq 2^{W \times W}$ ("agent relations"); and $\mathbf{G}$ is a Σ-algebra with universe $G \subseteq (2^R)^W$ ("group intensions").*

In a relational Σ-frame, the set of available agents is represented by a set of accessibility relations R. Functions $f \in G$ map possible worlds $w \in W$ to sets $f(w) \subseteq R$ corresponding to sets of agents. These functions can be seen as *intensions* of properties of agents: the intension f of a given property determines for each world w the *extension* $f(w)$ of the property at w, representing the set of agents that possess the given property in w. Crucially, properties may change their extensions from world to world.

Remark 1. We note that a relational Σ-frame can be seen as a Σ-algebra over a subset of a direct product of a family of Kripke frames. In particular, $G \subseteq \prod_{w \in W}(W, Q_w)$ where $Q_w \subseteq R$; every $\mathfrak{F}$ then gives rise to $\mathfrak{G}(\mathfrak{F}) = (G, \{o^{\mathfrak{G}(\mathfrak{F})} \mid o \in \Sigma\})$. Conversely, every $\mathfrak{G} = (G, \{o^{\mathfrak{G}} \mid o \in \Sigma\})$ where $G \subseteq \prod_{w \in W}(W, Q_w)$ such that $Q_w \subseteq R$ fo all $w \in W$ gives rise to a relational Σ-frame.

[1] Both ours and theirs formalisms implement term-indexed modalities in a two-sorted language, they are however languages with different expressive power—one algebraic, the other first-order—a precise comparison being a subject of future work.

Definition 2 (**Language**). *Let Pr, Gr be denumerable sets of propositional variables and group variables respectivelly. For each Σ, the* Σ-language *is two-sorted, consisting of group Σ-terms and Σ-formulas. The set of Σ-terms Tm_Σ, and the set of Σ-formulas Fm_Σ, are defined by the following grammars:*

$$Tm_\Sigma: \quad \alpha := \mathtt{a} \in Gr \mid o(\alpha_1, \ldots, \alpha_n) \qquad Fm_\Sigma: \quad \varphi := \mathtt{p} \in Pr \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathtt{[}\alpha\mathtt{]}\,\varphi \mid \langle\alpha\mathtt{]}\,\varphi.$$

Σ-terms represent structured intensional groups where the structure is articulated using the operators of Σ (number of examples follow). Formulas $\mathtt{[}\alpha\mathtt{]}\,\varphi$ read as "Everyone in the group (given by) α believes that φ" and $\langle\alpha\mathtt{]}\,\varphi$ read as "Someone in the group (given by) α believes that φ". We assume the standard definitions of Boolean operators ($\top, \bot, \vee, \rightarrow, \leftrightarrow$), and we define $\langle\alpha\rangle\,\varphi := \neg\mathtt{[}\alpha\mathtt{]}\neg\varphi$, $\mathtt{[}\alpha\rangle\,\varphi := \neg\langle\alpha\mathtt{]}\neg\varphi$.

Definition 3 (**Complex algebra**). *The* complex algebra *of $\mathfrak{F}$ is $\mathfrak{F}^+ = (\mathbf{F}, \mathbf{G}, \mathtt{[]}^+, \langle\mathtt{]}^+)$ where $\mathbf{F}$ is the Boolean algebra of (all) subsets of W and $\mathtt{[]}^+$, $\langle\mathtt{]}^+$ are functions of the type $2^W \times \mathbf{G} \to 2^W$ such that for $a \in G$ and $P \subseteq W$:*

$$\mathtt{[}a\mathtt{]}^+ P = \{w \mid \forall r \in a(w) : r(w) \subseteq P\} \qquad \langle a\mathtt{]}^+ P = \{w \mid \exists r \in a(w) : r(w) \subseteq P\}$$

(where $r(w) = \{u \mid (w,u) \in r\}$).

Definition 4 (**Relational model**). *A* model *based on a Σ-frame $\mathfrak{F} = (W, R, \mathbf{G})$ (Σ-model) is $\mathfrak{M} = (\mathfrak{F}, [\![\,]\!])$, where $[\![\,]\!]$ (the "interpretation function") is a homomorphism from $Tm_\Sigma \cup Fm_\Sigma$ to $\mathfrak{F}^{+2}$, that is,*

- $[\![\alpha]\!] \in \mathbf{G}$ *where* $[\![o(\alpha_1, \ldots, \alpha_n)]\!] = o^{\mathbf{G}}([\![\alpha_1]\!], \ldots, [\![\alpha_n]\!])$*;*
- $[\![\varphi]\!] \subseteq W$ *where*

$$[\![\neg\varphi]\!] = W \setminus [\![\varphi]\!] \quad [\![\varphi \wedge \psi]\!] = [\![\varphi]\!] \cap [\![\psi]\!] \quad [\![\mathtt{[}\alpha\mathtt{]}\,\varphi]\!] = \mathtt{[}[\![\alpha]\!]\mathtt{]}^+[\![\varphi]\!] \quad [\![\langle\alpha\mathtt{]}\,\varphi]\!] = \langle[\![\alpha]\!]\mathtt{]}^+[\![\varphi]\!].$$

A formula φ is valid *in a model $\mathfrak{M}$ iff $[\![\varphi]\!]_{\mathfrak{M}} = W_{\mathfrak{M}}$, and valid in a class of frames iff it is valid in each model based on a frame in the given class. If K is a class of frames, then $Log(\mathsf{K})$ is the set of formulas valid in all frames in K.*

Example 1. Consider a relational frame for epistemic logic with names [2]. Let N ("names"), A ("agents") and W ("worlds") be three non-empty sets. A *relational frame* is (W, A, N, Q, μ), where $Q : A \to 2^{W \times W}$ and $\mu : N \to (W \to 2^A)$. It is easy to see that each relational frame gives rise to a relational $\emptyset$-frame where $R = \{Q_i \mid i \in A\}$ and $G = \{\mu^\#(n) \mid n \in N\}$, where $\mu^\#(n)(w) = \{Q_i \mid i \in \mu(n)(w)\}$. Conversely, every relational $\emptyset$-frame can be seen as a relational frame where $A = R$, Q is the identity function on A, $N = G$ and $\mu(a)(w) = a(w)$ for all $a \in G$.

Example 2. Grove and Halpern [6] consider a version of their framework where groups are referred to by means of formulas of a Boolean language. A simplified version of this framework can be presented as an extension of the relational frames of the previous example. In these frames we require that N is a term algebra over terms in the signature $\Sigma_{\mathsf{BA}} = \{\bar{\ }, \wedge, \vee\}$, and that μ satisfies the following conditions (we use n, m as variables ranging over Σ_{BA}-term to highlight the relation to Grove and Halpern's framework):

$$\mu(\bar{n}, w) = W \setminus \mu(n, w) \qquad \mu(n \wedge m, w) = \mu(n, w) \cap \mu(m, w) \qquad \mu(n \vee m, w) = \mu(n, w) \cup \mu(m, w).$$

It is easy to see that every relational frame of this kind (Boolean relational frame) gives rise to a relational Σ_{BA}-frame where R and G are defined as in the previous example. Conversely, every relational Σ_{BA}-*model* gives rise to a Boolean relational model: $A = R$, Q is the identity function on A, N is the term algebra over Σ_{BA}-terms and $\mu(n) = [\![n]\!]$. The semantic clauses displayed above then follow from the assumption that the interpretation function $[\![\,]\!]$ is a homomorphism.

[2] Being a homomorphism, $[\![\cdot]\!]$ is determined by the values it assigns to variables.

Example 3. In their recent work [9] on logic with common knowledge of abstract groups AGEL, Humml and Schröder consider a rigid common knowledge operator for groups with membership defined by formulas. Technically, the common knowledge modality is labeled by formulas in an agent language $\mathscr{L}_{Ag}$ built over a fixed set Ag of agents, defining groups of agents by semantical means of an agent model A. A formula $C_\alpha \phi$ reads as ϕ is commonly known among agents satisfying α. The language is interpreted over AGEL frames of the form $(W, A, \sim)$ where W is a set of worlds, and $\sim$ is a set of agent epistemic indistinguishability relations. In the sense of this paper, their agent language $\mathscr{L}_{Ag}$ determines a signature Σ, and the complex algebra $\mathbf{A}$ of the agent model A, i.e. the algebra on group propositions $\{[\![\alpha]\!]_A \subseteq Ag \mid \alpha \in \mathscr{L}_{Ag}\}$, is a Σ-algebra. As the agent language conservatively extends classical propositional logic, this algebra carries a boolean structure. It gives rise to a Σ-relational frame where $R = \sim$ and the Σ-algebra $\mathbf{G}$ is determined by $\mathbf{A}$ on the universe consisting of assignments $g : W \to 2^R$, with $g(w) = \{\sim_{[\![\alpha]\!]_A} \mid \alpha \in \mathscr{L}_{Ag}\}$ for each $w \in W$, where $\sim_{[\![\alpha]\!]_A}$ is the union of relations of agents satisfying α.

Remark 2. Our framework covers also semantics of modal logics with operations on accessibility relations. A prominent example are models for (test-free) Propositional Dynamic Logic. A relational PDL-model corresponds to a relational Σ_{KA}-model, where $\Sigma_{\mathsf{KA}} = \{\cdot, +, {}^*, 1, 0\}$ is the signature of Kleene algebra, such that R is the set of all relations on a set of worlds W and the functions in G are constant and their values are singletons. In particular, $\mathbf{G}$ is the algebra of constant functions $f \in (2^R)^W$ such that $f(w)$ is a singleton (therefore we may identify f with the r such that $f(w) = \{r\}$) and $f \cdot g$ is relational composition of f and g, $f + g$ is the union of f and g, f^* is the reflexive transitive closure of f, 1 is the identity relation, and 0 is the empty relation.

Definition 5 (**Logic**)**.** *Let Σ be an algebraic signature. An* epistemic logic with structured intensional groups over Σ *(or simply a* Σ-logic*) is any set $L \subseteq Fm_\Sigma$ such that (for all $\alpha \in Tm_\Sigma$)*

1. *L contains all substitution instances of classical tautologies and is closed under Modus Ponens;*
2. *L contains all formulas of the form* (K) $[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi)$ *and is closed under the Necessitation rule* (Nec) $\dfrac{\varphi}{[\alpha]\varphi}$*;*
3. *L contains all formulas of the form* $\neg[\alpha]\bot \to \langle\alpha]\top$ *and* $\langle\alpha]\varphi \wedge [\alpha]\psi \to \langle\alpha](\varphi \wedge \psi)$[3].

2.1 Algebraic duality

In this section we introduce specific two-sorted algebras that generalize relational Σ-frames. In a sense to be specified below, completeness results for classes of relational Σ-frames correspond to specific representation results for these two-sorted algebras.

Definition 6 (**Frame**)**.** *Let Σ be an algebraic similarity type. A* Σ-frame *is* $\mathfrak{A} = (\mathbf{F}, \mathbf{G}, [], \langle])$*, where* $\mathbf{F} = (X, \wedge, \vee, \neg, \top, \bot)$ *is a Boolean algebra;* $\mathbf{G} = (A, \{o^{\mathbf{G}} \mid o \in \Sigma\})$ *is a Σ-type algebra; and* $[]$ *and* $\langle]$ *are functions of the type* $\mathbf{F} \times \mathbf{G} \to \mathbf{F}$ *such that*

$$[a]\top = \top \qquad (1) \qquad\qquad \neg[a]\bot \leq \langle a]\top \qquad (3)$$

$$[a](x \wedge y) = [a]x \wedge [a]y \qquad (2) \qquad\qquad \langle a]x \wedge [a]y \leq \langle a](x \wedge y) \qquad (4)$$

[3] It is easy to show that every Σ-logic contains all formulas of the form $[\alpha]\top \leftrightarrow \top$ and $[\alpha](\varphi \wedge \psi) \leftrightarrow ([\alpha]\varphi \wedge [\alpha]\psi)$, and that it is closed under the rule

$$\frac{\varphi \wedge \psi_1 \wedge \ldots \wedge \psi_n \to \chi}{\langle\varphi] \wedge [\alpha]\psi_1 \wedge \ldots \wedge [\alpha]\psi_n \to \langle\alpha]\chi}.$$

A Σ-frame is a two-sorted algebra bringing together a Boolean algebra of “propositions” with a Σ-algebra of “groups”. The modal operators `[]` and `⟨]`, resembling scalar multiplication in modules, take pairs consisting of a group and a proposition to a proposition. Formulas in Fm_Σ-can be seen as terms of the type corresponding to Σ-frames. In fact, we can define the following notion of an evaluation, leading to a natural definition of the equational theory of a class of Σ-frames.

Definition 7 (Equational theory). *An* evaluation *on a Σ-frame is any homomorphism $Tm_\Sigma \cup Fm_\sigma \to \mathfrak{A}$, that is, any function e such that $e(o(\varphi_1,\dots,\varphi_n)) = o^{\mathbf{F}}(e(\varphi_1),\dots,e(\varphi_n))$ for all Boolean operators o; $e(o(\alpha_1,\dots,\alpha_n)) = o^{\mathbf{G}}(e(\varphi_1),\dots,e(\varphi_n))$ for all Σ-operators o; and $e(\texttt{[}\alpha\texttt{]}\varphi) = \texttt{[}e(\alpha)\texttt{]}e(\varphi)$ and $e(\langle\alpha\texttt{]}\varphi) = \langle e(\alpha)\texttt{]}e(\varphi)$. A Σ-*formula equation *is an expression of the form $\varphi \approx \psi$ where $\varphi,\psi \in Fm_\Sigma$. An equation $\varphi \approx \psi$ is* valid in $\mathfrak{A}$ *iff $e(\varphi) = e(\psi)$ for all evaluations e on $\mathfrak{F}$. The* equational theory *of a class F of Σ-frames is the set of all Σ-formula equations that are valid in all frames in F, denoted as $Eq(\mathsf{F})$.*[4].

Remark 3. Dynamic algebras [11, 18], the algebraic counterparts of relational models for Propositional Dynamic Logic, are related to Σ-frames. A dynamic algebra is a pair $(\mathbf{F},\mathbf{G},\texttt{[]})$, where $\mathbf{F}$ is a Boolean algebra, $\mathbf{G}$ is a Kleene algebra, and $\texttt{[]} : \mathbf{G}\times\mathbf{F}\to\mathbf{F}$ satisfying our axioms (1–2) and further set of equations and quasi-equations. Therefore, dynamic algebras can be seen as a class of `⟨]`-free reducts of Σ_{KA}-frames.

Definition 8 (Ultrafilter frame). *Let $\mathfrak{A} = (\mathbf{F},\mathbf{G},\texttt{[]},\langle\texttt{]})$ be a Σ-frame. The* ultrafilter frame of $\mathfrak{A}$ *is $\mathfrak{A}_+ = (\mathrm{Uf}(\mathbf{F}), R_+, \mathbf{G}_+)$ where $\mathrm{Uf}(\mathbf{F})$ is the set of all ultrafilters on $\mathbf{F}$ (we define $\hat{x} = \{u \in \mathrm{Uf}(\mathbf{F}) \mid x \in u\}$);*

- $R_+ = \{r_{a,x} \mid x \in \mathbf{F} \;\&\; a \in \mathbf{G}\}$, *where* $\;r_{a,x} : w \mapsto \bigcap\{\hat{y} \mid \texttt{[}a\texttt{]}y \in w\} \cap \hat{x}$;
- $G_+ = \{G(a) \mid a \in \mathbf{G}\} \subseteq (2^{R_+})^{\mathrm{Uf}(\mathbf{F})}$ *such that* $\forall u \in \mathrm{Uf}(\mathbf{F}),\; G(a)(u) = \{r_{a,x} \mid \langle a\texttt{]}x \in u\}$ *(we will often write $G(a,u)$ instead of $G(a)(u)$);*
- $\mathbf{G}_+ = (G_+, \{o_+ \mid o \in O_\Sigma\})$ *where* $\;o_+(G(a_1),\dots,G(a_n))(u) = G(o(a_1,\dots,a_n))(u)$.

Definition 9 (Morphisms of Σ-frames). *Let $\mathfrak{A}_1,\mathfrak{A}_2$ be two Σ-frames. A (Σ-frame)* morphism *is a function $f : \mathfrak{A}_1 \to \mathfrak{A}_2$ such that*

(m1) f *is a homomorphism from* $\mathbf{F}_1$ *to* $\mathbf{F}_2$*;*

(m2) f *is a homomorphism from* $\mathbf{G}_1$ *to* $\mathbf{G}_2$*;*

(m3) $f(\texttt{[}a\texttt{]}_1 x) = \texttt{[}f(a)\texttt{]}_2 f(x)$*;*

(m4) $f(\langle a\texttt{]}_1 x) = \langle f(a)\texttt{]}_2 f(x)$*.*

A *quasi-embedding* of $\mathfrak{A}_1$ into $\mathfrak{A}_2$ is a morphism $f : \mathfrak{A}_1 \to \mathfrak{A}_2$ such that $f(x) = f(y) \to x = y$ for all x,y in $\mathbf{F}_1$. An *embedding* of $\mathfrak{A}_1$ into $\mathfrak{A}_2$ is a quasi-embedding where $f(a) = f(b) \to a = b$ for all a,b in $\mathbf{G}_1$. A *quasi-isomorphism* is a surjective quasi-embedding and an *isomorphism* is a surjective embedding. The *canonical embedding algebra* of $\mathfrak{A}$ is $(\mathfrak{A}_+)^+$ and the *ultrafilter extension* of $\mathfrak{F}$ is $(\mathfrak{F}^+)_+$. The *canonical morphism* is a function $f : \mathfrak{A} \to (\mathfrak{A}_+)^+$ with $f(x) = \hat{x}$ for $x \in \mathbf{F}$ and $f(a) = G(a)$ for $a \in \mathbf{G}$.

Lemma 1. *The canonical morphism is a quasi-embedding.*

For each signature Σ, Lemma 1 can be used to prove completeness of the basic Σ-logic with respect to all relational Σ-frames. In order to show this, we point out a useful example of a Σ-frame.

Example 4. Let L be a Σ-logic. Let $\equiv_L$ be a binary relation on Fm_Σ such that $\varphi \equiv_L \psi$ iff $\varphi \leftrightarrow \psi \in L$. Let $[\varphi]_L$ be the equivalence class of φ under $\equiv_L$. It can be shown that $\equiv_L$ is a congruence on Fm_Σ. Hence, we obtain the Boolean algebra $\mathbf{F}^L$ of equivalence classes $[\varphi]_L$, where $o^{\mathbf{F}^L}([\varphi_1]_L,\dots,[\varphi_n]_L) = [o(\varphi_1,\dots,\varphi_n)]_L$

[4] We note that it would make sense also to consider Σ-group equations as expressions of the form $\alpha \approx \beta$ where $\alpha,\beta \in Tm_\Sigma$, and define the group-equational theory of a class of frames, but we will not pursue this topic here.

for all Boolean operators o. We define $\mathbf{G}^L$ as the term algebra over Tm_Σ. Moreover, let $[\,]^L$ and $\langle\,]^L$ be functions of the type $\mathbf{F}^L \times \mathbf{G}^L \to \mathbf{F}^L$ such that $[\alpha]^L[\varphi]_L = [[\alpha]\,\varphi]_L$ and $\langle\alpha]^L[\varphi]_L = [\langle\alpha]\,\varphi]_L$ (note that these functions are well defined since $\equiv_L$ is a congruence). Let us define the *basic canonical L-frame* as $\mathfrak{B}^L = (\mathbf{F}^L, \mathbf{G}^L, [\,]^L, \langle\,]^L)$. It is clear that $\varphi \in L$ iff $\varphi \approx \top$ is valid in $\mathfrak{B}^L$.

Theorem 1 (Completeness). *For all Σ, the smallest Σ-logic is the set of Σ-formulas valid in all relational Σ-models.*

Proof. Fix a Σ and take the smallest Σ-logic L. Soundness is easily checked. To show completeness, take the relational Σ-frame $(\mathfrak{B}^L)_+$ (the canonical relational L-frame). Lemma 1 entails that if $\varphi \notin L$, then φ is not valid in $(\mathfrak{B}^L)_+$. (Define a model where $[\![\varphi]\!] = \widehat{[\varphi]_L}$ and $[\![\alpha]\!] = \alpha$. Lemma 1 implies that $[\![\,]\!]$ is indeed an interpretation function. Since $\varphi \leftrightarrow \top \notin L$, we have $[\![\varphi]\!] \neq [\![\top]\!]$ by the Prime Filter Theorem, and so φ is not valid in $(\mathfrak{B}^L)_+$.) □

2.2 Neighborhood semantics

The modalities $\langle\,]$ and $[\,]$ are monotone modalities of the $\exists\forall$ and $\forall\forall$ type and can therefore be studied in terms of monotone neighborhood semantics, if we understand sets $\{r(w) \mid r \in a(w)\}$ as so called core neighborhood sets [7, 14]. Relational Σ-frames generalize relational frames for epistemic logic with names [2, 6] (Example 1), which are categorialy equivalent to monotone neighborhood frames with neighborhood sets indexed by the set of names. Not surprisingly, a closely related connection arises between relational Σ-frames of this paper and monotone neighborhood frames where neighborhoods are indexed with algebraic terms. This will allow us to adapt and apply the well understood model theory of monotone neighborhood frames (for which we mainly refer to [7, 8, 14]) to study, among others, algebraic duality or modal definability on a convenient level of abstraction. A similar perspective has recently been adopted also by [3] on a logic containing a somebody-knows modality, previously studied by [1]. Neither of the approaches in [7, 3] however includes both $\exists\forall$ and $\forall\forall$ types of modalities, and therefore similar modifications of the general theory as those adopted in [2] are necessary, and the algebraic structure underlying the labelling of groups needs to be captured additionally.

Definition 10 (Neighborhood frames). *A neighborhood Σ-frame $\mathfrak{F}$ is a tuple $(W, \mathbf{G}, \{\nu_a\}_{a\in G})$ where W is a set of states, $\mathbf{G}$ is a Σ-algebra, and for each $a \in G$, $\nu_a : W \to 2^{2^W}$ is a neighborhood function that assigns to each state w a set of sets of states*[5].

Definition 11 (Semantics in neighborhood models). *The complex algebra $\mathfrak{F}^+$ of a neighborhood Σ-frame $\mathfrak{F}$ is given as the expansion of the boolean algebra of subsets of W by*

$$[a]^+P = \{w \mid \forall X \in \nu_a(w) : X \subseteq P\} \qquad \langle a]^+P = \{w \mid \exists X \in \nu_a(w) : X \subseteq P\}.$$

An interpretation function $[\![\,]\!]$ is a homomorphism from $Tm_\Sigma \cup Fm_\Sigma$ to $\mathfrak{F}^+$, i.e.

$$[\![[\alpha]\,\varphi]\!] = \{w \mid \forall X \in \nu_{[\![\alpha]\!]}(w)\; X \subseteq [\![\varphi]\!]\} \qquad [\![\langle\alpha]\,\varphi]\!] = \{w \mid \exists X \in \nu_{[\![\alpha]\!]}(w)\; X \subseteq [\![\varphi]\!]\}$$

Definition 12 (Neighborhood frame morphisms). *Neighborhood Σ-frame morphisms are pairs of maps $(g : \mathbf{G} \to \mathbf{G}', f : W \to W')$, where g is a homomorphism of Σ-algebras, satisfying*

$$\textit{(there)}\; X \in \nu_a(w) \Rightarrow f[X] \in \nu'_{g(a)}(f(w)) \qquad \textit{(back)}\; Y \in \nu'_{g(a)}(f(w)) \Rightarrow \exists X (f[X] = Y \;\&\; X \in \nu_a(w))$$

[5] For the minimal Σ-logic, we do not require any additional (algebraic) properties from the assignment $\nu_{\_} : \mathbf{G} \to [W, 2^{2^W}]$. They might however become desirable in the examples that follow, and we will treat them as additional properties defining particular classes of frames (modally definable or not).

Monotonicity of the modalities is built into the semantical definition rather than into the frame definition. As such, it corresponds to core neighborhood frames from [7], and the morphisms resemble core bounded morphisms of monotone neighborhood frames from [7, Definition 4.6], additionally involving the algebraic homomorphism $g : \mathbf{G} \to \mathbf{G}'$ which can be interpreted as allowing to "rename" the groups along frame morphisms in a structured way. Understanding frame validity as $\mathfrak{F}, w \Vdash \varphi$ if and only if $w \in [\![\varphi]\!]$ for each interpretation $[\![]\!]$ on $\mathfrak{F}$, we can prove that morphisms preserve frame validity:

Lemma 2 (Preservation of validity). *Let* $(f,g) : (W_1, \mathbf{G_1}, \{\nu_a\}_{a\in G_1}) \to (W_2, \mathbf{G_2}, \{\nu_a\}_{a\in G_2})$ *be a neighborhood* Σ*-frame morphism from* $\mathfrak{F}_1$ *to* $\mathfrak{F}_2$*. Then for each formula* φ *and each* $w \in W$*,*

$$\mathfrak{F}_1, w \Vdash \varphi \quad \Rightarrow \quad \mathfrak{F}_2, f(w) \Vdash \varphi.$$

A proof-sketch can be found in the Appendix A.2. For the sake of interest we also spell out in Appendix A.3 what bisimulations of neighborhood Σ-frames look like.

For a relational Σ-frame $\mathfrak{F} = (W, R, \mathbf{G})$, we can define the corresponding neighborhood Σ-frame $\mathfrak{F}^n = (W, \mathbf{G}, \{\nu^n_a\}_{a\in G})$ putting $\nu^n_a(w) = \{r(w) \mid r \in a(w)\}$. Conversely, for a neighborhood Σ-frame $\mathfrak{F} = (W, \mathbf{G}, \{\nu_a\}_{a\in G})$ we define the corresponding relational Σ-frame $\mathfrak{F}^r = (W, R^r, \mathbf{G})$ by $a^r(w) = \{r \mid r(w) \in \nu_a(w)\}$, $R^r = \bigcup_{a\in G, w\in W} a^r(w)$. We then obtain the following:

Theorem 2 (Categorial equivalence). *The categories of relational* Σ*-frames and neighborhood* Σ*-frames are equivalent.*

Given the completeness of the basic Σ-logic with respect to relational Σ-frames (Theorem 1), completeness with respect to all neighborhood Σ-frames follows[6] With the complex algebra/ultrafilter frame construction at hand, we can describe the algebraic duality, and obtain a definability theorem characterizing modally definable classes of neighborhood Σ-frames (cf. Theorem 2 of [2]).

3 Special cases

To illustrate some interesting special cases of the general framework discussed above, we introduce, in each case, a class of relational frames that captures some natural kind of structure imposed on intensional groups, provide an algebraic generalization of relational frames, and show that the respective classes of relational frames and their algebraic generalizations determine the same logic.

3.1 Unions and join-semilattices

One of the simplest forms of structure imposed on groups of agents corresponds to taking unions of sets of agents. On the intensional perspective, taking unions corresponds to an operation on intensional groups that, for each world w, gives the union of the extensions of the given intensional groups in w. It is then natural to impose a *semilattice* structure on the set of intensional groups, where the neutral element is an "inconsistent" intensional group that has an empty extension in each world. This case is also easily handled in the technical sense, and so we will discuss it as an introductory example.

Nevertheless, even this simple case has an interesting feature: the ultrafilter frame construction does not in general lead to a relational frame of the right kind. This may be surprising given the fact that unions are well-behaved in the extensional framework. This feature is discussed at the end of the section.

Definition 13 (JS-frame). *Let* $\Sigma_{\mathsf{SL}} = \{+, 0\}$ *be the join-semilattice signature. A* relational join-semilattice frame *(relational js-frame) is a relational* Σ_{SL}*-frame where* $0^{\mathbf{G}}(w) = \emptyset$ *and* $(f +^{\mathbf{G}} g)(w) = f(w) \cup g(w)$*. A* join-semilattice frame *(js-frame) is a* Σ_{SL}*-frame where* $\mathbf{G}$ *is a join semilattice and*

[6] It is also possible to define a canonical neighborhood Σ-frame directly, following similar pattern as in Definition 8.

$$[0]x = \top \qquad (5) \qquad\qquad [a+b]x = [a]x \wedge [b]x \qquad (7)$$

$$\langle 0]x = \bot \qquad (6) \qquad\qquad \langle a+b]x = \langle a]x \vee \langle b]x \qquad (8)$$

The class of (relational) js-frames will be denoted as FSL *(*rFSL*).*

Definition 14 (JS-logic). *The* join-semilattice logic *LSL is the smallest* Σ_{SL}*-logic that contains all formulas of the following forms:*

$$\top \to [0]\,\varphi \qquad (\text{a5}) \qquad\qquad [\alpha+\beta]\,\varphi \leftrightarrow [\alpha]\,\varphi \wedge [\beta]\,\varphi \qquad (\text{a7})$$

$$\langle 0]\,\varphi \to \bot \qquad (\text{a6}) \qquad\qquad \langle \alpha+\beta]\,\varphi \leftrightarrow \langle \alpha]\,\varphi \vee \langle \beta]\,\varphi \qquad (\text{a8})$$

Theorem 3. *(1)* $\varphi \in LSL$ *iff (2)* $\varphi \in Log(\mathsf{rFSL})$ *iff (3)* $(\top \approx \varphi) \in Eq(\mathsf{FSL})$.

Proof sketch. (1) implies (2) since the *LSL* axioms are valid in all relational js-frames. The fact that (2) implies (3) is established by showing that for each js-frame there is an equivalent relational js-frame. We cannot use the ultrafilter frame construction (Def. 8) for failure of canonicity[7]. However, a variant of the construction where 0 and + are defined exactly as in relational js-frames will do. That (3) implies (1) is established by contraposition, using a variant of the basic canonical *L*-frame of Example 4. Details are given in Appendix A.5. □

3.2 Meta-belief and right-unital magmas

Information about *meta-beliefs* ("i believes that j believes that p") is crucial to many multi-agent scenarios. The notion of meta-belief is often lifted to extensional groups of agents (sets). "Group I believes that group J believes that p" means that every agent in I believes that every agent in J believes that p. It is interesting to note that, if agents are seen as accessibility relations, the notion of meta-belief induces structure on sets of agents. In particular, every agent in I believes that every agent in J believes that p, iff every world accessible via $I \circ J = \{r \circ q \mid r \in I \ \& \ q \in J\}$ satisfies p. If the "environment" agent $E = \{\mathrm{id}_W\}$ is also included, we obtain a monoid structure. It is interesting to look at the notion of meta-belief, and the structure it induces, in the context of intensional groups.

Example 5. Adam ($\mathscr{A}$) is reviewing a paper for a journal, double-blind. Adam knows the researchers active in the particular area very well, and so he knows that either Bonnie ($\mathscr{B}$) or Carrie ($\mathscr{C}$) is the author or they are co-authoring the paper together. He knows that the authors of the paper, whoever they are, believe that the proof of a particular statement in the paper is correct (p), although Adam believes it is incorrect ($\neg p$). In reality, Bonnie and Carrie co-authored the paper and the proof is correct.

The scenario is represented by the relational model in Figure 1, with the actual world underlined. Adam's meta-beliefs concerning the authors of the paper are represented by the result of composing his relation with relations that "behave like" a relation corresponding to an author of the paper in any world accessible for Adam from the actual world. In particular, from the world $(\{\mathscr{C}\}, \neg p)$, representing the situation where only Carrie is the author and the proof is incorrect, only the $\mathscr{C}$-arrow is followed, and similarly for the world $(\{\mathscr{B}\}, \neg p)$ and the $\mathscr{B}$ arrow. This makes sense: beliefs of people who are not authors in the given world are disregarded. In the world $(\{\mathscr{B}, \mathscr{C}\}, \neg p)$, one could follow either $\mathscr{B}$ or $\mathscr{C}$, but the difference is not reflected by the accessibility arrows leading from that world.

[7] Axiom a8 $\langle \alpha+\beta]\,\varphi \leftrightarrow \langle \alpha]\,\varphi \vee \langle \beta]\,\varphi$ does not correspond to the condition $(f +^{\mathbf{G}} g)(w) = f(w) \cup g(w)$, but to the following one: $\forall w\ (\forall r \in (a+b)(w)\ \exists s \in a(w) \cup b(w)\ (s(w) \subseteq r(w))) \wedge \forall w\ (\forall s \in a(w) \cup b(w)\ \exists r \in (a+b)(w)\ (r(w) \subseteq s(w)))$. While the first conjunct is valid on an ultrafilter frame, the second one is not (unless we deal with an ultrafilter frame of a complete algebra). For the Σ_{SL}-neighborhood frames, a8 corresponds to the property: $\forall w(\nu_{a+b}(w)^{\uparrow} = \nu_a(w)^{\uparrow} \cup \nu_b(w)^{\uparrow})$.

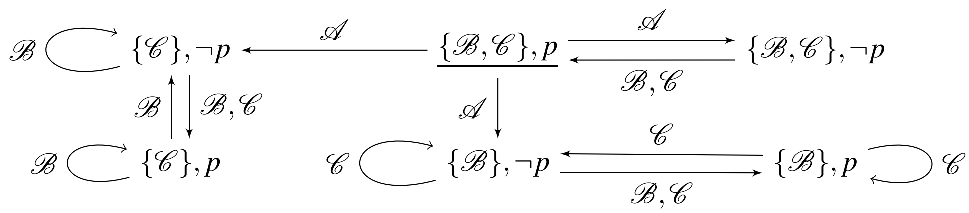


Figure 1: A relational model corresponding to Example 5.

Let W be a set and let $R \subseteq (2^W)^W$. Let $f \in (2^R)^W$ be an intensional group. A *variant* of f is a relation $r \in (2^W)^W$ such that, for each $w \in W$, if $f(w) \neq \emptyset$, then there is $q \in f(w)$ such that $r(w) = q(w)$, and if $f(w) = \emptyset$, then $r(w) = \emptyset$. We denote the set of all variants of f as $f^\dagger$.

Intuitively, a variant of an intensional group f is a relation that "behaves like" some relation in $f(w)$ whenever $f(w)$ is non-empty (not necessarily the same relation!) and which is "blind" in w where $f(w)$ is empty.

Definition 15 (**Intensional composition).** *Let R be a set of binary relations on a set W. We define the operation $\otimes : (2^R)^W \times (2^R)^W \to (2^R)^W$ point-wise by $(f \otimes g)(w) = f(w) \circ g^\dagger$.*

It can be shown that the structure on the set of intensional groups induced by intensional composition is rather weak. For instance, the natural candidate for the unit element, the function 1 that maps each w to $\{\mathrm{id}_W\}$ is the right unit, but not the left unit: $(f \otimes 1)(w) = f(w) \circ 1^\dagger = f(w) \circ \{\mathrm{id}_W\} = f(w)$, but in general $(1 \otimes f)(w) = 1(w) \circ f^\dagger = f^\dagger \neq f(w)$. (We note that in general $f^\dagger \neq f(w)$ since we can have $r \in f(w)$ and $r(v) \neq \emptyset$ for some $v \neq w$ such that $f(v) = \emptyset$.)

A *right-unital magma* (rum) is an algebra $(M, \cdot, 1)$ where $\cdot$ is a binary operation on M and $1 \in M$ such that $x \cdot 1 = x$ for all $x \in M$.

Definition 16 (**Rum-frame).** *Let $\Sigma_\mathsf{M} = \{\cdot, 1\}$ be the monoid signature. A* relational right-unital-magma frame *(a relational rum-frame) is a relational Σ_M-frame such that $1^\mathbf{G}(w) = \{\mathrm{id}_W\}$ and $(g \cdot^\mathbf{G} h)(w) = (g \otimes h)(w)$. A* rum-frame *is a Σ_M-frame where $\mathbf{G}$ is a right-unital-magma and*

$$[1]x = x \qquad (9) \qquad\qquad [a \cdot b]x = [a]\,[b]x \qquad (11)$$

$$\langle 1]x = x \qquad (10) \qquad\qquad \langle a \cdot b]x = \langle a]\,([b\rangle \bot \vee \langle b]x) \qquad (12)$$

The class of (relational) rum-frames will be denoted as FRUM (rFRUM, respectively).

Definition 17 (**Rum-logic).** *The* right-unital-magma logic *LRUM is the smallest Σ_M-logic that contains all formulas of the following forms:*

$$[1]\,\varphi \leftrightarrow \varphi \qquad (13) \qquad\qquad [\alpha \cdot \beta]\,\varphi \leftrightarrow [\alpha]\,[\beta]\,\varphi \qquad (15)$$

$$\langle 1]\,\varphi \leftrightarrow \varphi \qquad (14) \qquad\qquad \langle \alpha \cdot \beta]\,\varphi \leftrightarrow \langle \alpha]\,([\beta\rangle \bot \vee \langle \beta]\,\varphi) \qquad (16)$$

Remark 4. We note that a simpler variant of (15) for $\langle\,]$, namely, $\langle \alpha \cdot \beta]\,\varphi \leftrightarrow \langle \alpha]\,\langle \beta]\,\varphi$ is not valid. In particular, the left-to-right implication has the following counterexample. Let $[\![\alpha]\!](w) = \{r\}$ and let $r(w) = \{u, v\}$; moreover, let us assume that $[\![\beta]\!](u) = \{q\}$, $[\![\beta]\!](v) = \emptyset$, $q(u) = \{u\} = [\![p]\!]$. It is easily checked that $[\![\alpha \cdot \beta]\!](w) = \{\{(w,u)\}\}$, and so $w \models \langle \alpha \cdot \beta]\,p$. However, $w \not\models \langle \alpha]\,\langle \beta]\,p$, since this would require $[\![\beta]\!](v)$ to be non-empty. On the other hand, (16) is valid since worlds v accessible via α where $[\![\beta]\!](v) = \emptyset$ are taken care of by the extra disjunct $[\beta\rangle \bot$.

Theorem 4. *$\varphi \in LRUM$ iff $\varphi \in Log(\mathsf{rFRUM})$ iff $(\top \approx \varphi) \in Eq(\mathsf{FRUM})$.*

3.3 Closure semilattices and distributed knowledge

In the extensional setting, φ is distributed knowledge in a group iff it is satisfied in every world accessible using the intersection of the relations in the group. If all relations are reflexive, then φ is distributed knowledge iff there is a non-empty subset of the given group such that φ is satisfied in all worlds accessible using the intersection. On the relations-as-agents perspective, the intersection of each non-empty subset of a set of relations-agents gives rise to a new relation-agent. Hence, forming intersections of non-empty subsets of a group X transforms the group of relations-agents into a new group X'. Interestingly, distributed knowledge in X then corresponds to the "somebody knows" operator applied to X'. Hence, distributed knowledge induces structure on groups of agents even in the extensional setting. We will look at the structure induced by distributed knowledge in the intensional setting.

Definition 18 (CSL-frame)**.** *Let* $\Sigma_{\mathsf{CSL}} = \{+, 0, {}^{\cap}\}$ *where* $\{+, 0\}$ *is the join-semilattice signature and* ${}^{\cap}$ *is a unary operator. A* relational closure semilattice frame *(relational cs-frame) is a relational* Σ_{CSL}*-frame where all* $r \in R$ *are reflexive and* $0^{\mathbf{G}}(w) = \emptyset$*,* $(f +^{\mathbf{G}} g)(w) = f(w) \cup g(w)$*, and*

$$f^{\cap^{\mathbf{G}}}(w) = \{r \in R \mid r(w) = \bigcap_{r_i \in X} r_i(w) \textit{ for some } \emptyset \neq X \subseteq f(w)\}$$

[8] *A* closure semilattice frame *(cs-frame) is a* Σ_{CSL}*-frame where* $\mathbf{G}$ *is a join semilattice with partial order defined as usual (*$a \leq b$ *iff* $a + b = b$*),* ${}^{\cap}$ *is a closure operator on* $\mathbf{G}$*, the join-semilattice axioms (5–8) are satisfied as well as* $0^{\cap} = 0$*, and*

$$[a]x \leq x \tag{17}$$

$$\langle a^{\cap}]x \wedge \langle a^{\cap}]y \leq \langle a^{\cap}](x \wedge y) \tag{18}$$

$$[a^{\cap}]x = [a]x \tag{19}$$

$$\langle a^{\cap}]x \leq \langle a]\top \tag{20}$$

The class of (relational) cs-frames is denoted as FCS *(*rFCS*).*

Definition 19 (CS-logic). *The* closure semilattice logic *LCS is the smallest* Σ_{CSL}*-logic that extends LSL and contains all formulas of the following forms:*

$$[\alpha]\varphi \to \varphi \tag{21}$$

$$\langle \alpha^{\cap}]\varphi \wedge \langle \alpha^{\cap}]\psi \to \langle \alpha^{\cap}](\varphi \wedge \psi) \tag{22}$$

$$\langle \alpha]\varphi \to \langle \alpha^{\cap}]\varphi \tag{23}$$

$$[\alpha^{\cap}]\varphi \leftrightarrow [\alpha]\varphi \tag{24}$$

$$\langle \alpha^{\cap}]\varphi \to \langle \alpha]\top \tag{25}$$

$$\langle \alpha^{\cap\cap}]\varphi \to \langle \alpha^{\cap}]\varphi \tag{26}$$

and closed under the rule

$$\frac{\langle \alpha]\varphi \to \langle \beta]\varphi}{\langle \alpha^{\cap}]\varphi \to \langle \beta^{\cap}]\varphi}.$$

Theorem 5. $\varphi \in LCS$ *iff* $\varphi \in Log(\mathsf{rFCS})$ *iff* $(\top \approx \varphi) \in Eq(\mathsf{FCS})$.

4 Further work

With a reasonable notion of composition of intensional groups, we may use the standard fixpoint construction to introduce common knowledge into our framework. We intend to study the extension with common knowledge in the immediate future. An additional topic for future work is the exploration of variants of the notion of intensional composition. In particular, we are curious if there is a variant giving rise to a monoid structure on intensional groups.

[8] We assume that $r \in R$ in the frame is closed under this operation.

Acknowledgement. The authors are grateful to the reviewers for useful comments. A preliminary version of this paper was presented at the CELIA Project Meeting in Bayreuth, Germany on 23-24 February 2023. We acknowledge the valuable suggestions we received from the audience at the meeting.

References

[1] Thomas Ågotnes & Yì N. Wáng (2021): *Somebody knows*. In: *Proc. of (KR 2021)*, pp. 2–11, doi:10.24963/kr.2021/1.

[2] Marta Bílková, Zoé Christoff & Olivier Roy (2021): *Revisiting Epistemic Logic with Names*. In J. Halpern & A. Perea, editors: *TARK 2021*, pp. 39–54, doi:10.4204/EPTCS.335.4.

[3] Yifeng Ding, Jixin Liu & Yanjing Wang (2023): *Someone knows that local reasoning on hypergraphs is a weakly aggregative modal logic*. *Synthese* 201(46), doi:10.1007/s11229-022-04032-y.

[4] Melvin Fitting, Lars Thalmann & Andrei Voronkov (2001): *Term-Modal Logics*. *Studia Logica* 69, pp. 133–169, doi:10.1023/A:1013842612702.

[5] Adam J. Grove (1995): *Naming and identity in epistemic logic Part II: a first-order logic for naming*. *Artificial Intelligence* 74(2), pp. 311–350, doi:10.1016/0004-3702(95)98593-D.

[6] Adam J. Grove & Joseph Y. Halpern (1993): *Naming and identity in epistemic logics. Part I: The Propositional Case*. *Journal of Logic and Computation* 3(4), pp. 345–378, doi:10.1093/logcom/3.4.345.

[7] Helle Hvid Hansen (2003): *Monotonic modal logics*. Master's thesis, ILLC UVA. Available at https://eprints.illc.uva.nl/id/document/264.

[8] Helle Hvid Hansen, Clemens Kupke & Eric Pacuit (2007): *Bisimulation for neighbourhood structures*. In: *CALCO 2017*, Springer, pp. 279–293, doi:10.1007/978-3-540-73859-6_19.

[9] Merlin Humml & Lutz Schröder (2023): *Common Knowledge of Abstract Groups*. In: *AAAI '23*, doi:10.48550/arXiv.2211.16284.

[10] Barteld Kooi (2008): *Dynamic term-modal logic*. In J. Van Benthem, S. Ju & F. Veltman, editors: *A meeting of the minds. Proceedings of the workshop on Logic, Rationality and Interaction LORI 2007*, College publications, London, pp. 173–185. Available at https://research.rug.nl/files/2701112/Dynamitermmodallogic.pdf.

[11] Dexter Kozen (1980): *A representation theorem for models of *-free PDL*. In: *Proc. 7th Colloq. Automata, Languages, and Programming*, EATCS, pp. 351–362, doi:10.1007/3-540-10003-2_83. Available at https://dl.acm.org/doi/10.5555/646234.682547.

[12] Yoram Moses & Mark R Tuttle (1988): *Programming simultaneous actions using common knowledge*. *Algorithmica* 3(1), pp. 121–169, doi:10.1007/BF01762112.

[13] Pavel Naumov & Jia Tao (2018): *Everyone knows that someone knows: quantifiers over epistemic agents*. *The review of symbolic logic*, doi:10.1017/S1755020318000497.

[14] Eric Pacuit (2017): *Neighborhood semantics for modal logic*. Springer, doi:10.1007/978-3-319-67149-9.

[15] Anantha Padmanabha & Rangaraj Ramanujam (2019): *Propositional modal logic with implicit modal quantification*. In: *ICLA 2019*, Springer, pp. 6–17, doi:10.1007/978-3-662-58771-3_2.

[16] Anantha Padmanabha & Rangaraj Ramanujam (2019): *Two variable fragment of Term Modal Logic*. In: *MFCS 2019*, *LIPIcs* 138, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 1–14, doi:10.4230/LIPIcs.MFCS.2019.30.

[17] Anantha Padmanabha, Rangaraj Ramanujam & Yanjing Wang (2018): *Bundled Fragments of First-Order Modal Logic: (Un)Decidability*. In: *FSTTCS 2018*, *LIPIcs* 122, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 1–20, doi:10.4230/LIPIcs.FSTTCS.2018.43.

[18] Vaughan Pratt (1991): *Dynamic algebras: Examples, constructions, applications*. Studia Logica 50(3), pp. 571–605, doi:10.1007/BF00370685.

[19] Gennady Shtakser (2018): *Propositional epistemic logics with quantification over agents of knowledge*. Studia Logica 106(2), pp. 311–344, doi:10.1007/s11225-017-9741-0.

[20] Yanjing Wang & Jeremy Seligman (2018): *When names are not commonly known: epistemic logic with assignments*. In G. Bezhanishvili, G. D'Agostino, G. Metcalfe & T. Studer, editors: Proceedings of AiML *2018*, pp. 611–628. Available at http://www.aiml.net/volumes/volume12/Wang-Seligman.pdf.

A Appendix

A.1 Proof of Lemma 1

Proof. It is a standard observation that f embeds any Boolean algebra $\mathbf{F}$ into the power set algebra over $\mathrm{Uf}(\mathbf{F})$. This takes care of (m1) and the injectivity condition. (m2) holds by definition.

(m3) is established as follows (we write $\texttt{[]}$ instead of $\texttt{[]}^+$). The inclusion $f(\texttt{[}a\texttt{]}x) \subseteq \texttt{[}G(a)\texttt{]}\hat{x}$ means that if $\texttt{[}a\texttt{]}x \in u$ and $r \in G(a,u)$ for some u, then $r(u) \subseteq \hat{x}$. This holds by the definition of $G(a)$. The converse inclusion $\texttt{[}G(a)\texttt{]}\hat{x} \subseteq f(\texttt{[}a\texttt{]}x)$ is established by contraposition. Assume that $u \notin f(\texttt{[}a\texttt{]}x)$, that is, $\texttt{[}a\texttt{]}x \notin u$ for some a and x. Using (1–2), we can show that $v_0 = \{y \mid \texttt{[}a\texttt{]}y \in u\}$ is a filter on $\mathbf{F}$ such that $x \notin v_0$. Hence, v_0 extends to an ultrafilter v such that $x \notin v$. Take the relation $r_{a,\top}$, where $\top := x \vee \neg x$ for some $x \in \mathbf{F}$. By (2–3), $\texttt{[}a\texttt{]}x \notin u$ implies $\langle a\texttt{]}\top \in u$, and so $r_{a,\top} \in G(a,u)$. Moreover, $r_{a,\top}(u,v)$ by the construction of v. This means that $u \notin \texttt{[}G(a)\texttt{]}\hat{x}$.

(m4) is established as follows. The inclusion $f(\langle a\texttt{]}x) \subseteq \langle G(a)\texttt{]}\hat{x}$ means that if $\langle a\texttt{]}x \in u$ for some u, then there is $r \in G(a,u)$ such that $r(u) \subseteq \hat{x}$. Fix such a, x and u, and consider the relation $r_{a,x}$. It is clear that $r_{a,x} \in G(a,u)$ and $r_{a,x}(u) \subseteq \hat{x}$. The converse inclusion $\langle G(a)\texttt{]}\hat{x} \subseteq f(\langle a\texttt{]}x)$ is established as follows. Let us assume that $\langle G(a)\texttt{]}\hat{x}$, i.e. there is $r \in G(a,u)$ such that $r(u) \subseteq \hat{x}$. This means that $r = r_{a,z}$ for some z such that $\langle a\texttt{]}z \in u$, and $r(u) = \bigcap\{\hat{y} \mid \texttt{[}a\texttt{]}y \in u\} \cap \hat{z}$. This entails that

$$\{z\} \cup \{y \mid \texttt{[}a\texttt{]}y \in u\} \subseteq w \implies x \in w$$

for all $w \in \mathrm{Uf}(\mathbf{F})$. Hence, x is in the filter generated by $\{z\} \cup \{y \mid \texttt{[}a\texttt{]}y \in u\}$. By the properties of filters generated by (non-empty) subsets of a lattice, there is a finite $\{y_1, \ldots, y_n\} \subseteq \{y \mid \texttt{[}a\texttt{]}y \in u\}$ such that $z \wedge y_1 \wedge \ldots \wedge y_n \leq x$. This means that

$$\langle a\texttt{]}z \wedge \texttt{[}a\texttt{]}y_1 \wedge \ldots \wedge \texttt{[}a\texttt{]}y_n \leq \langle a\texttt{]}x,$$

using (2) and (4). Consequently, $\langle a\texttt{]}x \in u$ as we wanted to show. □

A.2 Proof of Lemma 2 (Preservation of validity for neighborhood frame morphisms)

Proof. To show for each formula φ and each $w \in W$, $(\mathfrak{F}_1, w \Vdash \varphi \quad \Rightarrow \quad \mathfrak{F}_2, f(w) \Vdash \varphi)$, it is enough to show that, once we fix valuations on the respective frames so that (i) for each $\mathtt{a} \in Gr$, $g(\llbracket \mathtt{a} \rrbracket_1) = \llbracket \mathtt{a} \rrbracket_2$, and (ii) for each $\mathtt{p} \in Pr$, $w \in \llbracket \mathtt{p} \rrbracket_1$ iff $f(w) \in \llbracket \mathtt{p} \rrbracket_2$, we obtain for each φ

$$w \in \llbracket \varphi \rrbracket_1 \;\Leftrightarrow\; f(w) \in \llbracket \varphi \rrbracket_2.$$

This is easily proven by a routine induction on the complexity of a given formula. □

A.3 Bisimulations of neighborhood frames

To see what a natural notion of bisimulation is for neighborhood Σ-models, compared to that of models for epistemic logic with names described in [2, Definition 6], we need to incorporate the algebraic component. For a binary relation B, let $X\,\overline{B}\,Y$ iff $\forall x \in X \exists y \in Y\; xBy$ and $\forall y \in Y \exists x \in X\; xBy$.

Definition 20 (Bisimulations). *Let $(W_1, \nu^1, \mathbf{G_1}, \llbracket\rrbracket_1)$ and $(W_2, \nu^2, \mathbf{G_2}, \llbracket\rrbracket_2)$ be neighborhood Σ-models. A pair $(\cong, B)$, with $\cong \subseteq \mathbf{G_1} \times \mathbf{G_2}$ being a congruence relation, and $B \subseteq W_1 \times W_2$, is a bisimulation of neighborhood Σ-models, if*

$$\forall \mathtt{a} \in Gr\; \llbracket \mathtt{a} \rrbracket_1 \cong \llbracket \mathtt{a} \rrbracket_2 \;\text{ and }\; \forall \mathtt{p} \in Pr\; (w_1 B w_2 \Rightarrow (w_1 \in \llbracket \mathtt{p} \rrbracket_1 \Leftrightarrow w_2 \in \llbracket \mathtt{p} \rrbracket_2))$$

$$(w_1 B w_2 \wedge a_1 \cong a_2) \Rightarrow (\forall X \in \nu^1_{a_1}(w_1) \exists Y \in \nu^2_{a_2}(w_2)\ X\,\overline{B}\,Y) \wedge (\forall Y \in \nu^2_{a_2}(w_2) \exists X \in \nu^1_{a_1}(w_1)\ X\,\overline{B}\,Y)$$

As expected, bisimilarity implies modal equivalence for the language of the basic Σ-logic, and the converse holds for image-finite models (where every core neighborhood set is a finite set of finite sets). Graphs of neighborhood Σ-frame morphisms are prominent examples of bisimulations, and functional bisimulations correspond to graphs of neighborhood Σ-frame morphisms.

A.4 Proof of Theorem 2 (Categorial equivalence)

Proof. For a relational Σ-frame $\mathfrak{F} = (W, R, \mathbf{G})$, we define the corresponding neighborhood Σ-frame $\mathfrak{F}^n = (W, \mathbf{G}, \{\nu^n_a\}_{a \in G})$ as follows:

$$\nu^n_a(w) = \{r(w) \mid r \in a(w)\}.$$

Conversely, for a neighborhood Σ-frame $\mathfrak{F} = (W, \mathbf{G}, \{\nu_a\}_{a \in G})$ we define the corresponding Σ-frame $\mathfrak{F}^r = (W, R^r, \mathbf{G})$ as follows:

$$a^r(w) = \{r \mid r(w) \in \nu_a(w)\},\ R^r = \bigcup_{a \in G, w \in W} a^r(w).$$

It is easy to see that $\nu^n_a(w) = \{r(w) \mid r \in a^r(w)\} = \nu_a(w)$ and $a^r(w) = \{r \mid r(w) \in \nu^n_a(w)\} = a(w)$. However, going there-and-back on a relational frame, we do not recover the same R, as we can in principle recover only those relations r in R that are in some $a(w)$, but we also include relations who agree with r on w. Still the resulting frame ends up to be isomorphic to the original one in terms of frame morphisms, which is what matters.

For the morphism part, we use the fact that the corresponding frames are defined over the same set W and Σ-algebra $\mathbf{G}$, and therefore we use the same underlying map in both directions. First, we observe that morphisms of relational Σ-frames (which can be read of the Definition 9 of morphisms of Σ-frames) can equivalently be understood as pairs of maps $(g : \mathbf{G} \to \mathbf{G}', f : W \to W')$, where g is a homomorphism of Σ-algebras, satisfying:

(there) $\forall r \in a(w)\ \exists r' \in g(a)(f(w))\ (f[r(w)] = r'(f(w)))$,

(back) $\forall r' \in g(a)(f(w))\ \exists r \in a(w)\ (f[r(w)] = r'(f(w)))$.

It is not hard to see now that the two notions of morphisms are equivalent, when applied to the translated frames respectively. □

A.5 Proof of Theorem 3 (Completeness of semilattice logic)

Proof. If $\varphi \in LSL$, then $\varphi \in Log(\mathsf{rFSL})$. It is sufficient to show that (a5–a8) are valid on relational js-frames. This is easily shown using the definition of js-frames. For instance, $w \models [0]\,\varphi$ iff $\forall r \in \llbracket 0 \rrbracket(w) : r(w) \subseteq \llbracket \varphi \rrbracket$. However, since $\llbracket 0 \rrbracket(w) = \emptyset$, this is trivially satisfied for all w. As another example, note that $w \models \langle \alpha + \beta]\,\varphi$ iff there is $r \in \llbracket \alpha + \beta \rrbracket(w)$ such that $r(w) \subseteq \llbracket \varphi \rrbracket$. However, $\llbracket \alpha + \beta \rrbracket(w) = \llbracket \alpha \rrbracket(w) \cup \llbracket \beta \rrbracket(w)$ and so the previous statement is equivalent to $\exists r \in \llbracket \alpha \rrbracket(w) : r(w) \subseteq \llbracket \varphi \rrbracket$ or $\exists r \in \llbracket \beta \rrbracket(w) : r(w) \subseteq \llbracket \varphi \rrbracket$ which is equivalent to $w \models \langle \alpha]\,\varphi \vee \langle \beta]\,\varphi$.

$\varphi \in Log(\mathsf{rFSL})$ implies $(\top \approx \varphi) \in Eq(\mathsf{FSL})$. We reason by contraposition. Fix a js-frame $\mathfrak{A} = (\mathbf{F}, \mathbf{G}, [\,], \langle\,])$ and an evaluation function e such that $e(\varphi) \neq e(\top)$ for some $\varphi \in Fm_\Sigma$. We define the relational Σ_{SL}-frame $\mathfrak{F} = (\mathrm{Uf}(\mathbf{F}), R, \mathbf{H})$ where $R = (2^{\mathrm{Uf}(\mathbf{F})})^{\mathrm{Uf}(\mathbf{F})}$ and $\mathbf{H} = (H, 0^{\mathbf{H}}, +^{\mathbf{H}})$ is specified as follows: $H = \{H(\alpha) \mid \alpha \in Tm_{\Sigma_{\mathsf{SL}}}\}$ where $H(\alpha) \in (2^R)^{\mathrm{Uf}(\mathbf{F})}$ such that

- $H(\mathtt{a})(u) = \{r_{e(\mathtt{a}), e(\varphi)} \mid e(\langle \mathtt{a}]\,\varphi) \in u\}$;[9]

[9] Recall the definition of $r_{a,x}$ in Def. 8: $r_{a,x}(u) = \bigcap\{\hat{y} \mid [a]y \in u\} \cap \hat{x}$.

- $H(0)(u) = 0^{\mathbf{H}}(u) = \emptyset$; and
- $H(\alpha+\beta)(u) = (H(\alpha) +^{\mathbf{H}} H(\beta))(u) = H(\alpha)(u) \cup H(\beta)(u)$.

It is clear that $\mathfrak{F}$ is a relational js-frame. We define $V : Tm \cup Fm \to 2^{\mathrm{Uf}(\mathbf{F})} \cup H$ by $V(\chi) = \widehat{e(\chi)}$ and $V(\alpha) = H(\alpha)$. We show that V is an interpretation function on $\mathfrak{F}$. V is a Boolean homomorphism by the properties of ultrafilters and it is a Σ-homomorphism from Tm to $\mathbf{H}$ by the definition of $H(\alpha)$. (For instance, $V(\alpha+\beta)(u) = H(\alpha+\beta)(u) = H(\alpha)(u) \cup H(\beta)(u) = (V(\alpha) +^{\mathbf{H}} V(\beta))(u)$ for all u; hence, $V(\alpha+\beta) = (V(\alpha) +^{\mathbf{H}} V(\beta)$.) It remains to show the leftmost equalities in the following (recall that we write $H(\gamma, w)$ instead of $H(\gamma)(w)$):

(i) $V([\gamma]\chi) = [V(\gamma)]V(\chi) = \{w \mid \forall r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$; and

(ii) $V(\langle\gamma]\chi) = \langle V(\gamma)]V(\chi) = \{w \mid \exists r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$.

We show both by induction on the complexity of γ. (i) The base case $\gamma \in Gr$ is established similarly as the corresponding case in the proof of Lemma 1, since $H(\gamma)$ is in this case defined as $G(\gamma)$ in the ultrafilter frame. The case $\gamma = 0$ is established as follows: $w \in V([0]\chi)$ iff $e([0]\chi) \in w$ iff $[e(0)]e(\chi) \in w$ iff $[0]e(\chi) \in w$ iff (using axiom 5) $\top \in w$ iff $\forall r \in \emptyset : r(w) \subseteq \widehat{e(\chi)}$ (both are true for all w) iff $\forall r \in H(0, w) : r(w) \subseteq V(\chi)$ iff $w \in [V(0)]V(\chi)$. The case $\gamma = \alpha+\beta$ is established as follows: $w \in V([\alpha+\beta]\chi)$ iff $w \in e([\alpha+\beta]\chi)$ iff $w \in [e(\alpha)+e(\beta)]e(\chi)$ iff (using axiom 7) $w \in [e(\alpha)]e(\chi)$ and $w \in [e(\beta)]e(\chi)$ iff $w \in V([\alpha]\chi)$ and $w \in V([\beta]\chi)$ iff (by the induction hypothesis) $\forall r \in H(\alpha, w) : r(w) \subseteq \widehat{e(\chi)}$ and $\forall r \in H(\beta, w) : r(w) \subseteq \widehat{e(\chi)}$ iff $\forall r \in H(\alpha, w) \cup H(\beta, w) : r(w) \subseteq V(\chi)$ iff $\forall r \in H(\alpha+\beta, w) : r(w) \subseteq V(\chi)$ iff $w \in [V(\alpha+\beta)]V(\chi)$. Part (ii) is established similarly, using axiom (6) in the case $\gamma = 0$ and axiom (8) in the case $\gamma = \alpha+\beta$.

Now since $e(\varphi) \neq e(\top)$, there is $u \in \mathrm{Uf}(\mathbf{F})$ such that $e(\varphi) \in u$ and $e(\top) \notin u$ by the Prime Filter Theorem. Hence, $V(\varphi) \neq V(\top)$ and so φ is not valid in the relational js-model $(\mathfrak{F}, V)$.

$(\top \approx \varphi) \in Eq(\mathsf{FSL})$ implies $\varphi \in LSL$. We define the *canonical LSL-frame* as follows.[10] Let $\equiv$ be a binary relation on formulas defined as $\equiv_L$ (for $L = LSL$) in Example 4, and let $\equiv^{Tm}$ be a binary relation on Tm such that $\alpha \equiv^{Tm} \beta$ iff $[\alpha]\varphi \leftrightarrow [\beta]\varphi \in LSL$ and $\langle\alpha]\varphi \leftrightarrow \langle\beta]\varphi \in LSL$ for all $\varphi \in Fm$. Let $[\varphi]$ be the equivalence class of φ under $\equiv$ and let $[\alpha]$ be the equivalence class of α under $\equiv^{Tm}$. It can be shown that $\equiv$ is a congruence on Fm (the usual argument) and $\equiv^{Tm}$ is a congruence on Tm. The latter is established using the "reduction axioms" for the semilattice operators: if $[\alpha]\varphi \leftrightarrow [\beta]\varphi \in LSL$ for all φ, then $[\alpha+\gamma]\varphi \leftrightarrow [\beta+\gamma] \in LSL$ since $[\alpha+\gamma]\varphi \leftrightarrow [\alpha]\varphi \wedge [\gamma]\varphi \in LSL$ using (a7), and so $[\alpha+\gamma]\varphi \leftrightarrow [\beta]\varphi \wedge [\gamma]\varphi \in LSL$ by the assumption which means that $[\alpha+\gamma]\varphi \leftrightarrow [\beta+\gamma]\varphi \in LSL$ using (a7) again. Hence, we obtain the Boolean algebra $\mathbf{F}$ of equivalence classes $[\varphi]$, where $o^{\mathbf{F}}([\varphi_1], \ldots, [\varphi_n]) = [o(\varphi_1, \ldots, \varphi_n)]$ for all Boolean operators o, and the join-semilattice $\mathbf{G}$ of equivalence classes $[\alpha]$, where $o^{\mathbf{G}}([\alpha_1], \ldots, [\alpha_n]) = [o(\alpha_1, \ldots, \alpha_n)]$ for all $o \in \{0, +\}$. (The fact that $\mathbf{G}$ is a join-semilattice is easily shown using the reduction axioms: for instance, $[\alpha+\alpha] = [\alpha]$ since $[\alpha+\alpha]\varphi \equiv [\alpha]\varphi$ and $\langle\alpha+\alpha]\varphi \equiv \langle\alpha]\varphi$ which means that $\beta \in [\alpha+\alpha]$ iff $\beta \in [\alpha]\varphi$. Moreover, let $[]$ and $\langle]$ be functions of the type $\mathbf{F} \times \mathbf{G} \to \mathbf{F}$ such that

$$[[\alpha]][\varphi] = [[\alpha]\varphi] \quad \text{and} \quad \langle[\alpha]][\varphi] = [\langle\alpha]\varphi]$$

(note that these functions are well defined since $\equiv$ and $\equiv^{Tm}$ are both congruences). The *canonical LSL-frame* is $\mathfrak{C}^{LSL} = (\mathbf{F}, \mathbf{G}, [], \langle])$. It is clear that $\varphi \in LSL$ iff $\varphi \approx \top$ is valid in $\mathfrak{C}^{LSL}$ (Prime Filter Theorem). Hence, if $\varphi \notin LSL$, then there is a js-frame that invalidates $\top \approx \varphi$, establishing our claim by contraposition. □

[10] The definition of the canonical LSL-frame resembles the definition of the basic canonical L-frame from Example 4. However, we cannot use $\mathfrak{B}^{LSL}$ here since the group algebra $\mathbf{G}^{LSL}$ in $\mathfrak{B}^{LSL}$ is not a join-semilattice – it is the term algebra. Hence, we have to define a suitable LSL-congruence on the term algebra and prove that it gives rise to a join-semilattice.

A.6 Proof of Theorem 4 (Completeness of right-unital magma logic)

Proof. $\varphi \in LRUM$ implies $\varphi \in Log(\mathsf{rFRUM})$. It is sufficient to check that the extra axioms (13–16) of *LRUM* are valid in all relational rum-frames. The validity of (13–14) is clear. To show that (15) is valid, we reason as follows: $w \not\models [\alpha][\beta]\varphi$ iff there are $r \in \llbracket\alpha\rrbracket(w)$, $u \in r(w)$ and $q \in \llbracket\beta\rrbracket(u)$ such that $q(u) \not\subseteq \llbracket\varphi\rrbracket$ iff[11] there are $r \in \llbracket\alpha\rrbracket(w)$, $u \in W$ and $q' \in \llbracket\beta\rrbracket^{\dagger}$ such that $r(w,u)$ and $q'(u) \not\subseteq \llbracket\varphi\rrbracket$ iff there is $s \in \llbracket\alpha \cdot \beta\rrbracket(w)$ such that $s(w) \not\subseteq \llbracket\varphi\rrbracket$ iff $w \not\models [\alpha \cdot \beta]\varphi$. To show that (16) is valid, we reason as follows: $w \models \langle\alpha \cdot \beta]\varphi$ iff there is $r \in \llbracket\alpha\rrbracket(w) \circ \llbracket\beta\rrbracket^{\dagger}$ such that $r(w) \subseteq \llbracket\varphi\rrbracket$ iff there are $q \in \llbracket\alpha\rrbracket(w)$ and $s \in \llbracket\beta\rrbracket^{\dagger}$ such that $(q \circ s)(w) \subseteq \llbracket\varphi\rrbracket$ iff there are $q \in \llbracket\alpha\rrbracket(w)$ and $s \in \llbracket\beta\rrbracket^{\dagger}$ such that $s(q(w)) \subseteq \llbracket\varphi\rrbracket$ iff there is $q \in \llbracket\alpha\rrbracket(w)$ such that for all $u \in q(w)$, if $\llbracket\beta\rrbracket(u) \neq \emptyset$, then there is $t \in \llbracket\beta\rrbracket(u)$ such that $t(u) \subseteq \llbracket\varphi\rrbracket$ iff $u \models \langle\alpha]([\beta\rangle\bot \vee \langle\beta]\varphi)$.

$\varphi \in Log(\mathsf{rFRUM})$ implies $(\top \approx \varphi) \in Eq(\mathsf{FRUM})$. We reason by contraposition. Fix a rum-frame $\mathfrak{A} = (\mathbf{F}, \mathbf{G}, [], \langle])$ and an evaluation function e such that $e(\varphi) \neq e(\top)$ for some $\varphi \in Fm_{\Sigma_{\mathsf{M}}}$. We define the relational Σ_{M}-frame $\mathfrak{F} = (\mathrm{Uf}(\mathbf{F}), R, \mathbf{H})$ where $R = (2^{\mathrm{Uf}(\mathbf{F})})^{\mathrm{Uf}(\mathbf{F})}$ and $\mathbf{H} = (H, 1^{\mathbf{H}}, \cdot^{\mathbf{H}})$ is specified as follows: $H = \{H(\alpha) \mid \alpha \in Tm_{\Sigma_{\mathsf{M}}}\}$ where $H(\alpha) \in (2^{R})^{\mathrm{Uf}(\mathbf{F})}$ such that

- $H(\mathtt{a})(u) = \{r_{e(\mathtt{a}),e(\varphi)} \mid e(\langle\mathtt{a}]\varphi) \in u\}$;
- $H(1)(u) = 1^{\mathbf{H}}(u) = \{\mathrm{id}_{\mathrm{Uf}(\mathbf{F})}\}$; and
- $H(\alpha \cdot \beta)(u) = (H(\alpha) \cdot^{\mathbf{H}} H(\beta))(u) = (H(\alpha) \otimes H(\beta)(u) = H(\alpha)(u) \circ H(\beta)^{\dagger}$.

It is clear that $\mathfrak{F}$ is a relational rum-frame. We define $V : Tm \cup Fm \to 2^{\mathrm{Uf}(\mathbf{F})} \cup H$ by $V(\chi) = \widehat{e(\chi)}$ and $V(\alpha) = H(\alpha)$. We show that V is an interpretation function on $\mathfrak{F}$. V is a Boolean homomorphism by the properties of ultrafilters and it is a Σ_{M}-homomorphism from Tm to $\mathbf{H}$ by the definition of $H(\alpha)$. (For instance, $V(\alpha \cdot \beta)(u) = H(\alpha \cdot \beta)(u) = H(\alpha)(u) \otimes H(\beta)^{\dagger} = (V(\alpha) \cdot^{\mathbf{H}} V(\beta))(u)$ for all u; hence, $V(\alpha \cdot \beta) = (V(\alpha) \cdot^{\mathbf{H}} V(\beta)$.) It remains to establish the leftmost equalities in the following:

(i) $V([\gamma]\chi) = [V(\gamma)]V(\chi) = \{w \mid \forall r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$; and

(ii) $V(\langle\gamma]\chi) = \langle V(\gamma)]V(\chi) = \{w \mid \exists r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$.

We show both by induction on the complexity of γ. (i) The base case $\gamma \in Gr$ is established similarly as the corresponding case in the proof of Lemma 1, since $H(\gamma)$ is in this case defined as $G(\gamma)$ in the ultrafilter frame. The case $\gamma = 1$ is established as follows: $w \in V([1]\chi)$ iff $e([1]\chi) \in w$ iff $[e(1)]e(\chi) \in w$ iff $[1]e(\chi) \in w$ iff (using axiom 9) $e(\chi) \in w$ iff $\forall r \in H(1, w) : r(w) \subseteq \widehat{e(\chi)}$ iff $\forall r \in H(1, w) : r(w) \subseteq V(\chi)$ iff $w \in [V(1)]V(\chi)$. The case $\gamma = \alpha \cdot \beta$ is established as follows: $w \in V([\alpha \cdot \beta]\chi)$ iff $[e(\alpha) \cdot e(\beta)]e(\chi) \in w$ iff (using axiom 11) $[e(\alpha)][e(\beta)]e(\chi) \in w$ iff $w \in V([\alpha][\beta]\chi)$ iff (by induction hypothesis applied to α) $\forall r \in H(\alpha, w) \forall u (r(w,u) \to u \in V([\beta]\chi))$ iff (by induction hypothesis applied to β) $\forall r \in H(\alpha, w) \forall u (r(w,u) \to (\forall q \in H(\beta, u) : q(u) \subseteq V(\chi)))$ iff $\forall r, q (r \in H(\alpha, w) \ \& \ q \in H(\beta)^{\dagger} \to q(r(w)) \subseteq V(\chi))$ iff $\forall r, q (r \in H(\alpha, w) \ \& \ q \in H(\beta)^{\dagger} \to (r \circ q)(w) \subseteq V(\chi))$ iff $\forall s \in H(\alpha, w) \circ H(\beta)^{\dagger} : s(w) \subseteq V(\chi)$ iff $\forall s \in H(\alpha \cdot \beta, w) : s(w) \subseteq V(\chi)$ iff $w \in [H(\alpha \cdot \beta)]V(\chi)$ iff $w \in [V(\alpha \cdot \beta)]V(\chi)$. Part (ii) is established similarly, using axiom (10) in the case $\gamma = 1$ and axiom (12) in the case $\gamma = \alpha \cdot \beta$.

$(\top \approx \varphi) \in Eq(\mathsf{FRUM})$ implies $\varphi \in LRUM$. We define the *canonical LRUM-frame* $\mathfrak{C}^{LM}$ similarly as we defined $\mathfrak{C}^{LSL}$ in the third part of the proof of Theorem 3, but we use *LRUM* instead of *LSL*, of course. We have to show in our specific setting that (i) $\equiv^{Tm}$ is a congruence and that (ii) $\mathbf{G}$ is a right-unital magma. (i) is established using axioms (15–16). To establish (ii), we need to show that $[1]$ is the right

[11] Left to right: define q' so that $q'(u) = q(u)$ and $q'(v)$ for $v \neq$ u is fixed in an arbitrary way so that $q' \in \llbracket\beta\rrbracket^{\dagger}$ (this can always be done). Right to left: If $q' \in \llbracket\beta\rrbracket^{\dagger}$ and $q'(u) \neq \emptyset$, then by definition there has to be $q \in \llbracket\beta\rrbracket(u)$ such that $q(u) = q'(u)$.

unit with respect to $\otimes^{\mathbf{G}}$: to see this, it is sufficient to observe that $\mathtt{[}\alpha\cdot 1\mathtt{]}\varphi \equiv \mathtt{[}\alpha\mathtt{]}\mathtt{[}1\mathtt{]}\varphi \equiv \mathtt{[}\alpha\mathtt{]}\varphi$ and $\langle\alpha\cdot 1\mathtt{]}\varphi \equiv \langle\alpha\mathtt{]}(\mathtt{[}1\rangle\bot \vee \langle 1\mathtt{]}\varphi) \equiv \langle\alpha\mathtt{]}\varphi$. Hence, $\alpha \equiv^{Tm} \alpha\cdot 1$.

As before, the Prime Filter Theorem entails that $\varphi \in LRUM$ iff $\varphi \approx \top$ is valid in $\mathfrak{C}^{LRUM}$. Hence, if $\varphi \notin LRUM$, then there is a rum-frame that invalidates $\top \approx \varphi$, establishing our claim by contraposition. □

A.7 Proof of Theorem 5 (Completeness of closure semilattice logic)

Proof. $\varphi \in LCS$ implies $\varphi \in Log(\mathsf{rFCS})$. Validity of axioms (21–26) in relational cs-models is easily checked. The rule $\dfrac{\langle\alpha\mathtt{]}\varphi \to \langle\beta\mathtt{]}\varphi}{\langle\alpha^{\cap}\mathtt{]}\varphi \to \langle\beta^{\cap}\mathtt{]}\varphi}$ preserves validity: Assume there is a counterexample to the conclusion of the rule (we may assume $\phi = \mathtt{p}$ is atomic). Assume $w \models \langle\alpha^{\cap}\mathtt{]p}$ and $w \not\models \langle\beta^{\cap}\mathtt{]p}$. Then $w \not\models \langle\beta\mathtt{]p}$. We want to show that there is $\models'$ such that $w \models' \langle\alpha\mathtt{]p}$ and $w \not\models' \langle\beta\mathtt{]p}$. We know that α cannot be 0 since $0^{\cap} = 0$. If $\alpha = \gamma^{\cap}$ then we are done ($\models'$ is $\models$). Then there is a non-empty $X \subseteq [\![\alpha]\!](\mathrm{w})$ such that $\bigcap_{r\in X} r(x) \subseteq [\![p]\!]$ (W.l.o.g. we consider X to be a minimal nonempty such set). Now define for each $\mathtt{a} \in Gr$ the relation $r_{\mathtt{a}} = \bigcap\{r \in X \mid r \in [\![\mathtt{a}]\!](w)\}$ and define $[\![\mathtt{a}]\!]'(w) = [\![\mathtt{a}]\!](w) \cup \{r_{\mathtt{a}}\}$ in case $r_{\mathtt{a}} \neq \emptyset$, and $[\![\mathtt{a}]\!]'(w) = [\![\mathtt{a}]\!](w)\}$ otherwise. The interpretations $[\![\gamma]\!]'$ of complex γ are computed as usual. We can then prove by induction on α, β that $w \models' \langle\alpha\mathtt{]p}$ while $w \not\models' \langle\beta\mathtt{]p}$.

$\varphi \in Log(\mathsf{rFCS})$ implies $(\top \approx \varphi) \in Eq(\mathsf{FCS})$. We reason by contraposition. Fix a cs-frame $\mathfrak{A} = (\mathbf{F}, \mathbf{G}, \mathtt{[]}, \langle\mathtt{]})$ and an evaluation function e such that $e(\varphi) \neq e(\top)$ for some $\varphi \in Fm_{\Sigma_{\mathsf{M}}}$. Let Γ be the smallest set that contains φ and $\top$, is closed under taking subformulas, and

- $\mathtt{[}\alpha\mathtt{]}\chi \in \Gamma$ iff $\langle\alpha\mathtt{]}\chi \in \Gamma$
- $\mathtt{[}\alpha+\beta\mathtt{]}\chi \in \Gamma$ only if $\mathtt{[}\alpha\mathtt{]}\chi \in \Gamma$ and $\mathtt{[}\beta\mathtt{]}\chi \in \Gamma$
- $\langle\alpha+\beta\mathtt{]}\chi \in \Gamma$ only if $\langle\alpha\mathtt{]}(\neg\langle\beta\mathtt{]}\top \vee \langle\beta\mathtt{]}\chi) \in \Gamma$
- $\mathtt{[}\alpha^{\cap}\mathtt{]}\chi \in \Gamma$ only if $\mathtt{[}\alpha\mathtt{]}\chi \in \Gamma$
- $\langle\alpha^{\cap}\mathtt{]}\chi \in \Gamma$ only if $\langle\alpha\mathtt{]}\top \in \Gamma$

It is easily seen that Γ is always finite. For $x \in \mathbf{F}$ and $a \in \mathbf{G}$ such that $x = e(\chi)$ and $a = e(\alpha)$ for some $\mathtt{[}\alpha\mathtt{]}\chi \in \Gamma$, we define

$$r^{\Gamma}_{a,x} : w \mapsto \bigcap\{\widehat{e(\psi)} \mid \mathtt{[}\alpha\mathtt{]}\varphi \in \Gamma \ \&\ e(\mathtt{[}\alpha\mathtt{]}\chi) \in w\} \cap \widehat{x}.$$

For other pairs of x, a we define $r^{\Gamma}_{a,x} : w \mapsto \emptyset$.

We define the relational frame $\mathfrak{F}^{\Gamma} = (\mathrm{Uf}(\mathbf{F}), R, \mathbf{H})$ as before, but this time $\mathbf{H} = (H, 0^{\mathbf{H}}, +^{\mathbf{H}}, {}^{\cap^{\mathbf{H}}})$ is specified using the relations $r^{\Gamma}_{a,x}$ as follows:

- $H(\mathtt{a})(u) = \{r^{\Gamma}_{e(\mathtt{a}),e(\varphi)} \mid e(\langle\mathtt{a}\mathtt{]}\varphi) \in u\}$;
- $H(0)(u) = 0^{\mathbf{H}}(u) = \emptyset$;
- $H(\alpha+\beta)(u) = (H(\alpha) +^{\mathbf{H}} H(\beta))(u) = H(\alpha)(u) \cup H(\beta)(u)$; and
- $H(\alpha^{\cap})(u) = H(\alpha)^{\cap^{\mathbf{H}}}(u) = \{r \in R \mid r(u) = \bigcap_{q\in X} q(w)$ for some non-empty $X \subseteq H(u)\}$.

It is clear that $\mathfrak{F}$ is a relational cs-frame. We define $V : Tm \cup Fm \to 2^{\mathrm{Uf}(\mathbf{F})} \cup H$ by $V(\chi) = \widehat{e(\chi)}$ and $V(\alpha) = H(\alpha)$. We show that V is an interpretation function on $\mathfrak{F}$. V is a Boolean homomorphism by the properties of ultrafilters and it is a Σ-homomorphism from Tm to $\mathbf{H}$ by the definition of $H(\alpha)$. We would like to establish the leftmost equalities in the following:

(i) If $[\gamma]\chi \in \Gamma$, then $V([\gamma]\chi) = [V(\gamma)]V(\chi) = \{w \mid \forall r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$; and

(ii) if $[\gamma]\chi \in \Gamma$, then $V(\langle\gamma]\chi) = \langle V(\gamma)]V(\chi) = \{w \mid \exists r \in H(\gamma, w) : r(w) \subseteq \widehat{e(\chi)}\}$.

(i) can be show easily by induction on the complexity of γ. However, similarly to the extensional case, only the $\supseteq$ inclusion of (ii) can be shown to hold for $\mathfrak{F}^\Gamma$. Here we can use the standard technique of splitting to transform $\mathfrak{F}^\Gamma$ into a correct relational cs-frame (see the extended version of [2]). We omit the details.

$(\top \approx \varphi) \in Eq(\mathsf{FCS})$ implies $\varphi \in LCS$. We define the *canonical LCS-frame* $\mathfrak{C}^{LCS}$ similarly as we defined $\mathfrak{C}^{LSL}$ in the third part of the proof of Theorem 3, but we use *LCS* instead of *LSL*, of course. We have to show in our specific setting that (i) $\equiv^{Tm}$ is a congruence and that (ii) $\mathbf{G}$ is a closure semilattice. Both items are checked using the corresponding closure axioms, and the rule (to show that $^\cap$ is a monotonic operator on the $\equiv^{Tm}$-quotient algebra).

□

Sequential Language-based Decisions

Adam Bjorndahl
Department of Philosophy
Carnegie Mellon University
Pittsburgh, USA
abjorn@cmu.edu

Joseph Y. Halpern
Department of Computer Science
Cornell Univeristy
Ithaca, USA
halpern@cs.cornell.edu

In earlier work, we introduced the framework of *language-based decisions*, the core idea of which was to modify Savage's classical decision-theoretic framework [6] by taking actions to be descriptions in some language, rather than functions from states to outcomes, as they are defined classically. Actions had the form "if ψ then $do(\varphi)$", where ψ and φ were formulas in some underlying language, specifying what effects would be brought about under what circumstances. The earlier work allowed only one-step actions. But, in practice, plans are typically composed of a sequence of steps. Here, we extend the earlier framework to *sequential* actions, making it much more broadly applicable. Our technical contribution is a representation theorem in the classical spirit: agents whose preferences over actions satisfy certain constraints can be modeled as if they are expected utility maximizers. As in the earlier work, due to the language-based specification of the actions, the representation theorem requires a construction not only of the probability and utility functions representing the agent's beliefs and preferences, but also the state and outcomes spaces over which these are defined, as well as a "selection function" which intuitively captures how agents disambiguate coarse descriptions. The (unbounded) depth of action sequencing adds substantial interest (and complexity!) to the proof.

1 Background and motivation

In earlier work, we introduced the framework of *language-based decisions* [2], the core idea of which was to modify Savage's classical decision-theoretic framework [6] by taking actions to be descriptions in some language, rather than functions from states to outcomes, as they are defined classically. Actions had the form "if φ then $do(\psi)$", where φ and ψ were formulas in some underlying language, specifying what effects would be brought about under what circumstances.[1] For example, a statement like "If there is a budget surplus then $do(MW = 15)$ else *no-op*" would be an action in this framework, where $MW = 15$ represents the minimum wage being \$15, and *no-op* is the action of doing nothing. The effect of the action $do(MW = 15)$ is to bring about a state where the minimum wage is \$15. But this does not completely specify the state. (Do businesses close? Is there more automation so jobs are lost? Are no jobs lost and more people move into the state?)

In this context, we proved a representation theorem in the classical spirit: agents whose preferences over actions satisfy certain constraints can be modeled as if they are expected utility maximizers. This requires constructing not only probability and utility functions (as is done classically), but also the state and outcome spaces on which these functions are defined, and a *selection function* that describes which state will result from an underspecified action like $do(MW = 15)$. In this construction the state and outcome spaces coincide; intuitively, this is because the tests that determine whether an action is performed ("If φ then...") and the actions themselves ("$do(\psi)$") are described using the same language.

[1] This work in turn extended previous work by Blume, Easley, and Halpern [3] in which the tests in actions, but not the effects of actions, were specified in a formal language.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 131–141, doi:10.4204/EPTCS.379.12

The earlier work allowed only one-step actions. But, in practice, plans are typically composed of a sequence of steps, and we must choose among such plans: Do I prefer to walk to the cafe and then call my friend if the cafe is open, or would it be better to call my friend first, then walk to the cafe and call them back if it's closed? Should I ring the doorbell once, or ring it once and then a second time if no one replies to the first? Here, we extend the earlier framework to *sequential* actions, making it much more broadly applicable.

At a technical level, a decision-theoretic framework in which the state and outcome spaces coincide is the perfect setting in which to implement sequential actions, since—given that the actions are understood as functions—we have an immediate and natural way to "put them in sequence", namely, by composing the corresponding functions.

Our contribution in this paper is, first, to lay the mathematical groundwork for reasoning about sequential, language-based actions (Section 2), and second, to prove a representation theorem analogous to earlier such results (Section 3): roughly speaking, agents whose preferences over sequential actions satisfy certain axioms can be understood as if their preferences are derived by maximizing the expected value of a suitable utility function. Proving this result is substantially harder in the present setting, owing to the more complex nature of sequential actions (including but not limited to the fact that we allow sequential nesting to be arbitrarily deep). The reader is thus forewarned that the main result depends on a fairly lengthy, multi-stage proof.

2 Sequential language-based actions

The framework presented in this section is an expansion of that developed in [2]. We begin with the same simple, formal language: let Φ denote a finite set of *primitive propositions*, and $\mathcal{L}$ the propositional language consisting of all Boolean combinations of these primitives. A **basic model (over $\mathcal{L}$)** is a tuple $M = (\Omega, [\![\cdot]\!]_M)$ where Ω is a nonempty set of *states* and $[\![\cdot]\!]_M : \Phi \to 2^{\Omega}$ is a *valuation function*. This valuation is recursively extended to all formulas in $\mathcal{L}$ in the usual way, so that intuitively, each formula φ is "interpreted" as the "event" $[\![\varphi]\!]_M \subseteq \Omega$. We sometimes drop the subscript when the model is clear from context, and write $\omega \models \varphi$ rather than $\omega \in [\![\varphi]\!]$. We say that φ is *satisfiable in M* if $[\![\varphi]\!]_M \neq \emptyset$ and that φ is *valid in M* if $[\![\varphi]\!]_M = \Omega$; we write $\models \varphi$ to indicate that φ is valid in all basic models.

Given a finite set of formulas $F \subseteq \mathcal{L}$, the set of **(sequential) actions (over F)**, denoted by $\mathcal{A}_F$, is defined recursively as follows:

(1) for each $\varphi \in F$, $do(\varphi)$ is an action (called a *primitive action*);

(2) *no-op* is an action (this is short for "no operation"; intuitively, it is a "do nothing" action);

(3) for all $\psi \in \mathcal{L}$ and $\alpha, \beta \in \mathcal{A}_F$, not both *no-op*, **if** ψ **then** α **else** β is an action;

(4) for all $\alpha, \beta \in \mathcal{A}_F$, not both *no-op*, $\alpha;\beta$ is an action (intuitively, this is the action "do α and then do β").

In [2], actions were defined only by clauses (1) and (3). The idea of "sequencing" actions is of course not new; the semicolon notation is standard in programming languages.

It will also be useful for our main result to have a notion of the *depth* of an action, which intuitively should capture how deeply nested the sequencing is. We do so by induction. The only **depth-**0 action is *no-op*. A **depth-**1 **action** is either (1) *no-op*; (2) a primitive action $do(\varphi)$; or (3) an action of the form **if** ψ **then** α **else** β, where α and β are depth-1 actions. Now suppose that we have defined depth-k actions for $k \geq 1$; a **depth-($k+1$)** action is either (1) a depth-k action; (2) an action of the form **if** ψ **then** α **else** β, where α and β are depth-($k+1$) actions; or (3) an action of the form $\alpha;\beta$, where α is a depth-k_1

action, β is a depth-k_2 action, and $k_1 + k_2 \leq k+1$. Note that we have defined depth in such a way that the depth-k actions include all the depth-k' actions for $k' < k$, and so that **if...then** constructions do not increase depth—only sequencing does.

As in [2], given a basic model $M = (\Omega, [\![\cdot]\!]_M)$, we want $do(\varphi)$ to correspond to a function from Ω to Ω whose range is contained in $[\![\varphi]\!]_M$. For this reason we restrict our attention to basic models in which each $\varphi \in F$ is satisfiable, so that $[\![\varphi]\!]_M \neq \emptyset$; call such models **$F$-rich**. Moreover, in order for $do(\varphi)$ to pick out a *function*, we need some additional structure that determines, for each $\omega \in \Omega$, which state in $[\![\varphi]\!]_M$ the function corresponding to $do(\varphi)$ should actually map to. This is accomplished using a *selection function* $sel : \Omega \times F \to \Omega$ satisfying $sel(\omega, \varphi) \in [\![\varphi]\!]_M$.

The intuition for selection functions is discussed in greater detail in [2]. Briefly: $do(\varphi)$ says that φ should be made true, but there may be many ways of making φ true (i.e., many states one could transition to in which φ is true); *sel* picks out which of these φ-states to actually move to. In this way we can think of *sel* as serving to "disambiguate" the meaning of the primitive actions, which are inherently underspecified.

Note that selection functions are formally identical to the mechanism introduced by Stalnaker [9] to interpret counterfactual conditionals. In our context, we can think of a selection function as another component of an agent's model of the world, to be constructed in the representation theorem: in addition to a probability measure (to represent their beliefs) and a utility function (to capture their preferences), we will also need a selection function (to specify how they interpret actions).

A **selection model (over** F**)** is an F-rich basic model M together with a selection function $sel : \Omega \times F \to \Omega$ satisfying $sel(\omega, \varphi) \in [\![\varphi]\!]_M$. Given a selection model (M, sel) over F, we define the *interpretation* of $do(\varphi)$ to be the function $[\![do(\varphi)]\!]_{M,sel} : \Omega \to \Omega$ given by:

$$[\![do(\varphi)]\!]_{M,sel}(\omega) = sel(\omega, \varphi).$$

This interpretation can then be extended to all sequential actions in $\mathcal{A}_F$ in the obvious way:

$$[\![\textbf{if } \psi \textbf{ then } \alpha \textbf{ else } \beta]\!]_{M,sel}(\omega) = \begin{cases} [\![\alpha]\!]_{M,sel}(\omega) & \text{if } \omega \in [\![\psi]\!] \\ [\![\beta]\!]_{M,sel}(\omega) & \text{if } \omega \notin [\![\psi]\!], \end{cases}$$

and

$$[\![\alpha;\beta]\!]_{M,sel} = [\![\beta]\!]_{M,sel} \circ [\![\alpha]\!]_{M,sel}.$$

3 Representation

Let $\succeq$ be a binary relation on $\mathcal{A}_F$, where we understand $\alpha \succeq \beta$ as saying that α is "at least as good as" β from the agent's subjective perspective. Intuitively, such a binary relation is meant to be reasonably "accessible" to observers, "revealed" by how an agent chooses between binary options. As usual, we define $\alpha \succ \beta$ as an abbreviation of $\alpha \succeq \beta$ and $\beta \not\succeq \alpha$, and $\alpha \sim \beta$ as an abbreviation of $\alpha \succeq \beta$ and $\beta \succeq \alpha$; intuitively, these relations represent "strict preference" and "indifference", respectively.

We assume that $\succeq$ is a *preference order*, so is *complete* (i.e., for all acts $\alpha, \beta \in \mathcal{A}_F$, either $\alpha \succeq \beta$ or $\beta \succeq \alpha$) and transitive. Note that completeness immediately gives reflexivity as well. While there are good philosophical reasons to consider incomplete relations (see [4] and the references therein), for the purposes of this paper we adopt the assumption of completeness in order to simplify the (already quite involved) representation result.

A **language-based SEU (Subjective Expected Utility) representation** for a relation $\succeq$ on $\mathcal{A}_F$ is a selection model (M, sel) together with a probability measure Pr on Ω and a utility function $u : \Omega \to \mathbb{R}$ such that, for all $\alpha, \beta \in \mathcal{A}_F$,

$$\alpha \succeq \beta \Leftrightarrow \sum_{\omega \in \Omega} \Pr(\omega) \cdot u([\![\alpha]\!]_{M,sel}(\omega)) \geq \sum_{\omega \in \Omega} \Pr(\omega) \cdot u([\![\beta]\!]_{M,sel}(\omega)). \tag{1}$$

Our goal is to show that such a representation exists if the preference order satisfies one key axiom, discussed below.

3.1 Canonical maps and canonical actions

For each $a \subseteq \Phi$, let

$$\varphi_a = \bigwedge_{p \in a} p \wedge \bigwedge_{q \notin a} \neg q,$$

so φ_a settles the truth values of each primitive propositions in the language $\mathcal{L}$: it says that p is true iff it belongs to a. An **atom** is a formula of the form φ_a.[2] Since we are working with a classical propositional logic, it follows that for all formulas $\varphi \in \mathcal{L}$ and atoms φ_a, the truth of φ is determined by φ_a: either $\models \varphi_a \to \varphi$, or $\models \varphi_a \to \neg\varphi$. In the framework of [2], it followed that every action could be identified with a function from atoms to elements of F, since atoms determine whether the tests in an action hold. In our context, however, things are not so simple: actions can be put in sequence, so even though an atom may tell us which tests at the "first layer" hold, so to speak, it may not be enough to tell us which later tests hold. For example, in an action like "if p then $do(r)$ else $do(r')$; if q then $do(\neg r)$", the atom that currently holds determines whether p holds, but tells us nothing about whether q will hold when we get around to doing the second action in the sequence.

To deal with this, we need an outcome space that is richer than just F (i.e., richer than the set of all primitive actions); roughly speaking, we will instead identify actions with functions from atoms to "canonical" ways of describing the sequential structure of actions. We now make this precise.

Suppose that $|2^\Phi| = N$, so there are N atoms; call them $a_1, \ldots, a_N$. For each subset A of atoms, let $\varphi_A = \bigvee_{a \in A} \varphi_a$. A basic fact of propositional logic is that for every formula φ, there is a unique set A of atoms such that φ is logically equivalent to φ_A. Let $\tilde{F} = \{\varphi_A : (\exists \varphi \in F)(\models \varphi_A \leftrightarrow \varphi)\}$.

We want to associate with each action α of depth k a **canonical action** γ_α of depth k that is, intuitively, equivalent to α. The canonical action γ_α makes explicit how α acts in a state characterized by an atom a. We define γ_α by induction on the structure of α. It is useful in the construction to simultaneously define the **canonical map** c_α associated with α, a function from atoms to actions such that, for all atoms a, $c_\alpha(a)$ has the form *no-op*, $do(\varphi_A)$, or $do(\varphi_A); \gamma_\beta$ for some set A of atoms and action β. Intuitively, c_α defines how α acts in a state characterized by an atom a. For example, if α is **if** a **then** $do(\varphi_A)$ **else** β, then $c_\alpha(a) = do(\varphi_A)$.

If $\alpha = \textit{no-op}$, then $\gamma_{\textit{no-op}} = \textit{no-op}$ and $c_{\textit{no-op}}$ is the constant function such that $c_{\textit{no-op}}(a) = \textit{no-op}$ for all atoms a. If α is a depth-1 action other than *no-op*, then we define c_α by induction on structure:

$$\begin{aligned} c_{do(\varphi)}(a) &= do(\varphi_A), \text{ where } A \text{ is the unique subset of atoms such that } \models \varphi_A \leftrightarrow \varphi \\ c_{\textbf{if } \psi \textbf{ then } \alpha \textbf{ else } \beta}(a) &= \begin{cases} c_\alpha(a) & \text{if } \models \varphi_a \to \psi \\ c_\beta(a) & \text{if } \models \varphi_a \to \neg\psi. \end{cases} \end{aligned}$$

[2] Not to be confused with *atomic propositions*, which is another common name for primitive propositions.

The action γ_α is the depth-1 action defined as follows:

$$\gamma_\alpha = \textbf{if } \varphi_{a_1} \textbf{ then } c_\alpha(a_1) \textbf{ else (if } \varphi_{a_2} \textbf{ then } c_\alpha(a_2) \textbf{ else (} \cdots \textbf{(if } \varphi_{a_{N-1}} \textbf{ then } c_\alpha(a_{N-1}) \textbf{ else } c_\alpha(a_N)\textbf{)} \cdots \textbf{))}$$

if at least one of $c_\alpha(a_{N-1})$ or $c_\alpha(a_N)$ is not *no-op*. If both are *no-op*, then **if** $\varphi_{a_{N-1}}$ **then** $c_\alpha(a_{N-1})$ **else** $c_\alpha(a_N)$ is not an action according to our definitions; in this case, we take

$$\gamma_\alpha = \textbf{if } \varphi_{a_1} \textbf{ then } c_\alpha(a_1) \textbf{ else (if } \varphi_{a_2} \textbf{ then } c_\alpha(a_2) \textbf{ else (} \cdots \textbf{(if } \varphi_{a_m} \textbf{ then } c_\alpha(a_m) \textbf{ else } \textit{no-op}\textbf{))},$$

where m is the least index such that $c_\alpha(a_m) \neq \textit{no-op}$. (If $c_\alpha(a_m) = \textit{no-op}$ for all m, then $\gamma_\alpha = \textit{no-op}$.)

If α is a depth-$(k+1)$ action other than *no-op*, then we again define c_α by induction on structure:

$$\begin{aligned} c_{\textbf{if } \psi \textbf{ then } \alpha \textbf{ else } \beta}(a) &= \begin{cases} c_\alpha(a) & \text{if } \models \varphi_a \rightarrow \psi \\ c_\beta(a) & \text{if } \models \varphi_a \rightarrow \neg\psi \end{cases} \\ c_{\alpha;\beta}(a) &= \begin{cases} c_\beta(a) & \text{if } c_\alpha(a) = \textit{no-op} \\ do(\varphi_A);\gamma_\beta & \text{if } c_\alpha(a) = do(\varphi_A) \\ do(\varphi_A);\gamma_{\beta';\beta} & \text{if } c_\alpha(a) = do(\varphi_A);\gamma_{\beta'}. \end{cases} \end{aligned}$$

The canonical action γ_α is defined as above for the depth-1 case.

We take CA^k to be the set of canonical actions of depth k, and CM^k to be the set of canonical maps that correspond to some depth-k action. Finally, let $CA^{k,-}$ consist of all depth k-actions of the form *no-op*, $do(\varphi_A)$, or $do(\varphi_A);\gamma_\beta$, where β is a depth $(k-1)$-action. Note that if α is a depth-k action and a is an atom, then $c_\alpha(a) \in CA^{k,-}$. Observe that since the set of atoms is finite, as is $\tilde{F}$, it follows that for all k, CM^k, CA^k, and $CA^{k,-}$ are also finite. This will be crucial in our representation proof.

3.2 Cancellation

As in [2, 3], the key axiom in our representation theorem is what is known as a *cancellation axiom*, although the details differ due to the nature of our actions. Simple versions of the cancellation axiom go back to [5, 7]; our version, like those used in [2, 3], has more structure. See [3] for further discussion of the axiom.

The axiom uses multisets. Recall that a *multiset*, intuitively, is a set that allows for multiple instances of each of its elements. Thus two multisets are equal just in case they contain the same elements *with the same multiplicities*. We use "double curly brackets" to denote multisets, so for instance $\{\{a,b,b\}\}$ is a multiset, and it is distinct from $\{\{a,a,b\}\}$: both have three elements, but the mulitiplicity of a and b differ. With that background, we can state the axiom:

(Canc) Let $\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n \in \mathcal{A}_F$, and suppose that for each $a \subseteq \Phi$ we have

$$\{\{c_{\alpha_1}(a),\dots,c_{\alpha_n}(a)\}\} = \{\{c_{\beta_1}(a),\dots,c_{\beta_n}(a)\}\}.$$

Then, if for all $i < n$ we have $\alpha_i \succeq \beta_i$, it follows that $\beta_n \succeq \alpha_n$.

Intuitively, this says that if we get the same outcomes (counting multiplicity) using the canonical maps for $\alpha_1,\dots,\alpha_n$ as for $\beta_1,\dots,\beta_n$ in each state, then we should view the collections $\{\{\alpha_1,\dots,\alpha_n\}\}$ and $\{\{\beta_1,\dots,\beta_n\}\}$ as being "equally good", so if α_i is at least as good as β_i for $i = 1,\dots,n-1$, then, to balance things out, β_n should be at least as good as α_n. How intuitive this is depends on how intuitive one finds the association $\alpha \mapsto c_\alpha$ defined above; if the map c_α really does capture "everything decision-theoretically relevant" about the action α, then cancellation does seem reasonable.

In particular, it is not hard to show that whenever α and β are such that $\gamma_\alpha = \gamma_\beta$ (which of course is equivalent to $c_\alpha = c_\beta$), cancellation implies that $\alpha \sim \beta$. In other words, any information about α and β that is lost in the transformation to canonical actions is also forced to be irrelevant to decisionmaking. This means that **(Canc)** entails, among other things, that agents do not distinguish between logically equivalent formulas (since, e.g., when $\models \varphi \leftrightarrow \varphi'$, it's easy to see that $\gamma_{do(\varphi)} = \gamma_{do(\varphi')}$).

3.3 Construction

Theorem 1. *If $\succeq$ is a preference order on $\mathcal{A}_F$ satisfying* **(Canc)***, then there is a language-based subjective expected utility representation of $\succeq$.*

Proof. As in [2], we begin by following the proof in [3, Theorem 2], which says that if a preference order on a set of acts mapping a finite state space to a finite outcome space satsifies the cancellation axiom, then it has a state-dependent representation. "State-dependent" here means that the utility function constructed depends jointly on both states and outcomes, in a sense made precise below. To apply this theorem in our setting, we first fix k and take CM^k to be the set of acts. With this viewpoint, the state space is the set of atoms and the outcome space is $CA^{k,-}$; as we observed, both are finite.

The relation $\succeq$ on $\mathcal{A}_F$ induces a relation $\succeq^k$ on CM^k defined in the natural way:

$$c_\alpha \succeq^k c_\beta \Leftrightarrow \alpha \succeq \beta.$$

As noted, **(Canc)** implies that $\alpha \sim \alpha'$ whenever $c_\alpha = c_{\alpha'}$, from which it follows that $\succeq^k$ is well-defined. To apply Theorem 2 in [3], it must also be the case that $\succeq^k$ is a preference order and satisfies cancellation, which is immediate from the definition of $\succeq^k$ and the fact that $\succeq$ is a preference order and satisfies cancellation. It therefore follows that $\succeq^k$ has a state-dependent representation; that is, there exists a real-valued utility function v^k defined on state-outcome pairs such that, for all depth-k actions α and β,

$$c_\alpha \succeq^k c_\beta \text{ iff } \textstyle\sum_{i=1}^N v^k(a_i, c_\alpha(a_i)) \geq \sum_{i=1}^N v^k(a_i, c_\beta(a_i)).a \tag{2}$$

It follows from our definitions that for all depth-k actions α and β,

$$\alpha \succeq \beta \text{ iff } \textstyle\sum_{i=1}^N v^k(a_i, c_\alpha(a_i)) \geq \sum_{i=1}^N v^k(a_i, c_\beta(a_i)).$$

As we observed, we needed to restrict to depth-k actions here in order to ensure that the outcome space is finite, which is necessary to apply Theorem 2 in [3].

Our next goal is to define a selection model $M = (\Omega^k, [\![\cdot]\!]_M, sel)$, a probability $\Pr^k$ on Ω^k, and a utility function u^k on Ω^k such that, for all actions α and β of depth k,

$$\alpha \succeq \beta \text{ iff } \sum_{\omega \in \Omega^k} \Pr^k(\omega) u^k([\![\alpha]\!]_{M,sel}(\omega)) \geq \sum_{\omega \in \Omega^k} \Pr^k(\omega) u^k([\![\beta]\!]_{M,\mathrm{sec}}(\omega)). \tag{3}$$

Eventually, we will construct a single (state and outcome) space Ω^*, a probability $\Pr^*$ on Ω^*, and a utility u^* on Ω^* that we will use to provide a single representation theorem for all actions, without the restriction to depth k, but we seem to need to construct the separate spaces first.

As a first step to defining Ω^k, define a *labeled k-tree* to be a balanced tree of depth k whose root is labeled by an atom such that each non-leaf node has exactly N children, labeled $a_1, \ldots, a_N$, respectively. An *ordered labeled k-tree (k-olt)* is a labeled k-tree where, associated with each non-leaf node, there is a total order on its children. We assume that in different labeled k-trees, the nodes come from the same set, and corresponding nodes have the same label, so there is a unique labeled k-tree and k-olts differ only in the total order associated with each non-leaf node and the label of the root. Let T^k consist of all k-olts. For $k' \geq k$, a (k')-olt $s^{k'}$ *extends* (or *is an extension of*) a k-olt s^k if s^k is the prefix of $s^{k'}$ of depth k; we call s^k the *projection* of $s^{k'}$ onto depth k.

The intuition behind a k-olt is the following: the atom associated with the root r describes what is true before an action is taken. For each non-leaf node t, the total order associated with t on the children of t describes the selection function at t (with children lower in the order considered "closer"). For example, suppose that there are two primitive propositions, p and q. Then there are four atoms. If we take the action $do(\varphi)$ starting at r, we want to "move" to the "closest" child of r satisfying φ, which is the child lowest in the ordering associated with r. For example, suppose that the total order on the atoms associated with r is $\neg p \wedge q < \neg p \wedge \neg q < p \wedge \neg q < p \wedge q$. Then if we take the action $do(p \vee q)$ starting at r, we move to the child labeled with the atom $\neg p \wedge q$; if we instead do $do(p \vee \neg q)$, we move to the child labeled $\neg p \wedge \neg q$; and if instead we do **if** q **then** $do(p \vee q)$ **else** $do(p \vee \neg q)$, which of these two children we move to depends on whether q is true at the atom labeling r. For an action $do(p \vee q); do(p \vee \neg q)$, we move further down the tree. The first action, $do(p \vee q)$, takes us to the child t of r labeled $\neg p \wedge q$. We then take the action $do(p \vee \neg q)$ from there, which gets us to a child of t. Which one we get to depends on the ordering of the children of t associated with t.

It turns out that our states must be even richer than this; they must in addition include a *k-progress function* g that maps each node t in a k-olt s^k to a descendant of t in s^k. We give the intuition behind progress functions shortly. We take Ω^k to consist of all pairs (s^k, g), where $s^k \in T^k$ and g is a k-progress function and for each primitive proposition $p \in \Phi$, we define

$$[\![p]\!] = \{(s^k, g) \,:\, p \in a, \text{ where } a \text{ labels } g(r) \text{ and } r \text{ is the root of } s^k\}.$$

We now want to associate with each depth-k action α a function $f_\alpha : \Omega^k \to \Omega^k$; intuitively, this is the transition on states that we want to be induced by the selection function. To begin, we define f_α only on states of the form (s^k, id), where id is the identity function. We take $f_\alpha(s^k, id) = (s^k, g_{\alpha,s^k})$, where g_{α,s^k} is defined formally below. Intuitively, if t is a node at depth k' of s^k, then $g_{\alpha,s^k}(t)$ describes the final state if the action α were to (possibly counterfactually) end up at the node t after running for k' steps, and then continued running.

Given a k-olt s^k whose root r is labeled a and an action α of depth at most k, we define $g_{\alpha,s^k}(t)$ by induction on the depth of α. For the base case, we take $g_{no\text{-}op,s^k} = id$. Now suppose inductively that α has depth m and we have defined g_{α',s^k} for all actions α' of depth $m-1$. There are three cases to consider. (1) If $c_\alpha(a) = no\text{-}op$, then $g_{\alpha,s^k} = id$. (2) If $c_\alpha(a) = do(\varphi_A)$, then $g_{\alpha,s^k}(r)$ is the "closest" (i.e., minimal) child t' of r among those labelled by an atom in A, according to the total order labeling r; $g_{\alpha,s^k}(t) = t$ for all nodes $t \neq r$. (3) If $c_\alpha(a) = do(\varphi_A); \gamma_\beta$ (which means β is an action of depth at most $m-1$), then $g_{\alpha,s^k}(r) = g_{\beta,s^{k,t'}}(t')$, where $t' = g_{do(\varphi_A),s^k}(r)$ and $s^{k,t'}$ is the $(k-1)$-subolt of s^k rooted at t'. The intuition here is that $g_{\alpha,s^k}(r)$ is supposed to output the descendent of r that is reached by doing α; the fact that $c_\alpha(a) = do(\varphi_A); \gamma_\beta$ tells us that the way α works (in a state where a holds) is by first making φ_A true, and then following up with β. This means we must first move to the "closest" child of r where φ_A holds, which is t', and subsequently moving to whichever descendant of t' that β directs us to (which is defined, by the inductive hypothesis). Finally, if $t \neq r$, let t'' be the first step on the (unique) path from r to t and let $s^{k,t''}$ be the $(k-1)$-subolt of s^k rooted at t''. Then $g_{\alpha,s^k}(t) = g_{\beta,s^{k,t''}}(t)$ (where, once again, this is defined by the inductive hypothesis). This essentially forces us to "follow" the unique path from r to t, and then continue from that point by doing whatever the remaining part of the action α demands. It is clear from this definition that if the root of s^k is labeled by a, then $g_{\alpha,s^k} = g_{c_\alpha(a),s^k}$.

We now extend f_α to states of the form (s^k, g_{β,s^k}) by setting $f_\alpha(s^k, g_{\beta,s^k}) = f_{\beta;\alpha}(s^k, id)$. Intuitively, the state (s^k, g_{β,s^k}) is a state where β has "already happened" (i.e., it's the state we would arrive at by doing β in (s^k, id)) so doing α in this state should be the same as doing first β then α in (s^k, id).

Observe that a k-progress function g_{α,s^k} not only tells us the node that α would reach if it started at the root of s^k, but also gives a great deal of counterfactual information about which nodes would be reached starting from anywhere in s^k. This is in the same spirit as *subgame-perfect equilibrium* [8], which can depend on what happens at states that are never actually reached in the course of play, but could have been reached if play had gone differently. Like this game-theoretic notion, our representation theorem requires a kind of counterfactual information.

In light of (2), to prove (3), it suffices to define our selection function sel so that $[\![\alpha]\!]_{M,sel} = f_\alpha$, and find $\Pr^k$ and u^k such that for all actions α of depth k,

$$\sum_{i=1}^{N} v^k(a_i, c_\alpha(a_i)) = \sum_{(s^k,g)\in\Omega^k} \Pr^k(s^k,g)u^k(f_\alpha(s^k,g)). \tag{4}$$

Our definition of f_α is set up to make defining the right selection function straightforward: we simply set $sel((s^k,g),\varphi) = f_{do(\varphi_A)}(s^k,g)$, where A is the unique set of atoms such that $\models \varphi_A \leftrightarrow \varphi$. It is then easy to check that $[\![\alpha]\!]_{M,sel} = f_\alpha$.

Define $\Pr^k(s^k,g) = 0$ if $g \neq id$, and $\Pr^k(s^k,id) = 1/|T^k|$ for all $s^k \in T^k$. Given this, to establish (4), it suffices to define u^k such that for all actions α of depth k,

$$|T^k|\sum_{i=1}^{N} v^k(a_i, c_\alpha(a_i)) = \sum_{s^k\in T^k} u^k(f_\alpha(s^k,id)). \tag{5}$$

Given an atom a, let T^k_a consist of all k-olts whose root is labeled by a. By definition of f_α, to prove (5), it suffices to prove, for each atom $a \in \{a_1,\dots,a_N\}$ and all actions α of depth k, that

$$|T^k|v^k(a, c_\alpha(a)) = \sum_{s^k\in T^k_a} u^k(s^k, g_{\alpha,s^k}) = \sum_{s^k\in T^k_a} u^k(s^k, g_{c_\alpha(a),s^k}), \tag{6}$$

where the second equality follows from the fact, observed above, that $g_{\alpha,s^k} = g_{c_\alpha(a),s^k}$ whenever $s^k \in T^k_a$.

Since v^k is given, for each depth-k action $c_\alpha(a)$, the left-hand side of (6) is just a number. Replace each term $u^k(s^k, g_{c_\alpha(a),s^k})$ for $s^k \in T^k_a$ by the variable $x_{s^k, g_{c_\alpha(a),s^k}}$. This gives us a system of linear equations, one for each action $c_\alpha(a)$, with variables $x_{s^k,g}$, where the coefficient of $x_{s^k,g}$ in the equation corresponding to action α is either 1 or 0, depending on whether $g_{\alpha,s^k} = g$. We want to show that this system has a solution.

We can describe the relevant equations as the product $MX = U$ of matrices, where M is a matrix whose entries are either 0 or 1, and X is a vector of variables (namely, the variables $x_{s^k,g}$). The matrix M has one row corresponding to each action in $CA^{k,-}$ (since, for all actions α of depth k, $c_\alpha(a) \in CA^{k,-}$), and one column corresponding to each state (s^k,g) with $s^k \in T^k_a$. The entry in M in the row corresponding to the action γ_α and the column corresponding to (s^k,g) is 1 if $g_{\gamma_a,s^k} = g$ (i.e., if $f_\alpha(s^k,id) = (s^k, g_{\alpha,s^k}) = (s^k,g)$) and 0 otherwise. A basic result of linear algebra tells us that this system has a solution if the rows of the matrix M (viewed as vectors) are independent. We now show that this is the case.

Let r_α be the row of M indexed by action $\alpha \in CA^{k,-}$. Suppose that a linear combination of rows is 0; that is, $\sum_\alpha d_\alpha r_\alpha = 0$, for some scalars d_α. The idea is to put a partial order $\sqsupset$ on $CA^{k,-}$ and show by induction on $\sqsupset$ that for all $\alpha \in CA^k$, the coefficient $d_\alpha = 0$.

We define $\sqsupset$ as follows. We take *no-op* to be the minimal element of $\sqsupset$. For actions $\alpha = do\varphi_A;\gamma_\beta$ and $\alpha' = do(\varphi_{A'});\gamma_{\beta'}$ (where we take γ_β to be *no-op* if $\alpha = do(\varphi_A)$ and similarly for $\gamma_{\beta'}$), $\alpha \sqsupset \alpha'$ iff either (1) $A \supsetneq A'$, (2) $A = A'$, $\beta \neq$ *no-op*, and $\beta' =$ *no-op*, or (3) $A = A'$, $c_\beta(a) \sqsupset c_{\beta'}(a)$ for all atoms a.

We show that $d_\alpha = 0$ by induction assuming that $d_{\alpha'} = 0$ for all actions $\alpha' \in CA^{k,-}$ such that $\alpha \sqsupset \alpha'$. For the base case, $\alpha =$ *no-op*. Consider the k-progress function $g^k_{no\text{-}op}$ such that $g^k_{no\text{-}op}(t) = t$ for all nodes t in a k-olt. Note that $g_{no\text{-}op}(s^k, id) = g^k_{no\text{-}op}$ for all k-olts s^k. It is easy to see that if β has the form $do(\varphi_A)$ or $do(\varphi_A);\gamma_{\beta'}$, then for all k-olts s^k, $g_{\beta,s^k} \neq g^k_{no\text{-}op}$ (since for the root r of s^k, $g_{\beta,s^k}(r) \neq r$). Thus, the entry of $r_{no\text{-}op}$ corresponding to the column $(s^k, g^k_{no\text{-}op})$ is 1, while the entry of d_β for $\beta \neq$ *no-op* corresponding to this column is 0. It follows that $d_{no\text{-}op} = 0$.

For the general case, suppose that we have an arbitrary action $\alpha \neq$ *no-op* in $CA^{k,-}$ and $d_{\alpha'} = 0$ for all $\alpha' \in CA^k$ such that $\alpha \sqsupset \alpha'$. We now define a k-olt $s^{k,\alpha} \in T^k_a$ such that if $g_{\alpha',s^{k,\alpha}} = g_{\alpha,s^{k,\alpha}}$ and $\alpha \neq \alpha'$, then $\alpha \sqsupset \alpha'$, so $d_{\alpha'} = 0$ by the induction hypothesis. Once we show this, it follows that $d_\alpha = 0$ (since otherwise the entry in $\sum_{\alpha'} d_{\alpha'} r_{\alpha'}$ corresponding to $g_{\alpha,s^{k,\alpha}}$ would be nonzero). We construct $s^{k,\alpha}$ by induction on the depth of α. If α has depth 1 and is not *no-op*, it must be of the form $do(\varphi_A)$ for some set A of atoms. Suppose that $b \in A$. Let the total order at the root of $s^{k,\alpha}$ be such that the final elements in the order are the elements in A, and b is the first of these. For example, if $A = \{b,c,d\}$, we could consider an order where the final three elements are b, c, and d (or b, d, and c). Note that if r is the root of $s^{k,\alpha}$, then $g_{\alpha,s^{k,\alpha}}(r)$ is the child t_b of r labeled b. Now consider an action α' of the form $do(\varphi_{A'});\beta$ (β may be *no-op*). If A' contains an element not in A, then $g_{\alpha,s^{k,\alpha}}(r) \neq t_b$ (because there will be an atom in A' that is greater than b in the total order at r). If $A' \subset A$, then $\alpha \succ \alpha'$, as desired. And if $A = A'$ and $\alpha \neq \alpha'$, then $\alpha' = \varphi_A;\gamma_\beta$ and $\beta \neq$ *no-op*, so it is easy to see that $g_{\alpha',s^{k,\alpha}}(r) \neq t_b = g_{\alpha,s^{k,\alpha}}(r)$.

Suppose that $m > 1$ and we have constructed $s^{k,\beta}$ for all actions $\beta \in CA^{k,-}$ of depth less than m. We now show how to construct $s^{k,\alpha}$ for actions $\alpha \in CA^{k,-}$ of depth m that are not of depth $m-1$. This means that α must have the form $do(\varphi_A);\beta$. We construct the total order at r as above, and at the subtree of $s^{k,\alpha}$ whose root is the child of r labeled a, we use the same orderings as in $s^{k-1,c_\beta(a)}$, which by the induction hypothesis we have already determined. It now follows easily from the induction hypothesis that if $g_{\alpha',s^{k,\alpha}} = g_{\alpha,s^{k,\alpha'}}$ and $\alpha \neq \alpha'$, then $\alpha \sqsupset \alpha'$. This completes the argument for (6).

The argument above gives us a representation theorem for each k that works for actions of depth k. However, we are interested in a single representation theorem that works for all actions of all depths simultaneously. The first step is to make the state-dependent utility functions $v^1, v^2, \ldots$ that we began with (one utility function for each k in the argument above) *v-compatible*, in the sense that if α is a depth-k action and $k' > k$, then $v^k(a, c_\alpha(a)) = v^{k'}(a, c_\alpha(a))$. That is, we want to construct a sequence $(v^1, v^2, v^3, \ldots)$ of v-compatible utility functions, each of which satisfies (2). We proceed as follows.

We can assume without loss of generality that each utility function has range in $[0,1]$, by applying an affine transformation. (Doing this would not affect (2).) For each utility function v^k let v^{ki}, for $i \leq k$, be the restriction of v^k to actions of depth i. Thus, $v^{kk} = v^k$. Now consider the sequence $v^{11}, v^{21}, v^{31}, \ldots$ It must have a convergent subsequence, say $v^{m_1,1}, v^{m_2,1}, v^{m_3,1}, \ldots$. Say it converges to w^1. Now consider the subsequence $v^{m_2,2}, v^{m_3,2}, \ldots$. (We omit $v^{m_1,2}$, since we may have $m_2 = 1$, in which case $v^{m_1,2}$ is not defined.) It too has a convergent subsequence. Say it converges to w^2. Continuing this process, for each k, we can find a convergent subsequence, which is a subsequence of the sequence we found for $k-1$. It is easy to check that the limits $w^1, w^2, w^3, \ldots$ of these convergent subsequences satisfy (2) and are v-compatible (since, in general, v^{ki} is v-compatible with v^{kj} for $i, j \leq k$). For the remainder of this discussion, we assume without loss of generality that the utility functions in the sequence $v^1, v^2, \ldots$ are v-compatible.

Note that it follows easily from our definition that probability measures in the sequence $\Pr^1, \Pr^2, \ldots$

are Pr-*compatible* in the following sense: If $k' > k$, $(s^k, id) \in \Omega^k$, and $E^{k'}(s^k, id)$ consists of all the pairs $(t^{k'}, id)$ such that s^k is the projection of $t^{k'}$ onto depth k, then $\Pr^k(s^k, id) = \Pr^{k'}(E^{k'}(s^k, id))$. We will also want a third type of compatibility among the utility functions. To make this precise, define a k'-progress function g to be *k-bounded* for $k < k'$ if for all nodes t of depth $\leq k$, we have that $g(t)$ has depth $\leq k$, and if the depth of t is greater than k, then $(t) = t$. Note that if α is a depth-k action, then $g_{\alpha,s^{k'}}$ is k-bounded. If $k' > k$ and g is a k-bounded k'-progress function, then g has an obvious projection to a k-progress function. We want the utility functions in the sequence $u^1, u^2, \ldots$ that satisfies (4) to be *u-compatible* in the following sense: if g' is a k'-progress function that is k-bounded, g is the projection of g' to a k-progress function, and s^k is the projection of $t^{k'}$ onto depth k, then $u^k(s^k, g) = u^{k'}(t^{k'}, g')$. We can assume without loss of generality that the utility functions in the sequence $u^1, u^2, \ldots,$ are u-compatible. For given a sequence $u^1, u^2, \ldots$, define the sequence $w^1, w^2, \ldots$ as follows. Let $w^1 = u^1$. Suppose that we have defined $w^1, \ldots, w^k$. If the $(k+1)$-progress function g' is k-bounded, define $w^{k+1}(t^{k+1}, g') = w^k(s^k, g)$, where s^k is the projection of t^{k+1} onto depth k and g is the projection of g' to a k-progress function; if g is not k-bounded, define $w^{k+1}(t^{k+1}, g') = u^{k+1}(t^{k+1}, g')$. Clearly the sequence $w^1, w^2, \ldots$ is u-compatible. Moreover, it is easy to check that $(\Pr^k, w^k)$ satisfies (4).

We are now ready to define a single state space. Define an *∞-olt* just like a k-olt, except that now the tree is unbounded, rather than having depth k. Let Ω^∞ consist of all pairs (s^∞, g), where s^∞ is an ∞-olt and g is a k-bounded progress function for some k. This will be our state space. Define $E^\infty(s^k, id)$ by obvious analogy to $E^{k'}(s^k, id)$: it consists of all the pairs (t^∞, id) such that t^∞ extends s^k. Then, by Carathéodory's extension theorem [1] there is a measure $\Pr^\infty$ on the smallest σ-algebra extending the algebra generated by the sets $E^\infty(s^k, id)$ which agrees with $\Pr^k$ for all k (i.e., $\Pr^k(s^k, id) = 1/|T^k| = \Pr^\infty(E^\infty(s^k, id))$. Let u^∞ be defined by taking $u^\infty(s^\infty, g) = u^k(s^k, g^k)$ if g is k-bounded and s^k is the unique k-olt that s^∞ extends. It is easy to check that this is well-defined (note that if g is k-bounded then g is k'-bounded for $k' > k$, so there is something to check here). Finally, it is easy to check that for a depth-k action α, we have that the expected utility of α is

$$\sum_{(s^k,g)\in\Omega^k} \Pr^\infty(E^\infty(s^k, id))u^\infty(s^k, g_{\alpha,s^k}) = \sum_a v^k(a, c_\alpha(a)),$$

giving us the desired result. □

4 Conclusion and Future Work

We have extended the results of [2] to allow for actions that are composed of sequences of steps, and proved a representation theorem in this setting. More precisely, we have shown that when an agent's language-based preferences satisfy a suitably formulated cancellation axiom, they are acting as if they are an expected utility maximizer with respect to some background state space Ω, a probability and utility over Ω, and a selection function on Ω that serves to "disambiguate" the results of actions described in the language. Allowing for (possibly unbounded) sequences of steps made the proof significantly more complicated.

In [2], we also considered axioms regarding the preference order $\succeq$ that restricted properties of the selection function in ways that are standard in the literature on counterfactual conditions (e.g., being *centered*, so that $sel(\omega, \varphi) = \omega$ whenever $\omega \models \varphi$). Although we have not checked details yet, we believe it will be straightforward to provide axioms that similarly restrict the selection function in our setting, and to extend the representation theorem appropriately.

We also believe it is of interest in some contexts to consider more complex sequential actions, such as "do φ until ψ". This opens the door for potentially *non-terminating* actions, which of course will add further complexity to the analysis.

Finally, and perhaps most urgently, while the cancellation axiom is quite amazing in the power it has, it is not particularly intuitive. As shown in [3], more intuitive axioms can be derived from cancellation, such as transitivity of the relation $\succeq$ or the classic principle of *independence of irrelevant alternatives* (see [6]). In order to bring the technical results of this project more in line with everyday intuitions about preference, it would be very beneficial to "factor" the cancellation axiom into weaker, but easier to intuit, components. This is the subject of ongoing research.

Acknowledgments

Joe Halpern was supported in part by ARO grant W911NF-22-1-0061, MURI grant W911NF-19-1-0217, AFOSR grant FA9550-12-1-0040, and a grant from the Algorand Centres of Excellence program managed by Algorand Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

[1] R. B. Ash (1970): *Basic Probability Theory*. Wiley, New York.

[2] A. Bjorndahl & J. Y. Halpern (2021): *Language-based decisions*. In: Theoretical Aspects of Rationality and Knowledge: Proc. Eighteenth Conference (TARK 2021). The proceedings are published as *Electronic Proceedings in Theoretical Computer Science*. doi:10.4204/EPTCS.335.5

[3] L. E. Blume, D. Easley & J. Y. Halpern (2021): *Connstructive decision theory*. Journal of Economic Theory 196. An earlier version, entitled "Redoing the Foundations of Decision Theory", appears in *Principles of Knowledge Representation and Reasoning: Proc. Tenth International Conference (KR '06)*. doi:10.1016/j.jet.2021.105306

[4] J. Dubra, F. Maccheroni & E.A. Ok (2004): *Expected utility theory without the completeness axiom*. Journal of Economic Theory 115, pp. 118–133. doi:10.1016/S0022-0531(03)00166-2

[5] D. H. Krantz, R. D. Luce, P. Suppes & A. Tversky (1971): *Foundations of Measurement, Vol 1: Additive and Polynomial Representations*. Academic Press, New York.

[6] L. J. Savage (1954): *Foundations of Statistics*. Wiley, New York.

[7] D. Scott (1964): *Measurement structures and linear inequalities*. Journal of Mathematical Psychology 1, pp. 233–247. doi:10.1016/0022-2496(64)90002-1

[8] R. Selten (1975): *Reexamination of the perfectness concept for equilibrium points in extensive games*. International Journal of Game Theory 4, pp. 25–55. doi:10.1007/BF01766400

[9] R. C. Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: Studies in Logical Theory, Blackwell, Oxford, U.K., pp. 98–112. doi:10.1007/978-94-009-9117-0_2

Characterization of AGM Belief Contraction in Terms of Conditionals

Giacomo Bonanno*

Department of Economics
University of California
Davis, California, USA

gfbonanno@ucdavis.edu

We provide a semantic characterization of AGM belief contraction based on frames consisting of a Kripke belief relation and a Stalnaker-Lewis selection function. The central idea is as follows. Let K be the initial belief set and $K \div \phi$ be the contraction of K by the formula ϕ; then $\psi \in K \div \phi$ if and only if, at the actual state, the agent believes ψ and believes that if $\neg\phi$ is (were) the case then ψ is (would be) the case.

1 Introduction

Belief contraction is the operation of removing from the set K of initial beliefs a particular belief ϕ. One reason for doing so is, for example, the discovery that some previously trusted evidence supporting ϕ was faulty. For instance, a prosecutor might form the belief that the defendant is guilty on the basis of his confession; if the prosecutor later discovers that the confession was extorted, she might abandon the belief of guilt, that is, become open minded about whether the defendant is guilty or not. In their seminal contribution to belief change, Alchourrón, Gärdenfors and Makinson ([1]) defined the notion of ”rational and minimal” contraction by means of a set of eight properties, known as the AGM axioms or postulates. They did so within a syntactic approach where the initial belief set K is a consistent and deductively closed set of propositional formulas and the result of removing ϕ from K is a new set of propositional formulas, denoted by $K \div \phi$.

We provide a new characterization of AGM belief contraction based on a so-far-unnoticed connection between the notion of belief contraction and the Stalnaker-Lewis theory of conditionals ([34, 21]). Stalnaker introduced the notion of a selection function f taking as input a possible world w and a set of worlds E (representing a proposition) and giving as output a world $w' = f(w,E) \in E$, interpreted as the closest E-world to w (an E-world is a world that belongs to E). Lewis generalized this by allowing $f(w,E)$ to be a set of worlds. In the Stalnaker-Lewis theory the (indicative or subjunctive) conditional ”if ϕ is (were) the case then ψ is (would be) the case”, denoted by $\phi > \psi$, is declared to be true at a world w if and only if ψ is true at all the worlds in $f(w, \|\phi\|)$ ($\|\phi\|$ denotes the set of worlds at which ϕ is true).

We consider semantic frames consisting of a Kripke belief relation on a set of states S, representing the agent's initial beliefs, and a Stalnaker-Lewis selection function on $S \times 2^S$ representing conditionals. Adding a valuation to such a frame yields a model. Given a model, we define the initial belief set K as the set of formulas that the agent believes at the actual state and $K \div \phi$ (the contraction of K by ϕ) as the set of formulas that the agent believes initially and also on the supposition that $\neg\phi$: $\psi \in K \div \phi$ if and only if, at the actual state, the agent (1) believes ψ and (2) believes the conditional $\neg\phi > \psi$. We show that, when the selection function satisfies some natural properties, the contraction operation so defined captures precisely the set of AGM belief contraction functions.

[*] I am grateful to three anonymous reviewers for their comments.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 142–156, doi:10.4204/EPTCS.379.13

2 AGM contraction functions

Let At be a countable set of atomic formulas. We denote by Φ_0 the set of Boolean formulas constructed from At as follows: At $\subset \Phi_0$ and if $\phi, \psi \in \Phi_0$ then $\neg\phi$ and $\phi \vee \psi$ belong to Φ_0. Define $\phi \to \psi$, $\phi \wedge \psi$, and $\phi \leftrightarrow \psi$ in terms of $\neg$ and $\vee$ in the usual way.

Given a subset K of Φ_0, its deductive closure $Cn(K) \subseteq \Phi_0$ is defined as follows: $\psi \in Cn(K)$ if and only if there exist $\phi_1, ..., \phi_n \in K$ (with $n \geq 0$) such that $(\phi_1 \wedge ... \wedge \phi_n) \to \psi$ is a tautology. A set $K \subseteq \Phi_0$ is *consistent* if $Cn(K) \neq \Phi_0$; it is *deductively closed* if $K = Cn(K)$. Given a set $K \subseteq \Phi_0$ and a formula $\phi \in \Phi_0$, the *expansion* of K by ϕ, denoted by $K + \phi$, is defined as follows: $K + \phi = Cn(K \cup \{\phi\})$.

Let $K \subseteq \Phi_0$ be a consistent and deductively closed set representing the agent's initial beliefs and let $\Psi \subseteq \Phi_0$ be a set of formulas representing possible candidates for withdrawal. A *belief contraction function* (based on K and Ψ) is a function $\div_\Psi : \Psi \to 2^{\Phi_0}$ (where 2^{Φ_0} denotes the set of subsets of Φ_0) that associates with every formula $\phi \in \Psi$ a set $K \div_\Psi \phi \subseteq \Phi_0$ (interpreted as the result of removing ϕ from K). If $\Psi \neq \Phi_0$ then $\div_\Psi$ is called a *partial* contraction function, while if $\Psi = \Phi_0$ then $\div_{\Phi_0}$ is called a *full-domain* contraction function; in this case we simplify the notation and omit the subscript Φ_0.

Definition 1. *Let $\div_\Psi : \Psi \to 2^{\Phi_0}$ be a partial contraction function and $\div' : \Phi_0 \to 2^{\Phi_0}$ a full-domain contraction function (both of them based on K). We say that $\div'$ is an* extension *of $\div_\Psi$ if, for every $\phi \in \Psi$, $K \div' \phi = K \div_\Psi \phi$.*

A *full-domain* contraction function is called an *AGM contraction function* if it satisfies the following properties, known as the AGM postulates:

$(K-1)$ [Closure] $K \div \phi = Cn(K \div \phi)$.
$(K-2)$ [Inclusion] $K \div \phi \subseteq K$.
$(K-3)$ [Vacuity] If $\phi \notin K$ then $K \subseteq K \div \phi$.
$(K-4)$ [Success] If ϕ is not a tautology, then $\phi \notin K \div \phi$.
$(K-5)$ [Recovery] If $\phi \in K$ then $K \subseteq (K \div \phi) + \phi$.
$(K-6)$ [Extensionality] If $\phi \leftrightarrow \psi$ is a tautology, then $K \div \phi = K \div \psi$.
$(K-7)$ [Conjunctive overlap] $(K \div \phi) \cap (K \div \psi) \subseteq K \div (\phi \wedge \psi)$.
$(K-8)$ [Conjunctive inclusion] If $\phi \notin K \div (\phi \wedge \psi)$, then $K \div (\phi \wedge \psi) \subseteq K \div \phi$.

$(K-1)$ requires the result of contracting K by ϕ to be a deductively closed set.

$(K-2)$ requires the contraction of K by ϕ not to contain any beliefs that were not in K.

$(K-3)$ requires that if ϕ is not in the initial belief set, then every belief in K should also be present in $K \div \phi$ (thus, by $(K-2)$ and $(K-3)$, if $\phi \notin K$ then the contraction of K by ϕ coincides with K).

$(K-4)$ requires that ϕ not be contained in $K \div \phi$, unless ϕ is a tautology (in which case, by $(K-1)$, it must be in $K \div \phi$).

$(K-5)$ is a conservativity requirement: when $\phi \in K$, contracting by ϕ and then expanding the resulting set $K \div \phi$ by ϕ should involve no loss of beliefs relative to K (the converse inclusion $(K \div \phi) + \phi \subseteq K$ follows from $(K-2)$ and the hypothesis that $K = Cn(K)$).

$(K-6)$ says that logically equivalent formulas should lead to the same result in terms of contraction.

By $(K-7)$, if a formula $\chi \in K$ is neither removed in the contraction of K by ϕ nor in the contraction of K by ψ, then χ should not be removed in the contraction of K by the conjunction $\phi \wedge \psi$.

$(K-8)$, on the other hand, requires that if ϕ is removed when we contract by $\phi \wedge \psi$, then every formula that survives the contraction of K by $\phi \wedge \psi$ survives also when K is contracted by ϕ alone.

For an extensive discussion of the above postulates see [11, 17, 6].

The notion of AGM belief contraction has been given alternative characterizations. One characterization is in terms of a binary relation $\leqslant$ of ”epistemic entrenchment” on K, with the interpretation of $\phi \leqslant \psi$ as ”ϕ is either less entrenched than, or as entrenched as, ψ”. Gärdenfors ([11, Theorem 4.30, p. 96]) shows that if the relation $\leqslant$ satisfies five properties and a contraction function is defined by ‘$\psi \in K \div \phi$ if and only if $\psi \in K$ and either ϕ is a tautology or $\phi < (\phi \vee \psi)$’, then such contraction function is an AGM contraction function and, conversely, if an AGM contraction function is used to define the relation $\leqslant$ by ‘$\phi \leqslant \psi$ if and only if either $\phi \notin K \div (\phi \wedge \psi)$ or $\phi \wedge \psi$ is a tautology’ then such relation satisfies those five properties. Another characterization makes use of the set W of possible worlds, where a possible world is defined as a maximally consistent set of formulas in Φ_0; within this approach, contraction has been characterized either in terms of systems of spheres ([13, 21]) or in terms of a plausibility relation on W or in terms of propositional selection functions (see [6, Chapter 4]).

In this paper we provide an alternative characterization in terms of Stalnaker-Lewis conditionals.

3 An alternative semantic characterization of AGM contraction

Given a binary relation $R \subseteq S \times S$ on a set S, for every $s \in S$ we define $R(s) = \{x \in S : (s,x) \in R\}$.

Definition 2. *A* pointed frame *is a quadruple $\langle S, s_@, \mathscr{B}, f \rangle$ where*

1. *S is a set of* states*; subsets of S are called* events.
2. *$s_@ \in S$ is a distinguished element of S interpreted as the* actual state.
3. *$\mathscr{B} \subseteq S \times S$ is a binary* belief relation *on S which is serial: $\forall s \in S$, $\mathscr{B}(s) \neq \varnothing$.*
4. *$f : \mathscr{B}(s_@) \times 2^S \setminus \varnothing \to 2^S$ is a* Stalnaker-Lewis selection function[1] *that associates with every state-event pair (s,E) (with $s \in \mathscr{B}(s_@)$ and $\varnothing \neq E \subseteq S$) a set of states $f(s,E) \subseteq S$ such that,*
   - *(a) (a.1) $f(s,E) \neq \varnothing$ and (a.2) (Success) $f(s,E) \subseteq E$,*
   - *(b) (Weak Centering) if $s \in E$ then $s \in f(s,E)$,*
   - *(c) (Doxastic Priority 1) if $\mathscr{B}(s_@) \cap E \neq \varnothing$ then $f(s,E) \subseteq \mathscr{B}(s_@) \cap E$,*
   - *(d) (Intersection) $f(s,E) \cap F \subseteq f(s, E \cap F)$,*
   - *(e) (Doxastic Priority 2) Let $B_{EF} = \{s \in \mathscr{B}(s_@) : f(s,E) \cap F \neq \varnothing\}$. If $B_{EF} \neq \varnothing$ then*
     - *(e.1) if $s \in B_{EF}$ then $f(s, E \cap F) \subseteq f(s,E) \cap F$,*
     - *(e.2) if $s \notin B_{EF}$ then $f(s, E \cap F) \subseteq f(\hat{s}, E \cap F)$ for some $\hat{s} \in B_{EF}$.*

The set $\mathscr{B}(s)$ is the set of states that the agent considers possible at state s, so that $\mathscr{B}(s_@)$ is the set of doxastic possibilities at the actual state $s_@$ and represents the agent's initial beliefs. $f(s,E)$ is the set of states that the agent considers closest, or most similar, to state s conditional on event E.

(4.a) of Definition 2 requires $f(s,E)$ to be non-empty and, furthermore, that every state in $f(s,E)$ be an E-state.

(4.b) postulates that if s is an E-state then it belongs to $f(s,E)$, that is, s itself is one of the E-states that are closest to s.

By (4.c) if there exists an E-state among those initially considered possible ($\mathscr{B}(s_@) \cap E \neq \varnothing$), then, for every $s \in \mathscr{B}(s_@)$, the closest E-states to s must belong to $\mathscr{B}(s_@) \cap E$.

By (4.d), the closest E-states to s that are also F-states must belong to the set of closest $(E \cap F)$-states to

[1] Note that, for the purpose of this paper, the domain of f can be taken to be $\mathscr{B}(s_@) \times 2^S \setminus \varnothing$ rather than $S \times 2^S \setminus \varnothing$. However, it can easily be extended to $S \times 2^S \setminus \varnothing$ as follows: first, fix an arbitrary function $g : S \setminus \mathscr{B}(s_@) \to \mathscr{B}(s_@)$ and then define, for every $s \in S \setminus \mathscr{B}(s_@)$ and every $\varnothing \neq E \subseteq S$, $f(s,E) = f(g(s),E)$.

s.
(4.e) can be viewed as an extension of (4.c): it says that if, among the states initially considered possible, there is at least one state, call it s, that satisfies the property that among its closest E-states there is at least one that is also an F-state, then (1) the closest $(E\cap F)$-states to s must belong to the intersection $f(s,E)\cap F$ and (2) for any other state that does not satisfy the property, the closest $(E\cap F)$-states to it are contained in the set of closest $(E\cap F)$-states to some state that does satisfy the property.

Adding a valuation to a pointed frame yields a model. Thus a *model* is a tuple $\langle S,s_@,\mathscr{B},f,V\rangle$ where $\langle S,s_@,\mathscr{B},f\rangle$ is a pointed frame and $V:\mathtt{At}\to 2^S$ is a valuation that assigns to every atomic formula $p\in\mathtt{At}$ the set of states where p is true. Given a model $\langle S,s_@,\mathscr{B},f,V\rangle$ define truth of a Boolean formula $\phi\in\Phi_0$ at a state $s\in S$, denoted by $s\models\phi$, in the usual way:

Definition 3. *Truth of a formula at a state is defined as follows:*

1. *if $p\in\mathtt{At}$ then $s\models p$ if and only if $s\in V(p)$,*
2. *$s\models\neg\phi$ if and only if $s\not\models\phi$,*
3. *$s\models\phi\vee\psi$ if and only if $s\models\phi$ or $s\models\psi$ (or both),*

We denote by $\|\phi\|$ the truth set of ϕ: $\|\phi\|=\{s\in S:s\models\phi\}$.

Fix a model $M=\langle S,s_@,\mathscr{B},f,V\rangle$ and let $K=\{\phi\in\Phi_0:\mathscr{B}(s_@)\subseteq\|\phi\|\}$ (to simplify the notation, we omit the subscript denoting the model and thus write K rather than K_M); thus a Boolean formula ϕ belongs to K if and only if at the actual state $s_@$ the agent believes ϕ. It is shown in the Appendix (Lemma 1) that the set $K\subseteq\Phi_0$ so defined is deductively closed and consistent. Next, for every $\phi\in\Phi_0$ such that $\|\neg\phi\|\neq\varnothing$, define $K\div\phi\subseteq\Phi_0$ as follows:

$$\psi\in K\div\phi \text{ if and only if } \quad \begin{array}{l}(1)\,\mathscr{B}(s_@)\subseteq\|\psi\|, \text{ and}\\ (2)\,\forall s\in\mathscr{B}(s_@), f(s,\|\neg\phi\|)\subseteq\|\psi\|.\end{array} \tag{1}$$

In (2) below we rewrite (1) in an extended language containing a belief operator and a conditional operator, thus making the interpretation more transparent: $\psi\in K\div\phi$ if and only if, at the actual state $s_@$, the agent believes ψ initially as well as on the supposition that $\neg\phi$.[2].

Since, in general, not every $\phi\in\Phi_0$ is such that $\|\neg\phi\|\neq\varnothing$, this definition gives rise to a *partial* belief contraction function. The next proposition says that this partial contraction function can be extended to a full-domain AGM contraction function; conversely, given a full-domain AGM contraction function based on a consistent and deductively closed set K, there exists a model $M=\langle S,s_@,\mathscr{B},f,V\rangle$ such that $K=\{\phi\in\Phi_0:\mathscr{B}(s_@)\subseteq\|\phi\|\}$ and, for every $\phi\in\Phi_0$ such that $\|\neg\phi\|\neq\varnothing$, $K\div\phi$ satisfies (1). Thus the proposed semantics provides an alternative characterization of AGM belief contraction. The proof of the following proposition is given in the Appendix.

Proposition 1.

(A) *Given a model $\langle S,s_@,\mathscr{B},f,V\rangle$ let $K=\{\phi\in\Phi_0:\mathscr{B}(s_@)\subseteq\|\phi\|\}$ and, for every $\phi\in\Phi_0$ such that $\|\neg\phi\|\neq\varnothing$, let $K\div\phi$ be defined by* (1). *Then K is consistent and deductively closed and the (partial) belief contraction function so defined can be extended to a full-domain AGM belief contraction function.*

(B) *Let $K\subset\Phi_0$ be consistent and deductively closed and let $\div:\Phi_0\to 2^{\Phi_0}$ be an AGM belief contraction function. Then there exists a model $\langle S,s_@,\mathscr{B},f,V\rangle$ such that $K=\{\phi\in\Phi_0:\mathscr{B}(s_@)\subseteq\|\phi\|\}$ and, for every $\phi\in\Phi_0$ such that $\|\neg\phi\|\neq\varnothing$, $K\div\phi$ satisfies* (1).

[2] We take "believing ψ on the supposition that $\neg\phi$" to mean "believing that if $\neg\phi$ is (were) the case then ψ is (would be) the case".

The proposed semantics becomes more transparent if we extend the language by introducing two modal operators: a unimodal belief operator $\mathbb{B}$, corresponding to the belief relation $\mathscr{B}$, and a bimodal conditional operator $>$, corresponding to the selection function f. Recall that Φ_0 is the set of Boolean (or factual) formulas. Let Φ_1 be the modal language constructed as follows.

- $\Phi_0 \subset \Phi_1$,
- if $\phi, \psi \in \Phi_0$ then $\phi > \psi \in \Phi_1$,
- all the Boolean combinations of formulas in Φ_1.

Thus, for the purpose of this paper, the conditional $\phi > \psi$ (interpreted as the indicative or subjunctive conditional ”if ϕ is (were) the case then ψ is (would be) the case”) is defined only for Boolean formulas. Finally, let Φ be the modal language constructed as follows:

- $\Phi_1 \subset \Phi$,
- if $\phi \in \Phi_1$ then $\mathbb{B}\phi \in \Phi$,
- all the Boolean combinations of formulas in Φ.

Thus formulas in Φ are either Boolean or formulas of the form $\phi > \psi$, with ϕ and ψ Boolean, or of the form $\mathbb{B}\phi$ where ϕ is either Boolean or of the form $\psi > \chi$ with ψ and χ Boolean, or a Boolean combination of such formulas. We can now extend the definition of truth of a formula at a state (Definition 3) to the set Φ as follows:

Definition 4. *If $\phi \in \Phi_0$ then $s \models \phi$ according to the rules of Definition 3. Furthermore,*

- *$s \models (\phi > \psi)$ (with $\phi, \psi \in \Phi_0$) if and only if either $\|\phi\| = \varnothing$, or $\|\phi\| \neq \varnothing$ and $f(s, \|\phi\|) \subseteq \|\psi\|$,*
- *$s \models \mathbb{B}\phi$ if and only if $\mathscr{B}(s) \subseteq \|\phi\|$.*

Then we can re-write the definition of $K \div \phi$ given in (1) in terms of the modal operators $\mathbb{B}$ and $>$ as follows:

$$\psi \in K \div \phi \text{ if and only if } \phi, \psi \in \Phi_0 \text{ and } s_@ \models \mathbb{B}\psi \wedge \mathbb{B}(\neg\phi > \psi)\,. \tag{2}$$

Thus, in the statement of Proposition 1, $K = \{\phi \in \Phi_0 : \mathscr{B}(s_@) \subseteq \|\phi\|\}$ can be replaced by $K = \{\phi \in \Phi_0 : s_@ \models \mathbb{B}\phi\}$ and reference to (1) can be replaced by reference to (2). Note that only a fragment of the extended language is used in the characterization result of Proposition 1. In particular, nesting of conditionals and beliefs is disallowed. The study of whether the extended language can be used to obtain generalizations of AGM-style belief change that go beyond merely Boolean expressions is a topic left for future research.

4 Related literature

There is a vast literature that deals with AGM belief contraction (for a survey see, for example, [5, 6]). Because of space limitations we will only focus on a few issues.

The recovery postulate (AGM axiom $(K-5)$) appears to be a natural way of capturing a “minimal” way of suspending belief in ϕ, but has been subject to extensive scrutiny (see [25, 9, 14, 19, 22, 28, 15, 16]). In Makinson’s terminology ([25]), contraction operations that do not satisfy the recovery postulates are called *withdrawals*. Alternative types of withdrawal operators have been studied in the literature: contraction without recovery ([4]), semi-contraction ([7]), severe withdrawal ([32]), systematic withdrawal ([26]), mild contraction ([20]). If one interprets belief contraction as a form of *actual belief change* (in response to some input), then perhaps the recovery postulate is open to scrutiny. However, in the interpretation of belief contraction proposed in this paper, the recovery postulate is entirely natural. Indeed, if ψ belongs to the contraction of K by ϕ then ψ is believed both initially and on the supposition that $\neg\phi$; if this supposition is removed then one naturally falls back to the initial beliefs K.

There have been attempts in the literature to establish a link between notions of AGM belief change and Stalnaker-Lewis conditionals. Within the context of AGM belief revision this was done by [10], who considered the language that we called Φ_1, which includes conditionals of the form $\phi > \psi$. Gäerdenfors introduced the following postulate (where $K * \phi$ denotes the revised belief set in response to information ϕ): $(\phi > \psi) \in K$ if and only if $\psi \in K{*}\phi$. This postulate was taken to be an expression of the so-called Ransey test.[3] Gäerdenfors showed that this postulate can be satisfied only in cases where the revision operation is trivial; in other words, there cannot be interesting revision theories based on conditionals if one requires that the conditionals themselves be incorporated in the initial belief set. Several attempts have been made to circumvent Gäerdenfors' "triviality result". Different routes have been taken: weakening or re-interpretating the theorem ([22, 24, 23, 30, 31], generalizing from belief revision functions to belief change systems (consisting of a set of epistemic states, an assignment of a belief set to each epistemic state and a transition function function that determines how the epistemic state changes as a result of learning new information: [8]), considering an alternative semantics, namely Moss and Parikh's epistemic logic of subsets logic ([27]), and augmenting it with conditionals ([12]), and, in the context of iterated belief contraction, defining the notion of "contractional" in the context of belief states ([33]: if Ψ denotes a belief state and $[\beta|\alpha]$ is interpreted as "belief in β even in the absence of α", then the contractional is defined as $\Psi \models [\beta|\alpha]$ if and only if $\Psi \div \alpha \models \beta$). None of the approaches described above coincides with the framework considered in this paper.

5 Conclusion

We proposed a semantic characterization of AGM belief contraction in terms of a semantics consisting of a Kripke belief relation $\mathscr{B}$ (with associated modal operator $\mathbb{B}$) and a Stalnaker-Lewis selection function f (with associated conditional bimodal operator $>$). The proposed semantics can also be used to characterize AGM belief revision (see [2]). Indeed all three operations: belief expansion, belief contraction and belief revision, can be captured within this framework. Letting $s_@$ denote the actual state, we have:

1. Expansion: $\psi \in K + \phi$ if and only if $s_@ \models \neg\mathbb{B}\neg\phi \wedge \mathbb{B}(\phi \rightarrow \psi)$,
2. Contraction: $\psi \in K \div \phi$ if and only if $s_@ \models \mathbb{B}\psi \wedge \mathbb{B}(\neg\phi > \psi)$,
3. Revision: $\psi \in K * \phi$ if and only if $s_@ \models \mathbb{B}(\phi > \psi)$.

There are several issues that can be studied within this framework and are left for future work, for example, whether the extended modal language can provide a way to generalize AGM-style belief change and whether the proposed framework can accommodate iterated belief contraction/revision.

A Appendix

In this Appendix we prove Proposition 1. In order to make the proof entirely self-contained we include the proofs of known auxiliary results (e.g. the lemmas).[4]

[3] The expression "Ramsey Test" refers to the following passage from [29, p. 247]: "If two people are arguing "If p will q?" and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q".

[4] Which can be found, for example, in [11, 17].

Lemma 1. *Fix a model $M = \langle S, s_@, \mathscr{B}, f, V\rangle$ and let $K = \{\phi \in \Phi_0 : \mathscr{B}(s_@) \subseteq \|\phi\|\}$. Then K is deductively closed and consistent.*

Proof. First we show that K is deductively closed, that is, $K = Cn(K)$. If $\psi \in K$ then $\psi \in Cn(K)$, because $\psi \to \psi$ is a tautology; thus $K \subseteq Cn(K)$. To show that $Cn(K) \subseteq K$, let $\psi \in Cn(K)$, that is, there exist $\phi_1, ..., \phi_n \in K$ ($n \geq 0$) such that $(\phi_1 \wedge ... \wedge \phi_n) \to \psi$ is a tautology. Since $\|\phi_1 \wedge ... \wedge \phi_n\| = \|\phi_1\| \cap ... \cap \|\phi_n\|$ and, for all $i = 1, ..., n$, $\phi_i \in K$ (that is, $\mathscr{B}(s_@) \subseteq \|\phi_i\|$), it follows that $\mathscr{B}(s_@) \subseteq \|\phi_1 \wedge ... \wedge \phi_n\|$. Since $(\phi_1 \wedge ... \wedge \phi_n) \to \psi$ is a tautology, $\|(\phi_1 \wedge ... \wedge \phi_n) \to \psi\| = S$, that is, $\|\phi_1 \wedge ... \wedge \phi_n\| \subseteq \|\psi\|$. Thus $\mathscr{B}(s_@) \subseteq \|\psi\|$, that is, $\psi \in K$. Next we show that $Cn(K) \neq \Phi_0$, that is, K is consistent. Let $p \in \mathtt{At}$ be an atomic formula. Then $\|p \wedge \neg p\| = \varnothing$. By seriality of $\mathscr{B}$, $\mathscr{B}(s_@) \neq \varnothing$ so that $\mathscr{B}(s_@) \nsubseteq \|p \wedge \neg p\|$, that is, $(p \wedge \neg p) \notin K$ and hence, since $K = Cn(K)$, $(p \wedge \neg p) \notin Cn(K)$. □

Proof of Part (A) of Proposition 1.
Fix a model $\langle S, s_@, \mathscr{B}, f, V\rangle$ and let $K = \{\phi \in \Phi_0 : \mathscr{B}(s_@) \subseteq \|\phi\|\}$ and, for every $\phi \in \Phi_0$ such that $\|\neg\phi\| \neq \varnothing$, let $K \div \phi$ be defined as follows ((A1) below reproduces (1) above):

$$\psi \in K \div \phi \text{ if and only if } \quad \begin{array}{l}(1)\, \mathscr{B}(s_@) \subseteq \|\psi\|, \text{ (that is, } \psi \in K\text{) and} \\ (2)\, \forall s \in \mathscr{B}(s_@), f(s, \|\neg\phi\|) \subseteq \|\psi\|.\end{array} \tag{A1}$$

Let '$\div'$' be following extension to Φ_0 of the operator '$\div$' defined in (A1):

$$K \div' \phi = \begin{cases} K \div \phi & \text{if } \|\neg\phi\| \neq \varnothing \\ K \cap Cn(\neg\phi) & \text{if } \|\neg\phi\| = \varnothing. \end{cases} \tag{A2}$$

We want to show that the contraction operator defined in (A2) satisfies the AGM axioms.

(K−1) We need to show that, for every $\phi \in \Phi_0$, $K \div' \phi = Cn(K \div' \phi)$. If $\|\neg\phi\| = \varnothing$ then this is true by construction, since K is deductively closed and the intersection of deductively closed sets is deductively closed. Assume, therefore, that $\|\neg\phi\| \neq \varnothing$, so that $K \div' \phi = K \div \phi$. Note first that, by (A1), letting

$$\Psi_{\neg\phi} = \{\psi \in \Phi_0 : f(s, \|\neg\phi\|) \subseteq \|\psi\|, \forall s \in \mathscr{B}(s_@)\}, \tag{A3}$$

$K \div \phi = K \cap \Psi_{\neg\phi}$. Since the intersection of two deductively closed sets is deductively closed and K is deductively closed, it suffices to show that $\Psi_{\neg\phi}$ is deductively closed, that is, $\Psi_{\neg\phi} = Cn(\Psi_{\neg\phi})$. The inclusion $\Psi_{\neg\phi} \subseteq Cn(\Psi_{\neg\phi})$ follows from the fact that, for every $\chi \in \Psi_{\neg\phi}$, $\chi \to \chi$ is a tautology. Next we show that $Cn(\Psi_{\neg\phi}) \subseteq \Psi_{\neg\phi}$. Since $\|\neg\phi\| \neq \varnothing$, $f(s, \|\neg\phi\|)$ is defined for every $s \in \mathscr{B}(s_@)$. Fix an arbitrary $\psi \in Cn(\Psi_{\neg\phi})$; then there exist $\phi_1, ..., \phi_n \in \Psi_{\neg\phi}$ ($n \geq 0$) such that $(\phi_1 \wedge ... \wedge \phi_n) \to \psi$ is a tautology, so that $\|(\phi_1 \wedge ... \wedge \phi_n) \to \psi\| = S$, that is, $\|\phi_1 \wedge ... \wedge \phi_n\| \subseteq \|\psi\|$. Fix an arbitrary $s \in \mathscr{B}(s_@)$ and an arbitrary $i = 1, ..., n$. Then, since $\phi_i \in \Psi_{\neg\phi}$, $f(s, \|\neg\phi\|) \subseteq \|\phi_i\|$. Hence $f(s, \|\neg\phi\|) \subseteq \|\phi_1 \wedge ... \wedge \phi_n\|$. Since $\|(\phi_1 \wedge ... \wedge \phi_n)\| \subseteq \|\psi\|$ it follows that $f(s, \|\neg\phi\|) \subseteq \|\psi\|$, that is, $\psi \in \Psi_{\neg\phi}$.

(K−2) We need to show that $K \div' \phi \subseteq K$. If $\|\neg\phi\| = \varnothing$ then $K \div' \phi = K \cap Cn(\neg\phi) \subseteq K$. If $\|\neg\phi\| \neq \varnothing$ then $K \div' \phi = K \div \phi = K \cap \Psi_{\neg\phi} \subseteq K$.

(K−3) We need to show that if $\phi \notin K$ then $K \subseteq K \div' \phi$. Assume that $\phi \notin K$, that is, $\mathscr{B}(s_@) \cap \|\neg\phi\| \neq \varnothing$. Then $\|\neg\phi\| \neq \varnothing$ and thus $K \div' \phi = K \div \phi$. Fix an arbitrary $\psi \in K$, that is, $\mathscr{B}(s_@) \subseteq \|\psi\|$. We need to show that, $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \|\psi\|$. Since $\mathscr{B}(s_@) \cap \|\neg\phi\| \neq \varnothing$, by 4(c) of Definition 2, for every $s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \mathscr{B}(s_@) \cap \|\neg\phi\|$ and thus, since $\mathscr{B}(s_@) \subseteq \|\psi\|$, $f(s, \|\neg\phi\|) \subseteq \|\psi\|$.

(K−4) We need to show that if ϕ is not a tautology then $\phi \notin K \dot{-}' \phi$. Suppose that ϕ is not a tautology, so that $\phi \notin Cn(\neg\phi)$. If $\|\neg\phi\| = \varnothing$ then $K \dot{-}' \phi = K \cap Cn(\neg\phi)$ and thus $\phi \notin K \dot{-}' \phi$. Next, suppose that $\|\neg\phi\| \neq \varnothing$ so that $K \dot{-}' \phi = K \dot{-} \phi$. Since $K \dot{-} \phi = K \cap \Psi_{\neg\phi}$ (where $\Psi_{\neg\phi}$ is given by (A3)) it is sufficient to show that $\phi \notin \Psi_{\neg\phi}$, that is, $f(s, \|\neg\phi\|) \nsubseteq \|\phi\|$, for some $s \in \mathscr{B}(s_@)$. This follows from the fact that, by 4(a) of Definition 2, for every $s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \|\neg\phi\|$.

(K−5) We need to show that if $\phi \in K$ then $K \subseteq (K \dot{-}' \phi) + \phi = Cn(K \dot{-}' \phi \cup \{\phi\})$. Assume that $\phi \in K$ and fix an arbitrary $\psi \in K$. Then $(\phi \to \psi) \in K$. If $\|\neg\phi\| = \varnothing$ then $K \dot{-}' \phi = K \cap Cn(\neg\phi)$. Since $\neg\phi \in Cn(\neg\phi)$, $\phi \to \psi \in Cn(\neg\phi)$ and thus $\phi \to \psi \in K \dot{-}' \phi$, from which it follows (since, by $(K-1)$, $K \dot{-}' \phi$ is deductively closed) that $\psi \in Cn(K \dot{-}' \phi \cup \{\phi\})$. Suppose now that $\|\neg\phi\| \neq \varnothing$ so that $K \dot{-}' \phi = K \dot{-} \phi = K \cap \Psi_{\neg\phi}$ (where $\Psi_{\neg\phi}$ is given by (A3)). By 4(a) of Definition 2, for every $s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \|\neg\phi\|$ and thus $f(s, \|\neg\phi\|) \subseteq \|\phi \to \psi\| = \|\neg\phi\| \cup \|\psi\|$. Hence (recall that $(\phi \to \psi) \in K$) $(\phi \to \psi) \in K \dot{-} \phi$ so that $\psi \in Cn(K \dot{-} \phi \cup \{\phi\})$.

(K−6) We need to show that if $\phi \leftrightarrow \psi$ is a tautology then $K \dot{-}' \phi = K \dot{-}' \psi$. Assume that $\phi \leftrightarrow \psi$ is a tautology. Then $Cn(\neg\phi) = Cn(\neg\psi)$ and $\|\neg\phi\| = \|\neg\psi\|$. Thus $\|\neg\phi\| = \varnothing$ if and only if $\|\neg\psi\| = \varnothing$, in which case $K \dot{-}' \phi = K \cap Cn(\neg\phi) = K \cap Cn(\neg\psi) = K \dot{-}' \psi$. Furthermore, $\|\neg\phi\| \neq \varnothing$ if and only if $\|\neg\psi\| \neq \varnothing$, in which case $\{\chi \in \Phi_0 : f(s, \|\neg\phi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\} = \{\chi \in \Phi_0 : f(s, \|\neg\psi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$, from which it follows that $K \dot{-} \phi = K \dot{-} \psi$.

(K−7) We have to show that $(K \dot{-}' \phi) \cap (K \dot{-}' \psi) \subseteq K \dot{-}' (\phi \wedge \psi)$. We need to consider several cases.

Case 1: $\|\neg\phi\| = \|\neg\psi\| = \varnothing$ so that $\|\neg\phi\| \cup \|\neg\psi\| = \|\neg\phi \vee \neg\psi\| = \|\neg(\phi \wedge \psi)\| = \varnothing$. In this case $K \dot{-}' \phi = K \cap Cn(\neg\phi)$, $K \dot{-}' \psi = K \cap Cn(\neg\psi)$ and $K \dot{-}' (\phi \wedge \psi) = K \cap Cn(\neg(\phi \wedge \psi))$. Since $Cn(\neg\phi) \cap Cn(\neg\psi) \subseteq Cn(\neg\phi \vee \neg\psi) = Cn(\neg(\phi \wedge \psi))$ it follows that $(K \dot{-}' \phi) \cap (K \dot{-}' \psi) \subseteq K \dot{-}' (\phi \wedge \psi)$.

Case 2: $\|\neg\phi\| = \varnothing$ and $\|\neg\psi\| \neq \varnothing$, so that $\|\neg(\phi \wedge \psi)\| = \|\neg\phi \vee \neg\psi\| = \|\neg\phi\| \cup \|\neg\psi\| = \|\neg\psi\| \neq \varnothing$. In this case $K \dot{-}' \phi = K \cap Cn(\neg\phi)$, $K \dot{-}' \psi = K \dot{-} \psi = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\psi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$ and $K \dot{-}' (\phi \wedge \psi) = K \dot{-} (\phi \wedge \psi) = K \cap \{\chi \in \Phi_0 : f(s, \|\neg(\phi \wedge \psi)\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$. Since $\|\neg(\phi \wedge \psi)\| = \|\neg\psi\|$, $f(s, \|\neg(\phi \wedge \psi)\|) = f(s, \|\neg\psi\|)$ and thus $K \dot{-} (\phi \wedge \psi) = K \dot{-} \psi$. Hence the inclusion $(K \dot{-}' \phi) \cap (K \dot{-} \psi) \subseteq K \dot{-} (\phi \wedge \psi)$ reduces to $(K \dot{-}' \phi) \cap (K \dot{-} \psi) \subseteq K \dot{-} \psi$, which is trivially true.

Case 3: $\|\neg\phi\| \neq \varnothing$ and $\|\neg\psi\| = \varnothing$, so that $\|\neg\phi \vee \neg\psi\| = \|\neg\phi\| \cup \|\neg\psi\| = \|\neg\phi\| \neq \varnothing$. In this case, by an argument similar to the one used in Case 2, $K \dot{-}' (\phi \wedge \psi) = K \dot{-} (\phi \wedge \psi) = K \dot{-} \phi = K \dot{-}' \phi$, so that the inclusion $(K \dot{-}' \phi) \cap (K \dot{-}' \psi) \subseteq K \dot{-}' (\phi \wedge \psi)$ reduces to $(K \dot{-} \phi) \cap (K \dot{-}' \psi) \subseteq K \dot{-} \phi$, which is trivially true.

Case 4: $\|\neg\phi\| \neq \varnothing$ and $\|\neg\psi\| \neq \varnothing$, so that $\|\neg(\phi \wedge \psi)\| = \|\neg\phi \vee \neg\psi\| = \|\neg\phi\| \cup \|\neg\psi\| \neq \varnothing$. In this case $K \dot{-}' \phi = K \dot{-} \phi = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\phi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$, $K \dot{-}' \psi = K \dot{-} \psi = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\psi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$ and $K \dot{-}' (\phi \wedge \psi) = K \dot{-} (\phi \wedge \psi) = K \cap \{\chi \in \Phi_0 : f(s, \|\neg(\phi \wedge \psi)\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$. Fix an arbitrary $\chi \in (K \dot{-} \phi) \cap (K \dot{-} \psi)$ (thus, in particular, $\chi \in K$). We need to show that $\chi \in K \dot{-} (\phi \wedge \psi)$, that is, that, $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\neg(\phi \wedge \psi)\|) \subseteq \|\chi\|$. Since $\chi \in (K \dot{-} \phi) \cap (K \dot{-} \psi)$,

$$f(s, \|\neg\phi\|) \subseteq \|\chi\| \text{ and } f(s, \|\neg\psi\|) \subseteq \|\chi\|. \tag{A4}$$

By Property 4(a) of Definition 2, $f(s, \|\neg(\phi \wedge \psi)\|) \subseteq \|\neg(\phi \wedge \psi)\| = \|\neg\phi\| \cup \|\neg\psi\|$. It follows from this that

$$f(s, \|\neg(\phi \wedge \psi)\|) = \left(f(s, \|\neg(\phi \wedge \psi)\|) \cap \|\neg\phi\|\right) \bigcup \left(f(s, \|\neg(\phi \wedge \psi)\|) \cap \|\neg\psi\|\right). \tag{A5}$$

By Property 4(d) of Definition 2 (with $E = \|\neg\phi\| \cup \|\neg\psi\| = \|\neg(\phi \wedge \psi)\|$ and $F = \|\neg\phi\|$)

$$f(s, \|\neg(\phi \wedge \psi)\|) \cap \|\neg\phi\| \subseteq f(s, \|\neg\phi\|). \tag{A6}$$

A second application of Property 4(d) of Definition 2 (with $E = \|\neg\phi\| \cup \|\neg\psi\| = \|\neg(\phi \wedge \psi)\|$ and, this time, with $F = \|\neg\psi\|$) gives

$$f(s, \|\neg(\phi \wedge \psi)\|) \cap \|\neg\psi\| \subseteq f(s, \|\neg\psi\|). \tag{A7}$$

It follows from (A5), (A6), (A7) that $f(s, \|\neg(\phi \wedge \psi)\| \subseteq (f(s, \|\neg\phi\|) \cup f(s, \|\neg\psi\|))$ and thus, by (A4), $f(s, \|\neg(\phi \wedge \psi)\|) \subseteq \|\chi\|$.

(K−8) We need to show that if $\phi \notin K \div' (\phi \wedge \psi)$ then $K \div' (\phi \wedge \psi) \subseteq K \div' \phi$. Assume that $\phi \notin K \div' (\phi \wedge \psi)$.
Suppose first that $\|\neg\phi\| = \varnothing$, that is, $\|\phi\| = S$. Then $\mathscr{B}(s_@) \subseteq \|\phi\|$ and thus $\phi \in K$. If $\|\neg(\phi \wedge \psi)\| = \|\neg\phi\| \cup \|\neg\psi\| \neq \varnothing$ then $K \div' (\phi \wedge \psi) = K \div (\phi \wedge \psi) = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\phi\| \cup \|\neg\psi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$ and, since $\|\phi\| = S$, for all $s \in \mathscr{B}(s_@)$ we have that $f(s, \|\neg\phi\| \cup \|\neg\psi\|) \subseteq \|\phi\|$, implying that $\phi \in K \div (\phi \wedge \psi)$, contradicting our assumption. Thus the case where $\|\neg\phi\| = \varnothing$ and $\|\neg\phi\| \cup \|\neg\psi\| \neq \varnothing$ is ruled out and we are left with only two cases to consider.

Case 1: $\|\neg\phi\| \cup \|\neg\psi\| = \varnothing$ so that $\|\neg\phi\| = \varnothing$. In this case $K \div' (\phi \wedge \psi) = K \cap Cn(\neg(\phi \wedge \psi))$ and $K \div' \phi = K \cap Cn(\neg\phi)$. Fix an arbitrary $\chi \in K \div' (\phi \wedge \psi)$. Then $\chi \in K$ and $\chi \in Cn(\neg(\phi \wedge \psi))$. We need to show that $\chi \in K \div' \phi$, that is, that $\chi \in Cn(\neg\phi)$. Since $\chi \in Cn(\neg(\phi \wedge \psi))$, $\neg(\phi \wedge \psi) \to \chi$ is a tautology. Thus, since $\neg\phi \to \neg(\phi \wedge \psi)$ is also a tautology, $\neg\phi \to \chi$ is a tautology and thus $\chi \in Cn(\neg\phi)$.

Case 2: $\|\neg\phi\| \neq \varnothing$ and thus $\|\neg(\phi \wedge \psi)\| = \|\neg\phi\| \cup \|\neg\psi\| \neq \varnothing$. Then $K \div' (\phi \wedge \psi) = K \div (\phi \wedge \psi) = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\phi\| \cup \|\neg\psi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$ and $K \div' \phi = K \div \phi = K \cap \{\chi \in \Phi_0 : f(s, \|\neg\phi\|) \subseteq \|\chi\|, \forall s \in \mathscr{B}(s_@)\}$. Recall the assumption that $\phi \notin K \div (\phi \wedge \psi)$. Then two sub-cases are possible.
Case 2.1: $\phi \notin K$, that is, $\mathscr{B}(s_@) \cap \|\neg\phi\| \neq \varnothing$. Then, by 4(c) of Definition 2,

$$\forall s \in \mathscr{B}(s_@), f(s, \|\neg\phi\|) \subseteq \mathscr{B}(s_@) \cap \|\neg\phi\| \subseteq \mathscr{B}(s_@). \tag{A8}$$

Fix an arbitrary $\chi \in K \div (\phi \wedge \psi)$. Then $\chi \in K$, that is, $\mathscr{B}(s_@) \subseteq \|\chi\|$ and thus, by (A8), $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \|\chi\|$ so that $\chi \in K \div \phi$.
Case 2.2: $\phi \in K$ and $B_{\neg\phi\neg\psi} \neq \varnothing$, where $B_{\neg\phi\neg\psi} = \{s \in \mathscr{B}(s_@) : f(s, \|\neg\phi\| \cup \|\neg\psi\|) \cap \|\neg\phi\| \neq \varnothing\}$.[5] Then, by 4(e.1) of Definition 2 (with $E = \|\neg\phi\| \cup \|\neg\psi\|$ and $F = \|\neg\phi\|$)

$$\forall s \in \mathscr{B}_{\neg\phi\neg\psi}, f(s, \|\neg\phi\|) \subseteq f(s, \|\neg\phi\| \cup \|\neg\psi\|) \cap \|\neg\phi\| \tag{A9}$$

and, by 4(e.2) of Definition 2 (again, with $E = \|\neg\phi\| \cup \|\neg\psi\|$ and $F = \|\neg\phi\|$),

$$\forall s \in \mathscr{B}(s_@), f(s, \|\neg\phi\|) \subseteq f(s', \|\neg\phi\|) \text{ for some } s' \in B_{\neg\phi\neg\psi}. \tag{A10}$$

Fix an arbitrary $\chi \in K \div (\phi \wedge \psi)$. Then, $\chi \in K$ and (recall that $\|\neg(\phi \wedge \psi)\| = \|\neg\phi\| \cup \|\neg\psi\|$) $f(s, \|\neg\phi\| \cup \|\neg\psi\|) \subseteq \|\chi\|$, $\forall s \in \mathscr{B}(s_@)$; it follows from this, (A9) and (A10) that, $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\neg\phi\|) \subseteq \|\chi\|$. Thus $\chi \in K \div \phi$.

Before we proceed to the proof of Part (B) of Proposition 1, we establish the following lemma.

[5] Note that the case where $\phi \in K$ and $B_{\neg\phi\neg\psi} = \varnothing$ is ruled out by our initial assumption that $\phi \notin K \div (\phi \wedge \psi)$. In fact, $B_{\neg\phi\neg\psi} = \varnothing$ means that, $\forall s \in \mathscr{B}(s_@), f(s, \|\neg\phi\| \cup \|\neg\psi\|) \cap \|\neg\phi\| = \varnothing$, that is, $f(s, \|\neg\phi\| \cup \|\neg\psi\|) \subseteq \|\phi\|$, which, in conjunction with the hypothesis that $\phi \in K$, yields $\phi \in K \div (\phi \wedge \psi)$.

Lemma 2. *Let $A \subseteq \Phi_0$ be such that $A = Cn(A)$. Then, $\forall \alpha \in \Phi_0$, $\|Cn(A \cup \{\alpha\})\| = \|A\| \cap \|\alpha\|$.*

Proof. Since A is deductively closed, $\forall \beta \in \Phi_0$,

$$\beta \in Cn(A \cup \{\alpha\}) \text{ if and only if } (\alpha \to \beta) \in A. \tag{A12}$$

First we show that $\|A\| \cap \|\alpha\| \subseteq \|Cn(A \cup \{\alpha\})\|$. Fix an arbitrary $s \in \|A\| \cap \|\alpha\|$; we need to show that $s \in \|Cn(A \cup \{\alpha\})\|$, that is, that $\forall \beta \in Cn(A \cup \{\alpha\}), \beta \in s$. Since $s \in \|\alpha\|$, $\alpha \in s$. Fix an arbitrary $\beta \in Cn(A \cup \{\alpha\})$; then, by (A12), $(\alpha \to \beta) \in A$; thus, since $s \in \|A\|$, $(\alpha \to \beta) \in s$. Hence, since both α and $\alpha \to \beta$ belong to s and s is deductively closed, $\beta \in s$.
Next we show that $\|Cn(A \cup \{\alpha\})\| \subseteq \|A\| \cap \|\alpha\|$. Let $s \in \|Cn(A \cup \{\alpha\})\|$. Then, since $\alpha \in Cn(A \cup \{\alpha\})$, $\alpha \in s$, that is, $s \in \|\alpha\|$. It remains to show that $s \in \|A\|$, that is, that, for every $\beta \in A$, $\beta \in s$. Fix an arbitrary $\beta \in A$; then, since A is deductively closed, $(\alpha \to \beta) \in A$. Thus, by (A12), $\beta \in Cn(A \cup \{\alpha\})$ and thus, since $s \in \|Cn(A \cup \{\alpha\})\|$, $\beta \in s$. □

Proof of Part (B) of Proposition 1.
We need to show that if $K \subset \Phi_0$ is consistent and deductively closed and $\div : \Phi_0 \to 2^{\Phi_0}$ is an AGM belief contraction function based on K, then there exists a model $\langle S, s_@, \mathscr{B}, f, V \rangle$ such that $K = \{\phi \in \Phi_0 : \mathscr{B}(s_@) \subseteq \|\phi\|\}$ and, for all $\phi, \psi \in \Phi_0$, $\psi \in K \div \phi$ if and only if (A1) is satisfied. Define the following model $\langle S, s_@, \mathscr{B}, f, V \rangle$:

1. S is the set of maximally consistent sets of formulas in Φ_0.
2. The valuation $V : At \to S$ is defined by $V(p) = \{s \in S : p \in s\}$, so that, for every $\phi \in \Phi_0$, $\|\phi\| = \{s \in S : \phi \in s\}$. If $\Psi \subseteq \Phi_0$, define $\|\Psi\| = \{s \in S : \forall \phi \in \Psi, \phi \in s\}$.
3. Choose an arbitrary $s_@ \in S$ and define $\mathscr{B}(s_@) = \|K\|$.
4. Let $\mathscr{E} = \{E \subseteq S : \varnothing \neq E = \|\phi\| \text{ for some } \phi \in \Phi_0\}$. Define $f : \mathscr{B}(s_@) \times \mathscr{E} \to 2^S$ as follows:
$$\forall s \in \mathscr{B}(s_@),\ f(s, \|\phi\|) = \|K \div \neg\phi\| \cap \|\phi\|. \tag{A11}$$

Remark 1. *If ϕ is a tautology then $\neg\phi$ is a contradiction and thus (since, by hypothesis, K is consistent) $\neg\phi \notin K$. It follows from $(K-2)$ and $(K-3)$ that $K \div \neg\phi = K$. Furthermore, since ϕ is a tautology and K is deductively closed, $\phi \in K$, that is $\|K\| \subseteq \|\phi\|$ so that $\|K\| \cap \|\phi\| = \|K\|$. Hence, by* (A11), *$\forall s \in \mathscr{B}(s_@)$, $f(s, \|\phi\|) = \|K\|$. On the other hand, if $\neg\phi$ is a tautology then $\|\phi\| = \varnothing$ and thus $\|\phi\| \notin \mathscr{E}$, that is, $\|\phi\|$ is not in the domain of f.*

First we show that the selection function defined in (A11) satisfies Properties 4(a)-4(e) of Definition 2. In view of Remark 1, we can restrict attention to contingent formulas, that is, to formulas ϕ such that neither ϕ nor $\neg\phi$ is a tautology. Denote by $\Phi_{cont} \subseteq \Phi_0$ the set of contingent formulas.
Recall that S is the set of maximally consistent sets of formulas in Φ_0 and, for $A \subseteq \Phi_0$, $\|A\| = \{s \in S : \chi \in s, \forall \chi \in A\}$.

Property 4(a) We need to show that if $\phi \in \Phi_{cont}$ then $\|K \div \neg\phi\| \cap \|\phi\| \subseteq \|\phi\|$, which is obviously true, and $\|K \div \neg\phi\| \cap \|\phi\| \neq \varnothing$. Since $\phi \in \Phi_{cont}$, $\|\phi\| \neq \varnothing$ and, by $(K-4)$, $\neg\phi \notin K \div \neg\phi$. By $(K-1)$ $K \div \neg\phi = Cn(K \div \neg\phi)$ and thus $\neg\phi \notin Cn(K \div \neg\phi)$, that is, $K \div \neg\phi$ is consistent and hence $\|K \div \neg\phi\| \neq \varnothing$.

Property 4(b) Fix an arbitrary $s \in \mathscr{B}(s_@)$ and an arbitrary $\phi \in \Phi_{cont}$. We need to show that if $s \in \|\phi\|$ then $s \in f(s, \|\phi\|) = \|K \div \neg\phi\| \cap \|\phi\|$. By construction, $\mathscr{B}(s_@) = \|K\|$; thus, $s \in \|K\|$. By $(K-2)$, $K \div \neg\phi \subseteq K$ so that $\|K\| \subseteq \|K \div \neg\phi\|$. Hence $s \in \|K \div \neg\phi\|$. Thus if $s \in \|\phi\|$ then $s \in \|K \div \neg\phi\| \cap \|\phi\|$.

Property 4(c) We need to show that if $\mathscr{B}(s_@) \cap \|\phi\| \neq \varnothing$ then (since $\mathscr{B}(s_@) = \|K\|$ and, $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\phi\|) = \|K \div \neg\phi\| \cap \|\phi\|)$ $\|K \div \neg\phi\| \cap \|\phi\| \subseteq \|K\| \cap \|\phi\|$. If $\|K\| \cap \|\phi\| \neq \varnothing$ then $\neg\phi \notin K$ and thus, by $(K-3)$, $K \subseteq K \div \neg\phi$, so that $\|K \div \neg\phi\| \subseteq \|K\|$ and thus $\|K \div \neg\phi\| \cap \|\phi\| \subseteq \|K\| \cap \|\phi\|$.

Property 4(d) We need to show that if $\phi \in \Phi_{cont}$ and $\psi \in \Phi_0$, then $\forall s \in \mathscr{B}(s_@)$, $f(s, \|\phi\|) \cap \|\psi\| \subseteq f(s, \|\phi\| \cap \|\psi\|)$, that is, using (A11) and the fact that $\|\phi\| \cap \|\psi\| = \|\phi \wedge \psi\|$,

$$\|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\| \subseteq \|K \div \neg(\phi \wedge \psi)\| \cap \|\phi \wedge \psi\| \tag{A13}$$

By $(K-7)$, $\forall \alpha, \beta \in \Phi_0$, $(K \div \alpha) \cap (K \div \beta) \subseteq K \div (\alpha \wedge \beta)$. Thus applying $(K-7)$ to $\alpha = \neg(\phi \wedge \psi)$ and $\beta = \phi \to \psi$ we get

$$K \div \neg(\phi \wedge \psi) \cap K \div (\phi \to \psi) \subseteq K \div (\neg(\phi \wedge \psi) \wedge (\phi \to \psi)) \tag{A14}$$

Since $\neg(\phi \wedge \psi) \wedge (\phi \to \psi)$ is logically equivalent to $\neg\phi$, by $(K-6)$ $K \div (\neg(\phi \wedge \psi) \wedge (\phi \to \psi)) = K \div \neg\phi$. Thus, by (A14)

$$K \div \neg(\phi \wedge \psi) \cap K \div (\phi \to \psi) \subseteq K \div \neg\phi. \tag{A15}$$

Next we show that

$$Cn(K \div \neg(\phi \wedge \psi) \cup \{\phi \wedge \psi\}) \subseteq Cn(K \div \neg\phi \cup \{\phi \wedge \psi\}). \tag{A16}$$

Fix an arbitrary $\chi \in Cn(K \div \neg(\phi \wedge \psi) \cup \{\phi \wedge \psi\})$. Then, since, by $(K-1)$, $K \div \neg(\phi \wedge \psi)$ is deductively closed,

$$((\phi \wedge \psi) \to \chi) \in K \div \neg(\phi \wedge \psi). \tag{A17}$$

By $(K-2)$, $K \div \neg(\phi \wedge \psi) \subseteq K$ and thus, by (A17),

$$((\phi \wedge \psi) \to \chi) \in K. \tag{A18}$$

Next we show that

$$((\phi \wedge \psi) \to \chi) \in K \div (\phi \to \psi). \tag{A19}$$

If $(\phi \to \psi) \notin K$ then, by $(K-3)$, $K \subseteq K \div (\phi \to \psi)$ and thus (A19) follows from (A18). If $(\phi \to \psi) \in K$ then, by $(K-5)$, $K \subseteq Cn(K \div (\phi \to \psi) \cup \{\phi \to \psi\})$ so that, by (A18), $((\phi \wedge \psi) \to \chi) \in Cn(K \div (\phi \to \psi) \cup \{\phi \to \psi\})$, that is (since, by $(K-1)$, $K \div (\phi \to \psi)$ s deductively closed) $(\phi \to \psi) \to ((\phi \wedge \psi) \to \chi) \in K \div (\phi \to \psi)$. Since $(\phi \to \psi) \to ((\phi \wedge \psi) \to \chi)$ is logically equivalent to $((\phi \to \psi) \wedge (\phi \wedge \psi)) \to \chi$, which, in turn is logically equivalent to $(\phi \wedge \psi)) \to \chi$, (A19) is satisfied. It follows from (A18), (A19) and (A15) that $\big((\phi \wedge \psi) \to \chi\big) \in K \div \neg\phi$, that is, that $\chi \in Cn\big(K \div \neg\phi \cup \{\phi \wedge \psi\}\big)$, thus establishing (A16). From (A16) we get that

$$\|Cn(K \div \neg\phi \cup \{\phi \wedge \psi\})\| \subseteq \|Cn(K \div \neg(\phi \wedge \psi) \cup \{\phi \wedge \psi\})\| \tag{A20}$$

By Lemma 2 (with $A = K \div \neg\phi$ and $\alpha = \phi \wedge \psi$), $\|Cn(K \div \neg\phi \cup \{\phi \wedge \psi\})\| = \|K \div \neg\phi\| \cap \|\phi \wedge \psi\|$ which in turn (since $=\|\phi \wedge \psi\| = \|\phi\| \cap \|\psi\|$) is equal to $\|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\|$. By Lemma 2 again (with $A = K \div \neg(\phi \wedge \psi)$ and $\alpha = \phi \wedge \psi$), $\|Cn(K \div \neg(\phi \wedge \psi) \cup \{\phi \wedge \psi\})\| = \|K \div \neg(\phi \wedge \psi)\| \cap \|\psi \wedge \psi\|$. Hence (A13) follows from (A20).

Property 4(e) Since, by (A11), $\forall s, s' \in \mathscr{B}(s_@)$, $f(s, \|\phi\|) = f(s', \|\phi\|) = \|K \div \neg\phi\| \cap \|\phi\|$, it is sufficient to show that if $\|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\| \neq \varnothing$ then $\|K \div \neg(\phi \wedge \psi)\| \cap \|\phi \wedge \psi\| \subseteq \|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\|$. Assume that $\|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\| = \|K \div \neg\phi\| \cap \|\phi \wedge \psi\| \neq \varnothing$. Then

$$\neg(\phi \wedge \psi) \notin K \div \neg\phi. \tag{A21}$$

Since $\neg\phi$ is logically equivalent to $\neg(\phi \wedge \psi) \wedge \neg\phi$, by $(K-6)$

$$K \div \neg\phi = K \div (\neg(\phi \wedge \psi) \wedge \neg\phi). \tag{A22}$$

Thus, by (A21) and (A22),

$$\neg(\phi \wedge \psi) \notin K \div (\neg(\phi \wedge \psi) \wedge \neg\phi). \tag{A23}$$

By $(K-8)$, $\forall \alpha, \beta \in \Phi_0$, if $\alpha \notin K \div (\alpha \wedge \beta)$ then $K \div (\alpha \wedge \beta) \subseteq K \div \alpha$. Thus, by (A23) and $(K-8)$ (with $\alpha = \neg(\phi \wedge \psi)$ and $\beta = \neg\phi$), $K \div (\neg\phi \wedge \neg(\phi \wedge \psi)) \subseteq K \div \neg(\phi \wedge \psi)$. It follows from this and (A22) that $K \div \neg\phi \subseteq K \div \neg(\phi \wedge \psi)$ and thus

$$\|K \div \neg(\phi \wedge \psi)\| \subseteq \|K \div \neg\phi\|. \tag{A24}$$

Intersecting both sides of (A24) with $\|\phi \wedge \psi\| = \|\phi\| \cap \|\psi\|$ we get $\|K \div \neg(\phi \wedge \psi)\| \cap \|\phi \wedge \psi\| \subseteq \|K \div \neg\phi\| \cap \|\phi\| \cap \|\psi\|$, as desired.

To complete the proof of Part (B) of Proposition 1 we need to show that

$$\psi \in K \div \phi \text{ if and only if } \begin{array}{l} (1)\, \mathscr{B}(s_@) \subseteq \|\psi\|, \text{ and} \\ (2)\, \forall s \in \mathscr{B}(s_@), f(s, \|\neg\phi\|) \subseteq \|\psi\|. \end{array}$$

By (A11), $\forall s \in \mathscr{B}(s_@) = \|K\|$, $f(s, \|\neg\phi\|) = \|K \div \phi\| \cap \|\neg\phi\|$. Thus we have to show that

$$\psi \in K \div \phi \text{ if and only if } \|K\| \subseteq \|\psi\| \text{ and } \|K \div \phi\| \cap \|\neg\phi\| \subseteq \|\psi\|. \tag{A25}$$

First we establish a lemma.

Lemma 3. *$\forall \phi \in \Phi_0$,*

(i) if $A \subseteq \Phi_0$ is such that $A = Cn(A)$, then $A = Cn(A \cup \{\phi\}) \cap Cn(A \cup \{\neg\phi\})$

(ii) $K \div \phi = K \cap Cn(K \div \phi \cup \{\neg\phi\})$

Proof. (*i*) Let $A \subseteq \Phi_0$ be such that $A = Cn(A)$. Since $A \subseteq Cn(A \cup \{\phi\})$ and $A \subseteq Cn(A \cup \{\neg\phi\})$, $A \subseteq Cn(A \cup \{\phi\}) \cap Cn(A \cup \{\neg\phi\})$. Conversely, suppose that $\chi \in Cn(A \cup \{\phi\}) \cap Cn(A \cup \{\neg\phi\})$. Then both $\phi \to \chi$ and $\neg\phi \to \chi$ belong to A and thus so does their conjunction. Since $(\phi \to \chi) \wedge (\neg\phi \to \chi)$ is logically equivalent to χ it follows that $\chi \in A$.
(*ii*) We need to consider two cases.
Case 1: $\phi \in K$. Then, by $(K-5)$, $K \subseteq Cn(K \div \phi \cup \{\phi\})$. By $(K-2)$, $K \div \phi \subseteq K$, so that $Cn\big(K \div \phi \cup \{\phi\}\big) \subseteq Cn\big(K \cup \{\phi\}\big) = Cn(K) = K$ (by hypothesis, K is deductively closed). Thus

$$K = Cn(K \div \phi \cup \{\phi\}) \tag{A26}$$

By Part (*i*) (with $A = K \div \phi$, which, by $(K-1)$, is deductively closed),

$$K \div \phi = Cn(K \div \phi \cup \{\phi\}) \cap Cn(K \div \phi \cup \{\neg\phi\}) \tag{A27}$$

Thus, by (A26) and (A27), $K \div \phi = K \cap Cn(K \div \phi \cup \{\neg\phi\})$.
Case 2: $\phi \notin K$. Then, by $(K-2)$ and $(K-3)$,

$$K \div \phi = K \tag{A28}$$

By Part (i) (with $A = K$)

$$K = Cn(K \cup \{\phi\}) \cap Cn(K \cup \{\neg\phi\}) \tag{A29}$$

From (A29) we get that $K \cap Cn(K \cup \{\neg\phi\}) = Cn(K \cup \{\phi\}) \cap Cn(K \cup \{\neg\phi\}) = K$. Thus, by (A28), $K \div \phi = K \cap Cn(K \cup \{\neg\phi\})$, from which, by using (A28) again to replace the second instance of K with $K \div \phi$, we get $K \div \phi = K \cap Cn(K \div \phi \cup \{\neg\phi\})$ □

Now we are ready to prove (A25), namely that

$$\psi \in K \div \phi \text{ if and only if } \|K\| \subseteq \|\psi\|, \text{ and } \|Cn(K \div \phi \cup \{\neg\phi\})\| \subseteq \|\psi\|.$$

Let $\psi \in K \div \phi$. By (ii) of Lemma 3, $K \div \phi = K \cap Cn(K \div \phi \cup \{\neg\phi\})$; thus $\psi \in K$, that is, $\|K\| \subseteq \|\psi\|$, and $\psi \in Cn(K \div \phi \cup \{\neg\phi\})$, that is, $\|Cn(K \cup \{\neg\phi\})\| \subseteq \|\psi\|$. Conversely, suppose that $\|K\| \subseteq \|\psi\|$ and $\|Cn(K \div \phi \cup \{\neg\phi\})\| \subseteq \|\psi\|$, that is, $\psi \in K \cap Cn(K \div \phi \cup \{\neg\phi\})$. Then, by ($ii$) of Lemma 3, $\psi \in K \div \phi$. □

References

[1] Carlos Alchourrón, Peter Gärdenfors & David Makinson (1985): *On the logic of theory change: partial meet contraction and revision functions*. The Journal of Symbolic Logic 50, pp. 510–530, doi:10.2307/2274239.

[2] Giacomo Bonanno (2023): *A Kripke-Stalnaker-Lewis semantics for AGM belief revision*. Technical Report, REPEC preprint No. 354. Available at https://econpapers.repec.org/paper/cdawpaper/354.htm.

[3] Samir Chopra, Aditya Ghose, Thomas Meyer & Ka-Shu Wong (2008): *Iterated Belief Change and the Recovery Axiom*. Journal of Philosophical Logic 37(5), pp. 501–520, doi:10.1007/s10992-008-9086-2.

[4] Eduardo Fermé (1998): *On the Logic of Theory Change: Contraction without Recovery*. Journal of Logic, Language, and Information 7(2), pp. 127–137, doi:10.1023/A:1008241816078.

[5] Eduardo Fermé & Sven Ove Hansson (2011): *AGM 25 Years*. Journal of Philosophical Logic 40, pp. 295–331, doi:10.1007/S10992-011-9171-9.

[6] Eduardo Fermé & Sven Ove Hansson (2018): *Belief change: introduction and overview*. Springer, doi:10.1007/978-3-319-60535-7.

[7] Eduardo Fermé & Ricardo Rodriguez (1998): *Semi-Contraction: Axioms and Construction*. Notre Dame Journal of Formal Logic 39(3), pp. 332–345, doi:10.1305/ndjfl/1039182250.

[8] Nir Friedman & Joseph Halpern (1994): *Conditional logics of belief change*. In Barbara Hayes-Roth & Richard Korf, editors: *AAAI'94: Proceedings*, AAAI Press, pp. 915–921. Available at https://dl.acm.org/doi/proceedings/10.5555/2891730.

[9] André Fuhrmann (1991): *Theory contraction through base contraction*. Journal of Philosophical Logic 20, pp. 175–203, doi:10.1007/BF00284974.

[10] Peter Gärdenfors (1986): *Belief Revisions and the Ramsey Test for Conditionals*. Philosophical Review 95(1), pp. 81–93, doi:10.2307/2185133.

[11] Peter Gärdenfors (1988): *Knowledge in flux: modeling the dynamics of epistemic states*. MIT Press. Available at https://www.collegepublications.co.uk/logic/lcs/?00004.

[12] Konstantinos Georgatos (2017): *Epistemic Conditionals and the Logic of Subsets*. In Ramaswamy Ramanujam, Lawrence Moss & Can Başkent, editors: Rohit Parikh on Logic, Language and Society, Springer Verlag, pp. 259–277, doi:10.1007/978-3-319-47843-2.

[13] Adam Grove (1988): *Two modellings for theory change*. Journal of Philosophical Logic 17, pp. 157–170, doi:10.1007/BF00247909.

[14] Sven Ove Hansson (1991): *Belief contraction without recovery*. Studia Logica 50(2), pp. 251–260, doi:10.1007/BF00370186.

[15] Sven Ove Hansson (1996): *Hidden structures of belief*. In Andre Fuhrmann & Hans Rott, editors: Logic, Actions and Information, de Gruyter, pp. 79–100. Available at https://www.degruyter.com/document/isbn/9783110868890/html?lang=en.

[16] Sven Ove Hansson (1999): *Recovery and epistemic residue*. Journal of Logic, Language and Information 8, pp. 421–428, doi:10.1023/A:1008316915066.

[17] Sven Ove Hansson (1999): *A textbook of belief dynamics: Theory change and database updating*. Springer Dordrecht, Dordrecht, doi:10.1007/978-94-007-0814-3.

[18] Sébastien Konieczny & Ramón Pino Pérez (2017): *On Iterated Contraction: Syntactic Characterization, Representation Theorem and Limitations of the Levi Identity*. In Serafín Moral, Olivier Pivert, Daniel Sánchez & Nicolás Marín, editors: Scalable Uncertainty Management, Springer International Publishing, pp. 348–362, doi:10.1007/978-3-319-67582-4_25.

[19] Isaac Levi (1991): *The fixation of belief and its undoing*. Cambridge University Press, doi:10.1017/CBO9780511663819.

[20] Isaac Levi (2004): *Mild Contraction*. Oxford University Press, doi:10.1093/0199270708.001.0001.

[21] David Lewis (1973): *Counterfactuals*. Harvard University Press. Available at https://www.wiley.com/en-us/Counterfactuals-p-9780631224259.

[22] Sten Lindström & Wlodek Rabinowicz (1991): *Epistemic entrenchment with incomparabilities and relational belief revision*. In André Fuhrmann & Michael Morreau, editors: The Logic of Theory Change, Springer, pp. 93–126, doi:10.1007/BFb0018418.

[23] Sten Lindström & Wlodek Rabinowicz (1998): *Conditionals and the Ramsey Test*. In Didier Dubois & Henri Prade, editors: Belief Change, Springer Netherlands, Dordrecht, pp. 147–188, doi:10.1007/978-94-011-5054-5_4.

[24] Sten Linström & Wlodzimierz Rabinowicz (1992): *The Ramsey test revisited**. Theoria 58(2-3), pp. 131–182, doi:10.1111/j.1755-2567.1992.tb01138.x.

[25] David Makinson (1987): *On the status of the postulate of recovery in the logic of theory change*. Journal of Philosophical Logic 16, pp. 383–394, doi:10.1007/BF00431184.

[26] Thomas Meyer, Johannes Heidema, Willem Labuschagne & Louise Leenen (2002): *Systematic Withdrawal*. Journal of Philosophical Logic 31(5), pp. 415–443, doi:10.1023/A:1020199115746.

[27] Lawrence S Moss & Rohit Parikh (1992): *Topological reasoning and the logic of knowledge: preliminary report*. In Yoram Moses, editor: Proceedings of the 4th Conference on Theoretical Aspects of Reasoning about Knowledge (TARK 1992), Morgan Kaufmann, pp. 95– 105.

[28] Reinhard Niederée (1991): *Multiple contraction a further case against Gärdenfors' principle of recovery*. In André Fuhrmann & Michael Morreau, editors: The Logic of Theory Change, Springer, pp. 322–334, doi:10.1007/BFb0018427.

[29] Frank P. Ramsey (1950): *General Propositions and Causality*. In R. B. Braithwaite, editor: The Foundations of Mathematics and other Logical Essays, Humanities Press, pp. 237–257, doi:10.4324/9781315887814.

[30] Hans Rott (1986): *Ifs, though, and because*. Erkenntnis 25(3), pp. 345–370, doi:10.1007/BF00175348.

[31] Hans Rott (2017): *Preservation and postulation: lessons from the new debate on the Ramsey test*. Mind 126(502), pp. 609–626, doi:10.1093/mind/fzw028.

[32] Hans Rott & Maurice Pagnucco (1999): *Severe Withdrawal (and Recovery)*. Journal of Philosophical Logic 28(5), pp. 501–547, doi:10.1023/A:1004344003217.

[33] Kai Sauerwald, Gabriele Kern-Isberner & Christoph Beierle (2020): *A conditional perspective for iterated belief contraction*. In G.D. Giacomo et al, editor: *ECAI 2020*, IOS Press, Berlin, Heidelberg, pp. 889–896. Available at 10.3233/FAIA200180.

[34] Robert Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: Studies in logical theory, Blackwell, pp. 98–112, doi:10.1007/978-94-009-9117-0_2.

Comparing the Update Expressivity of Communication Patterns and Action Models

Armando Castañeda
Instituto de Matemáticas
Universidad Nacional Autónoma de México
armando.castaneda@im.unam.mx

Hans van Ditmarsch
University of Toulouse, CNRS, IRIT
hans.van-ditmarsch@irit.fr

David A. Rosenblueth
Instituto de Inv. en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
drosenbl@unam.mx

Diego A. Velázquez
Posgr. en Ciencia e Ingeniería de la Computación
Universidad Nacional Autónoma de México
velazquez-diego@ciencias.unam.mx

Any kind of dynamics in dynamic epistemic logic can be represented as an action model. Right? Wrong! In this contribution we prove that the update expressivity of communication patterns is incomparable to that of action models. Action models, as update mechanisms, were proposed by Baltag, Moss, and Solecki in 1998 and have remained the nearly universally accepted update mechanism in dynamic epistemic logics since then. Alternatives, such as arrow updates that were proposed by Kooi and Renne in 2011, have update equivalent action models. More recently, the picture is shifting. Communication patterns are update mechanisms originally proposed in some form or other by Ågotnes and Wang in 2017 (as resolving distributed knowledge), by Baltag and Smets in 2020 (as reading events), and by Velázquez, Castañeda, and Rosenblueth in 2021 (as communication patterns). All these logics have the same expressivity as the base logic of distributed knowledge. However, their update expressivity, the relation between pointed epistemic models induced by such an update, was conjectured to be different from that of action model logic. Indeed, we show that action model logic and communication pattern logic are incomparable in update expressivity. We also show that, given a history-based semantics and when restricted to (static) interpreted systems, action model logic is (strictly) more update expressive than communication pattern logic. Our results are relevant for distributed computing wherein oblivious models involve arbitrary iteration of communication patterns.

1 Introduction

It is well known that the expressivity of public announcement logic is the same as that of epistemic logic [15]. This is proved by way of a reduction system showing that every public announcement formula is equivalent to one without public announcement modalities. Similarly, the expressivity of the logic of distributed knowledge with public announcements is the same as that of the logic of distributed knowledge [1]. Again, this is shown by a reduction. A reduction also exists for the logic of distributed knowledge with action models [2]; see [18, Fig. 5 and Th. 15] and the reduction axiom called AD in [18, Fig. 9].

Distributed knowledge can also be extended with dynamic modalities for communication patterns, an update mechanism proposed in [17]. The resulting communication pattern logic is as expressive as the logic of distributed knowledge: we can reduce formulas with dynamic modalities to formulas without [6]. This logic is a slight generalization of logics with similar modalities also showing this by reduction [1, 3]. A detailed comparison to these other proposals is found in [6].

We conclude that the logic of communication patterns and distributed knowledge has the same expressivity as the logic of action models and distributed knowledge, because they both reduce to the logic of distributed knowledge. A different matter, however, is so-called *update expressivity* [10, 12, 7].

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 157–172, doi:10.4204/EPTCS.379.14

We will compare the update expressivity of communication pattern logic and action model logic. Communication patterns, like action models, are (induce) *updates* transforming pointed epistemic models into other pointed epistemic models. Is there for each communication pattern an action model defining the same update, and vice versa? Communication patterns can always be executed, but action models cannot always be executed, for example a truthful public announcement of p requires p to be true in some world. We can therefore expect a trivial difference in update expressivity. It becomes non-trivial if we also consider union of relations, such as non-deterministic choice between the announcement of p and the announcement of $\neg p$.

This is an overview of the structure of our contribution. Sect. 2 recalls communication pattern logic, action model logic, and update expressivity. In Sect. 3 we show that for each communication pattern there is an update equivalent action model when executed on epistemic models that are interpreted systems. However, the resulting model may not be an interpreted system. In Sect. 4 we then show that communication pattern logic and action model logic are indeed incomparable in update expressivity on the class of epistemic models. Finally, in Sect. 5 we propose a history-based semantics for communication pattern logic for which the class of interpreted systems is, after all, closed under updates, and we show that for each iterated communication pattern there is then an update equivalent action model.

2 Communication pattern logic and action model logic

2.1 Language

Given are a finite set of *agents* A and a set of *propositional variables* $P \subseteq P' \times A$, where P' is a countable set. For $B \subseteq A$ and $Q \subseteq P$, $Q \cap (P' \times B)$ is denoted Q_B (where Q_a is $Q_{\{a\}}$), and $(p,a) \in P$ is denoted p_a. The set P_a consists of the *local variables* of agent a. In this work we consider the following languages.

Definition 1 (Language) *Given A and P, the language $\mathscr{L}^{\times\circ}$ is defined by BNF (where $p_a \in P$, $B \subseteq A$):*

$$\varphi := p_a \mid \neg\varphi \mid \varphi \wedge \varphi \mid D_B\varphi \mid [\boldsymbol{R},R]\varphi \mid [\boldsymbol{U},e]\varphi$$

where $(\boldsymbol{R},R)$ and $(\boldsymbol{U},e)$ are structures defined below, with $R \in \boldsymbol{R}$ and e in the domain of $\boldsymbol{U}$. Furthermore, $\mathscr{L}^{\circ}$ is the language without $[\boldsymbol{U},e]\varphi$, $\mathscr{L}^{\times}$ without $[\boldsymbol{R},R]\varphi$, and $\mathscr{L}^{-}$ without either.

Epistemic formula $D_B\varphi$ is read as 'the agents in B have distributed knowledge of φ'. We write $K_a\varphi$ for $D_{\{a\}}\varphi$, for 'agent a knows φ'. *Dynamic* formula $[\boldsymbol{R},R]\varphi$ means 'after execution of communication graph R from communication pattern $\boldsymbol{R}$, φ is true', and $[\boldsymbol{U},e]\varphi$ means 'after execution of action e from action model $\boldsymbol{U}$, φ is true'. Dynamic modalities will be interpreted as updates of epistemic models.

By notational abbreviation we define $[\boldsymbol{U}]\varphi := \bigwedge_{e \in E}[\boldsymbol{U},e]\varphi$ and $[\boldsymbol{R}]\varphi := \bigwedge_{R \in \boldsymbol{R}}[\boldsymbol{R},R]\varphi$. The *modal depth* of a formula $\varphi \in \mathscr{L}^{\circ\times}$ is inductively defined as: $md(p_a) = 0$, $md(\neg\varphi) = md(\varphi)$, $md(\varphi \wedge \psi) = \max\{md(\varphi), md(\psi)\}$, $md(D_B\varphi) := md(\varphi) + 1$, $md([\boldsymbol{R},R]\varphi) := md(\varphi)$, $md([\boldsymbol{U},e]\varphi) = md(\boldsymbol{U}) + md(\varphi)$, where $md(\boldsymbol{U}) = \max\{md(\mathsf{pre}(f)) \mid f \in E\}$. In $md(\boldsymbol{U})$, the formulas $\mathsf{pre}(f)$ are defined below.

If P is finite and $Q \subseteq P$, *description* δ_Q (of valuation Q) is defined as $\bigwedge_{p_a \in Q} p_a \wedge \bigwedge_{p_a \in P \setminus Q} \neg p_a$. If P is infinite and $Q \subseteq Q' \subset P$ are finite subsets of P, description $\delta_{Q,Q'}$ is defined as $\bigwedge_{p_a \in Q} p_a \wedge \bigwedge_{p_a \in Q' \setminus Q} \neg p_a$.

2.2 Structures

Definition 2 (Epistemic model) *An* epistemic model *M is a triple $(W, \sim, L)$, where for all $a \in A$, $\sim_a$ is an* equivalence relation *on the* domain *W (also denoted $\mathscr{D}(M)$) consisting of* states *(or* worlds*), and* *where $L : W \to \mathscr{P}(P)$ is the* valuation *(function). For $\bigcap_{a \in B} \sim_a$ we write $\sim_B$, and for $\{w' \in W \mid w' \sim_a w\}$ we write $[w]_a$. We further require epistemic models to be* local*: for all $a \in A$ and $v, w \in W$, $v \sim_a w$ implies $L(v)_a = L(w)_a$; if for all a, v, w also $L(v)_a = L(w)_a$ implies $v \sim_a w$, it is a* (static) interpreted system.

An epistemic model encodes uncertainty among the agents about the value of other agents' local variables and about the knowledge of other agents.

Definition 3 (Communication pattern) *A* communication graph *R is a reflexive binary relation on the set of agents A, that is, $R \in \mathscr{P}(A \times A)$ and such that for all $a \in A$, $(a,a) \in R$. A* communication pattern *$\boldsymbol{R}$ is a set of communication graphs, that is, $\boldsymbol{R} \subseteq \mathscr{P}(A \times A)$.*

Expression $(a,b) \in R$ means that the message sent by a is received by b. For $(a,b) \in R$ we write aRb. We let $Rb := \{a \in A \mid aRb\}$, $RB := \bigcup_{b \in B} Rb$, and $R'B \equiv RB$ if $R'a = Ra$ for all $a \in B$. The *identity relation I* is $\{(a,a) \mid a \in A\}$. The *universal relation U* is $A \times A$. A communication graph is a reflexive relation, because we assume that an agent always receives her own message. But not every other agent may receive the message. We could alternatively have defined a communication pattern as a structure with equivalence relations $\sim_a$ for each agent, namely by defining that $R \sim_a R'$ iff $Ra = R'a$, as in [17].

Definition 4 (Action model) *An* action model *$\boldsymbol{U} = (E, \sim, \mathsf{pre})$ consists of a* domain *E of* actions*, an* accessibility function *$\sim : A \to \mathscr{P}(E \times E)$, where each $\sim_a$ is an equivalence relation, and a* precondition function *$\mathsf{pre} : E \to \mathscr{L}^-$.*

An action model [2] is a structure like an epistemic model but with a precondition function, associating a formula with each state. The restriction to language $\mathscr{L}^-$ for preconditions excuses us from explanations involving mutual recursion.

For all the above structures X we also consider *pointed* and *multi-pointed* versions that are pairs (X,x) with $x \in X$ (or $x \in \mathscr{D}(X)$) resp. (X,Y) with $Y \subseteq X$ ($Y \subseteq \mathscr{D}(X)$), so we have pointed epistemic models (M,w), multi-pointed action models $(\boldsymbol{U},T)$, etcetera.

Communication patterns are fairly novel in dynamic epistemic logic. We note that similar structures or modalities were proposed in [1] (resolving distributed knowledge), in [3] (reading events), and in [17] (communication patterns). The communication patterns in [17] have preconditions, just as action models. The reading events in [3] and resolution in [1] are communication patterns without uncertainty over the reception of messages. Then again, communication patterns permit less uncertainty than the arbitrary reading events in [3]. These differences are discussed in [6]. Examples are given in Sect. 3.

One can update an epistemic model with a communication pattern and one can also update an epistemic model with an action model. The updated epistemic model encodes how the knowledge has changed after agents have informed each other according to the update.

Given an epistemic model $M = (W, \sim, L)$ and a communication pattern $\boldsymbol{R}$, the *updated* epistemic model $M \odot \boldsymbol{R} = (\dot{W}, \dot{\sim}, \dot{L})$ (the *update* of M with $\boldsymbol{R}$) is defined as:

$$\begin{array}{lcl} \dot{W} & = & W \times \boldsymbol{R} \\ (w,R) \dot{\sim}_a (w',R') & \text{iff} & w \sim_{Ra} w' \text{ and } Ra = R'a \\ \dot{L}(w,R) & = & L(w) \end{array}$$

The relation $\dot{\sim}_a$ is the intersection $\sim_{Ra}$ of the relations of all agents from which a received messages.

Given an epistemic model $M = (W, \sim, L)$ and an action model $\boldsymbol{U} = (E, \sim, \mathsf{pre})$, the updated epistemic model $M \otimes \boldsymbol{U} = (W^\times, \sim^\times, L^\times)$ is defined as:

$$\begin{array}{lcl} W^\times & = & \{(v,f) \mid v \in W, f \in E, \text{ and } M, v \models \mathsf{pre}(f)\} \\ (v,f) \sim_a^\times (v',f') & \text{iff} & v \sim_a v' \text{ and } f \sim_a f' \\ L^\times(v,f) & = & L(v) \end{array}$$

The satisfaction relation $\models$ to determine $M, v \models \mathsf{pre}(f)$ is defined below, by mutual recursion.

In order to compare the information content of epistemic models we need the notions of *(collective) bisimulation* and *bounded (collective) bisimulation* (*n-bisimulation*) [5, 16].

Definition 5 (Collective bisimulation) *A relation Z between the domains of epistemic models $M = (W, \sim, L)$ and $M' = (W', \sim', L')$ is a* (collective) bisimulation*, notation $Z : M \leftrightarroweq M'$, if for all $(w, w') \in Z$:*

- **atoms***: for all $p_a \in P$, $p_a \in L(w)$ iff $p_a \in L'(w')$;*
- **forth***: for all nonempty $B \subseteq A$ and for all $v \in W$, if $w \sim_B v$ then there is $v' \in W'$ such that $(v, v') \in Z$ and $w' \sim_B v'$;*
- **back***: for all nonempty $B \subseteq A$ and for all $v' \in W'$, if $w' \sim_B v'$ then there is $v \in W$ such that $(v, v') \in Z$ and $w \sim_B v$.*

We additionally define a collective bisimulation bounded by *n, as a set of relations $Z^0 \supseteq Z^1 \cdots \supseteq Z^n$ of i-bisimulations for $0 \leq i \leq n$. Relation Z^0 merely satisfies* **atoms***, and for all $(w, w') \in Z^{n+1}$:*

- **atoms***: for all $p_a \in P$, $p_a \in L(w)$ iff $p_a \in L'(w')$;*
- **forth**-$(n+1)$*: for all nonempty $B \subseteq A$ and for all $v \in W$, if $w \sim_B v$ then there is $v' \in W'$ such that $(v, v') \in Z^n$ and $w' \sim_B v'$.*
- **back**-$(n+1)$*: for all nonempty $B \subseteq A$ and for all $v' \in W'$, if $w' \sim_B v'$ then there is $v \in W$ such that $(v, v') \in Z^n$ and $w \sim_B v$.*

If there is a bisimulation Z between M and M' we write $M \leftrightarroweq M'$, and if there is one containing (w, w') we write $(M, w) \leftrightarroweq (M', w')$. We then say that M and M', respectively (M, w) and (M', w'), are bisimilar*. If Z is bounded by n we write $(M, w) \leftrightarroweq^n (M', w')$ and we say that (M, w) and (M', w') are n-*bisimilar.

Bounded bisimulations are used to compare models (M, w) and (M', w') up to a depth n from the respective points w and w'. Collective n-bisimilarity implies that both models satisfy the same $\mathscr{L}^-$ formulas of modal depth at most n, as a minor variation of the standard result in [5].

To compare dynamic modalities we define *updates* and *update expressivity*.

Definition 6 (Update, update expressivity) *An* update *(or* update relation*) is a binary relation X on a class of pointed epistemic models. Given updates X and Y, X* is update equivalent to *Y, if for all pointed epistemic models (M, w) the update of (M, w) with X is collectively bisimilar to the update of (M, w) with Y. Update modalities $[X]$ and $[Y]$ are update equivalent, if X and Y are update equivalent. (For more refined notions see [7].)*

A language $\mathscr{L}$ is at least as update expressive as *$\mathscr{L}'$ if for every update modality $[X]$ of $\mathscr{L}'$ there is an update modality $[Y]$ of $\mathscr{L}$ such that X is update equivalent to Y. Language $\mathscr{L}$ is* equally update expressive as *$\mathscr{L}'$ (or 'as update expressive as'), if $\mathscr{L}$ is at least as update expressive as $\mathscr{L}'$ and $\mathscr{L}'$ is at least as update expressive as $\mathscr{L}$. Language $\mathscr{L}$ is* (strictly) more update expressive than *$\mathscr{L}'$, if $\mathscr{L}$ is at least as update expressive as $\mathscr{L}'$ and $\mathscr{L}'$ is not at least as update expressive as $\mathscr{L}$. Languages $\mathscr{L}$ and $\mathscr{L}'$ are incomparable in update expressivity if if $\mathscr{L}$ is not at least as update expressive as $\mathscr{L}'$ and $\mathscr{L}'$ is not at least as update expressive as $\mathscr{L}$.*

2.3 Semantics

Definition 7 (Semantics on epistemic models) *Given $M = (W, \sim, L)$ and $w \in W$, the* satisfaction relation *$\models$ is defined by induction on $\varphi \in \mathscr{L}^{\times\circ}$ (where $p \in P$, $a \in A$, $B \subseteq A$, $(\boldsymbol{R}, R)$ a pointed communication* *pattern and $(\boldsymbol{U},e)$ a pointed action model).*

$$\begin{array}{lll}
M,w \models p_a & \textit{iff} & p_a \in L(w) \\
M,w \models \neg\varphi & \textit{iff} & M,w \not\models \varphi \\
M,w \models \varphi \land \psi & \textit{iff} & M,w \models \varphi \textit{ and } M,w \models \psi \\
M,w \models D_B\varphi & \textit{iff} & M,v \models \varphi \textit{ for all } v \sim_B w \\
M,w \models [\boldsymbol{R},R]\varphi & \textit{iff} & M \odot \boldsymbol{R},(w,R) \models \varphi \\
M,w \models [\boldsymbol{U},e]\varphi & \textit{iff} & M,w \models \mathsf{pre}(e) \textit{ implies } M \otimes \boldsymbol{U},(w,e) \models \varphi
\end{array}$$

Formula φ is valid on *M iff for all $w \in W$, $M,w \models \varphi$; formula φ is* valid *iff for all (M,w), $M,w \models \varphi$.*

The (required) locality of epistemic models causes distributed knowledge to have slightly different properties in our semantics. In the standard semantics of distributed knowledge $D_B\varphi \leftrightarrow \varphi$ is invalid for any $B \subseteq A$. Whereas in our semantics $D_A\varphi \leftrightarrow \varphi$ is valid although $D_B\varphi \leftrightarrow \varphi$ for $B \subset A$ remains invalid.

A complete axiomatization of the validities of $\mathscr{L}^\circ$ (communication pattern logic), reducing the dynamics, is given in [6] (similar to [1, 3]). A complete axiomatization of the validities of $\mathscr{L}^\times$ (action model logic), reducing the dynamics, is given in [2]. The language $\mathscr{L}^{\times\circ}$ is not of independent interest.

3 Induced action models for interpreted systems

In this section, let P be finite. From each communication pattern we will construct an *induced action model*. We will show that communication patterns are update equivalent to induced action models when executed in an interpreted system. However, the update of an interpreted system with a communication pattern may not be an interpreted system, and the update of an epistemic model that is not an interpreted system with a communication pattern may not have the same update effect as its induced action model, of which we will give an example.

Definition 8 (Action model induced by a communication pattern) *Given a communication pattern $\boldsymbol{R}$, define* induced action model $\boldsymbol{U}(\boldsymbol{R}) = (E,\sim,\mathsf{pre})$ *as follows (where $R,R' \in \boldsymbol{R}$, $Q,Q' \subseteq P$, $a \in A$).*

$$\begin{array}{lll}
E & = & \boldsymbol{R} \times \mathscr{P}(P) \\
(R,Q) \sim_a (R',Q') & \textit{iff} & Ra = R'a \textit{ and } Q_{Ra} = Q'_{R'a} \\
\mathsf{pre}(R,Q) & = & \delta_Q
\end{array}$$

Informally, this says that two actions are indistinguishable for an agent if the agent receives messages from the same agents ($Ra = R'a$) and if the messages it receives from those agents are the same ($Q_{Ra} = Q'_{R'a}$). As $\boldsymbol{R}$ and P are finite, $\boldsymbol{U}(\boldsymbol{R})$ has a finite domain, so that modality $[\boldsymbol{U}(\boldsymbol{R})]$ is in $\mathscr{L}^\times$. The size of the action model $\boldsymbol{U}(\boldsymbol{R})$ is $|\boldsymbol{R} \times \mathscr{P}(P)| = |\boldsymbol{R}| \cdot 2^{|P|}$. Therefore, $\boldsymbol{U}(\boldsymbol{R})$ is exponentially larger than $\boldsymbol{R}$.

Proposition 9 *Let an interpreted system M and $\boldsymbol{R}$ be given. Then $M \odot \boldsymbol{R}$ is bisimilar to $M \otimes \boldsymbol{U}(\boldsymbol{R})$.*

Proof Let $M = (W,\sim,L)$. Define the following relation Z between (the domains of) $M \odot \boldsymbol{R}$ and $M \otimes \boldsymbol{U}(\boldsymbol{R})$: $Z : (w,R) \mapsto (w,(R,L(w)))$. We show that Z defines a bisimulation.

Let $((w,R),(w,R,L(w)) \in Z$.

atoms: Straightforwardly, $\dot{L}(w,R) = L(w) = L^\times(w,(R,L(w)))$.

forth: Assume $(w,R)\dot{\sim}_B(v,S)$. We claim that $(v,(S,L(v)))$ is the required witness to show **forth**. Obviously $((v,S),(v,(S,L(v))) \in Z$. We also have:

$(w,R)\dot{\sim}_B(v,S) \quad \Leftrightarrow$

for all $a \in B$, $(w,R)\dot{\sim}_a(v,S) \quad \Leftrightarrow$ by definition of $\dot{\sim}_a$

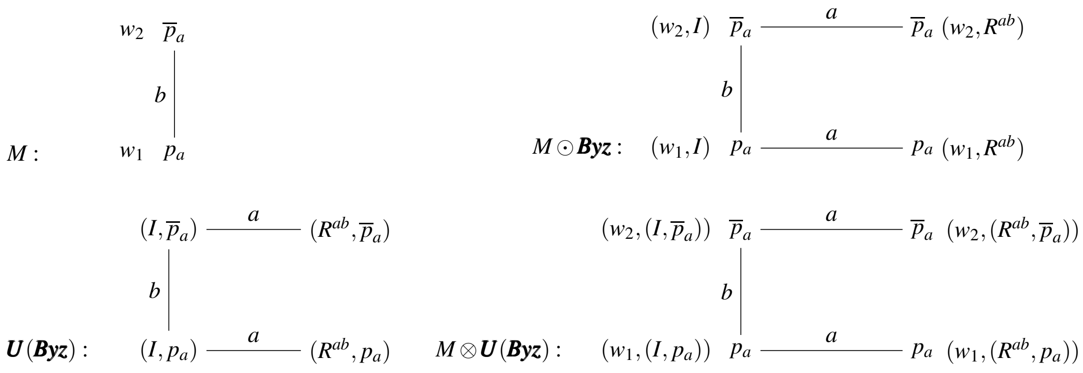

M:
w_2 $\overline{p}_a$
b
w_1 p_a
$M \odot$ **Byz**:
(w_2,I) $\overline{p}_a$
a
$\overline{p}_a$ (w_2,R^{ab})
b
(w_1,I) p_a
a
p_a (w_1,R^{ab})
U(**Byz**):
$(I,\overline{p}_a)$
a
$(R^{ab},\overline{p}_a)$
b
(I,p_a)
a
(R^{ab},p_a)
$M \otimes$ **U**(**Byz**):
$(w_2,(I,\overline{p}_a))$ $\overline{p}_a$
a
$\overline{p}_a$ $(w_2,(R^{ab},\overline{p}_a))$
b
$(w_1,(I,p_a))$ p_a
a
p_a $(w_1,(R^{ab},p_a))$

Figure 1: Communication pattern and action model for Byzantine Generals

for all $a \in B$, $w \sim_{Ra} v$ and $Ra = Sa$ $\quad\Leftrightarrow$ (*)
for all $a \in B$, $w \sim_a v$, $Ra = Sa$, and $L(w)_{Ra} = L(v)_{Sa}$ $\quad\Leftrightarrow$
for all $a \in B$, $w \sim_a v$ and $(R,L(w)) \sim_a (S,L(v))$ $\quad\Leftrightarrow$
for all $a \in B$, $(w,(R,L(w))) \sim_a^\times (v,(S,L(v)))$ $\quad\Leftrightarrow$
$(w,(R,L(w))) \sim_B^\times (v,(S,L(v)))$.

(*): As M is an interpreted system, for all agents $b \in Ra$, $w \sim_b v$ iff $L(w)_b = L(v)_b$, in other words: $w \sim_{Ra} v$ iff $L(w)_{Ra} = L(v)_{Sa}$. As in particular $a \in Ra$, $w \sim_a v$ on the right-hand side of the equation also follows from $L(w)_{Ra} = L(v)_{Sa}$.

back: Similar to **forth**. □

Example 10 (Byzantine generals) *Byzantine attack [13, 9] is a communication pattern given in [17]. Let $A = \{a,b\}$ and $P = \{p_a\}$. Generals a and b wish to schedule an attack, where b desires to learn whether a wants to 'attack at dawn' (p_a) or 'attack at noon' ($\neg p_a$). General a now sends her decision to general b in a message that may fail to arrive. This fits the communication pattern $\boldsymbol{R} = \{I,R^{ab}\}$ where $R^{ab} = I \cup \{(a,b)\}$, which models that a is uncertain whether her message has been received by b. In this instantiation of Byzantine generals, general b has no local variable.*

*The communication pattern **Byz** $= \{I,R^{ab}\}$ where $R^{ab} = I \cup \{(a,b)\}$. We have that $Ia = R^{ab}a = \{a\}$ whereas $Ib = \{b\}$ and $R^{ab}b = \{a,b\}$ (see also [17, Figure 1] and [6, Example 7]).*

*Fig. 1 depicts the initial epistemic model M wherein agent b is uncertain about the value of a variable p_a of agent a, the updated model $M \odot$ **Byz**, the action model **U**(**Byz**), and updated model $M \otimes$ **U**(**Byz**). The states in epistemic models are also labelled with valuations, where p_a stands for $\{p_a\}$ and $\overline{p}_a$ stands for $\emptyset$. Model M is an interpreted system in the vacuous sense that if agent b were to have local variables we could assume their value to be the same in both states. In **U**(**Byz**), the precondition of actions $(I,\{p_a\})$ and $(R^{ab},\{p_a\})$ is p_a, and that of actions $(I,\emptyset)$ and $(R^{ab},\emptyset)$ is $\neg p_a$. (In the figure, for visual consistency, these actions are written as (I,p_a), (R^{ab},p_a), $(I,\overline{p}_a)$, and $(R^{ab},\overline{p}_a)$.) Model $M \otimes$ **U**(**Byz**) is bisimilar, as required, to $M \odot$ **Byz** and even isomorphic.*

When model M is an interpreted system, $M \odot \boldsymbol{R}$ may not be an interpreted system, as, in a way, $M \odot$ **Byz** in Example 10. If agent b were to have local variables, their value would be the same in w_1 and in w_2 and thus also in the four worlds of the updated model. But now agent b has three equivalence classes. It is therefore no longer an interpreted system.

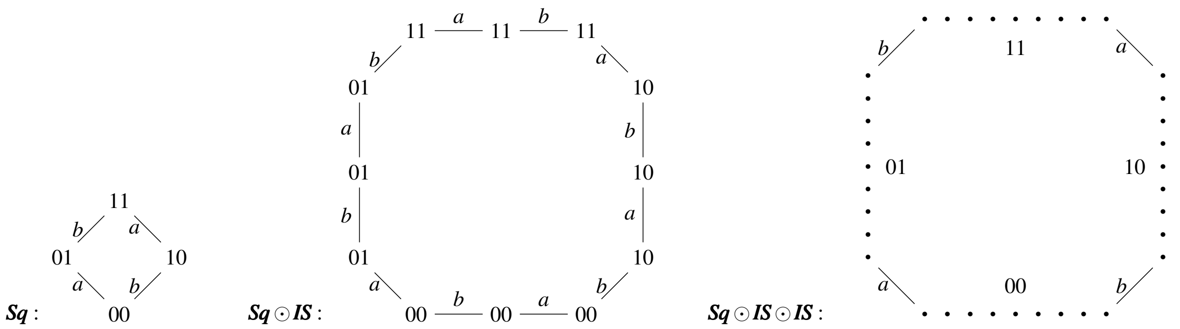


Figure 2: Iterated immediate snapshot for two agents a,b. In world 10 local variable p_a is true and p_b is false (a slightly simpler depiction than $p_a\overline{p}_b$), etcetera. In $\boldsymbol{Sq} \odot \boldsymbol{IS}$ and $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ it is implicit which communication graph is executed, and in $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ valuations are only indicated schematically.

Example 11 (Iterated Immediate Snapshot) *Consider the model* $\boldsymbol{Sq}$ *where* a *knows the truth about* p_a *and* b *knows the truth about* p_b*. This is the interpreted system for two agents each having a single variable. We recall the immediate snapshot (*$\boldsymbol{IS}$*) [11] for two agents* $\{a,b\}$*, defined as* $\{R^{ab},R^{ba},U\}$*, where* $R^{ab} = I \cup \{(a,b)\}$ *and* $R^{ba} = I \cup \{(b,a)\}$*. These three communication graphs, as points of* $\boldsymbol{IS}$*, are commonly denoted as* schedules *consisting of* concurrency classes *a.b, b.a, and ab, respectively. Fig. 2 shows the models* $\boldsymbol{Sq}$*,* $\boldsymbol{Sq} \odot \boldsymbol{IS}$*, and* $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$*. Lemma 12 below shows that iteration of* $\boldsymbol{IS}$ *preserves circularity, as in the figure.*

It follows from Prop. 9 that $\boldsymbol{Sq} \odot \boldsymbol{IS}$ *is bisimilar to* $\boldsymbol{Sq} \otimes \boldsymbol{U}(\boldsymbol{IS})$*. However,* $(\boldsymbol{Sq} \otimes \boldsymbol{U}(\boldsymbol{IS})) \otimes \boldsymbol{U}(\boldsymbol{IS})$ *is not bisimilar to* $(\boldsymbol{Sq} \odot \boldsymbol{IS}) \odot \boldsymbol{IS}$ *and these models therefore satisfy different formulas in comparable worlds. In view of Prop. 9 it is sufficient to show that* $(\boldsymbol{Sq} \odot \boldsymbol{IS}) \otimes \boldsymbol{U}(\boldsymbol{IS})$ *is not bisimilar to* $(\boldsymbol{Sq} \odot \boldsymbol{IS}) \odot \boldsymbol{IS}$*.*

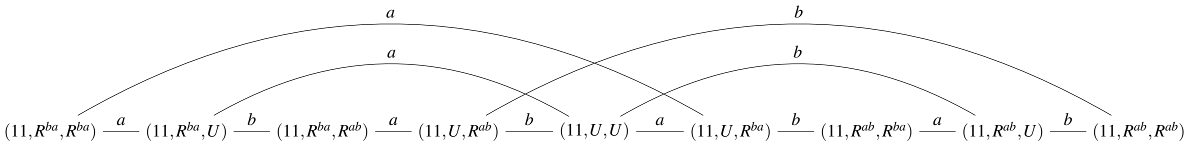
Consider the fragment

$$(11,R^{ba}) \overset{a}{\text{———}} (11,U) \overset{b}{\text{———}} (11,R^{ab})$$

of model $\boldsymbol{Sq} \odot \boldsymbol{IS}$*. This is the top row in Fig. 2. In the model* $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ *this becomes*

$$(11,R^{ba},R^{ba}) \overset{a}{\text{——}} (11,R^{ba},U) \overset{b}{\text{——}} (11,R^{ba},R^{ab}) \overset{a}{\text{——}} (11,U,R^{ab}) \overset{b}{\text{——}} (11,U,U) \overset{a}{\text{——}} (11,U,R^{ba}) \overset{b}{\text{——}} (11,R^{ab},R^{ba}) \overset{a}{\text{——}} (11,R^{ab},U) \overset{b}{\text{——}} (11,R^{ab},R^{ab})$$

Let us now, instead, calculate $\boldsymbol{Sq} \odot \boldsymbol{IS} \otimes \boldsymbol{U}(\boldsymbol{IS})$*. Instead of* $(11,R^{ba},R^{ba})$*—a—*$(11,R^{ba},U)$*, we obtain* $(11,R^{ba},(R^{ba},11))$*—a—*$(11,R^{ba},(U,11))$*. Apart from this edge and other expected edges as above, we now obtain additional edges as below (where we also assume transitivity).*



For example, $(11,R^{ba},(U,11)) \sim_a (11,U,(U,11))$*, because by the semantics of action model execution,* $(11,R^{ba}) \sim_a (11,U)$ *in* $\boldsymbol{Sq} \odot \boldsymbol{IS}$ *and* $(U,11) \sim_a (U,11)$ *in* $\boldsymbol{U}(\boldsymbol{IS})$*. Similarly,* $(11,R^{ba},(R^{ba},11)) \sim_a (11,U,(U,11))$*, because* $(11,R^{ba}) \sim_a (11,U)$ *in* $\boldsymbol{Sq} \odot \boldsymbol{IS}$ *and* $(R^{ba},11) \sim_a (U,11)$ *in* $\boldsymbol{U}(\boldsymbol{IS})$*, where the latter holds because* $R^{ba}a = Ua$ *(namely* $\{a,b\}$*) and* $11_{R^{ba}a} = 11_{Ua}$ *(namely* $11_{\{ab\}}$*, which is* 11*).*

Intuitively, in $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ *the agents learn in the second round whether the communication succeeded in the previous, first, round. But in* $\boldsymbol{Sq} \odot \boldsymbol{IS} \otimes \boldsymbol{U}(\boldsymbol{IS})$ *they do not learn this in the second round.*

It is easy to see that $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ is not bisimilar to $\boldsymbol{Sq} \odot \boldsymbol{IS} \otimes \boldsymbol{U}(\boldsymbol{IS})$ wherein we can reach states in the model with a different valuation in fewer steps. There are then distinguishing formulas, e.g., $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}, (11, U, R^{ba}) \not\models \widehat{K}_a \widehat{K}_b \neg p_a$, whereas $\boldsymbol{Sq} \odot \boldsymbol{IS} \otimes \boldsymbol{U}(\boldsymbol{IS}), (11, U, (R^{ba}, 11)) \models \widehat{K}_a \widehat{K}_b \neg p_a$.

On squares and circles A *circular ab-chain* is an epistemic model consisting of an even number of worlds $0, \dots, 2n-1$, where $n \in \mathbb{N}$ with $n \geq 2$, and such that for all $i \leq n$, $2i \sim_a 2i+1$ and $2i \sim_b 2i-1$ (modulo $2n$).

Lemma 12 *Define $\boldsymbol{Sq} \odot \boldsymbol{IS}^0 := \boldsymbol{Sq}$ and $\boldsymbol{Sq} \odot \boldsymbol{IS}^{n+1} := (\boldsymbol{Sq} \odot \boldsymbol{IS}^n) \odot \boldsymbol{IS}$. For all $n \in \mathbb{N}$, $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$ is a circular ab-chain.*

Proof We prove this by induction.

Model $\boldsymbol{Sq}$ is a (minimal) circular *ab*-chain.

Assuming that $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$ is a circular *ab*-chain, take any world w in that chain and let neighbouring worlds w', w'' be such that $w' \sim_a w$ and $w \sim_b w''$ (where w, w', w'' have arbitrary valuation). We now execute $\boldsymbol{IS}$ once more. Consider the new worlds $(w, R^{ab}), (w, U), (w, R^{ba})$. Then:

- $(w', R^{ab}) \sim_a (w, R^{ab})$ because $R^{ab}a = R^{ab}a$ $(= \{a\})$ and $w' \sim_a w$. No other world than (w', R^{ab}) is indistinguishable for a from (w, R^{ab}). If $R \neq R^{ab}$ then $Ra \neq R^{ab}a$ so $(w', R) \nsim_a (w, R^{ab})$. If $v \neq w, w'$ then $v \nsim_a w$ so $(v, R^{ab}) \nsim_a (w, R^{ab})$.
- $(w, R^{ba}) \sim_b (w'', R^{ba})$ because $R^{ba}b = R^{ba}b$ $(= \{b\})$ and $w \sim_b w''$. Similarly to the previous case this is the unique indistinguishable other world in the updated model.
- $(w, R^{ab}) \sim_b (w, U)$ because $R^{ab}b = Ub$ $(= \{a, b\})$ and $w \sim_{ab} w$. No other world than (w, U) is indistinguishable for b from (w, R^{ab}). We note that $R^{ba}b \neq R^{ab}b$ and $R^{ba}b \neq Ub$, so $(w, R^{ab}) \nsim_b (w, R^{ba})$ and $(w, U) \nsim_b (w, R^{ba})$. If $v \neq w$ then $v \nsim_{ab} w$ so $(w, R^{ab}) \nsim_b (v, R^{ab})$ and $(w, U) \nsim_b (v, U)$.
- $(w, R^{ba}) \sim_a (w, U)$ because $R^{ba}a = Ua$ $(= \{a, b\})$ and $w \sim_{ab} w$. Similarly to the previous case this is the unique indistinguishable other world in the updated model.

□

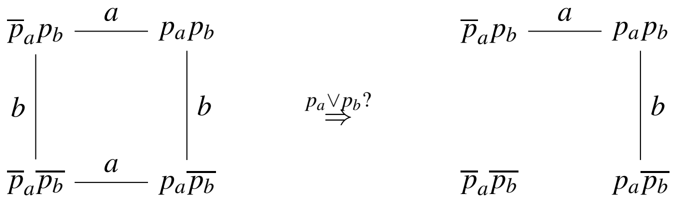
This result is not surprising. In the corresponding representation as simplicial complexes, an application of $\boldsymbol{IS}$ is a so-called *subdivision* [11]. A circular *ab*-chain corresponds to a circular graph (1-dimensional complex) with alternating a and b nodes, such that each edge *a*—*b* gets replaced by three edges *a*—*b*—*a*—*b* at each iteration of $\boldsymbol{IS}$ (and duplicated nodes keeps their old labels).

4 Communication patterns and action models are incomparable

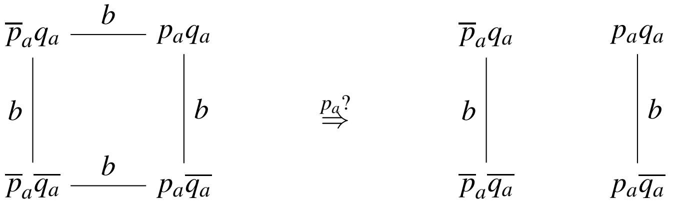
Proposition 13 *Communication pattern logic is not at least as update expressive as action model logic.*

Proof We can prove this in different ways, which seems instructive.

First, in a public announcement, the environment may reveal something that cannot be revealed by the agents individually or jointly, such as the announcement whether $p_a \vee p_b$ in a model where a knows whether p_a and b knows whether p_b.



Second, agents may choose to reveal some but not all of their local variables, such as, if a knows whether p_a and whether q_a, a informing b of the truth about p_a but not about q_a.



Third, there are action models that produce more uncertainty than any communication pattern. Here we should note that although the composition of two action models is again an action model (therefore, for all $\boldsymbol{U},\boldsymbol{U}'$ there is a $\boldsymbol{U}''$, namely the composition of $\boldsymbol{U}$ and $\boldsymbol{U}'$, such that $[\boldsymbol{U}][\boldsymbol{U}']\varphi \leftrightarrow [\boldsymbol{U}'']\varphi)$, sequentially executing two communication patterns is typically not the same as executing a single communication pattern (it is not the case that for all $\boldsymbol{R},\boldsymbol{R}'$ there is a $\boldsymbol{R}''$ such that $[\boldsymbol{R}][\boldsymbol{R}']\varphi \leftrightarrow [\boldsymbol{R}'']\varphi)$. For example, consider the models $\boldsymbol{Sq}$ and $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ (Example 11). The domain of model $\boldsymbol{Sq}$ consists of four worlds and that of $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ consists of 36 worlds; it is nine times larger (and it is bisimulation minimal). Now there are only four different communication patterns for two agents (namely I, R^{ba}, R^{ab}, and U). So the maximum size of a model resulting from updating $\boldsymbol{Sq}$ with a communication pattern is 16. Therefore there is no such communication pattern. In other words, there is no $\boldsymbol{R}$ such that $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ is bisimilar to $\boldsymbol{Sq} \odot \boldsymbol{R}$ which implies that there is no $\boldsymbol{R}$ that has the same update effect as updating twice with $\boldsymbol{IS}$.

However, there is an action model $\boldsymbol{U}$ such that $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ is bisimilar to $\boldsymbol{Sq} \otimes \boldsymbol{U}$: its domain is the domain of $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$; its relations for a and b are the relations for a and b on the model $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$, and its preconditions are such that the precondition of a world (ij,R,R') in the domain of $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ is the description δ_{ij} of the valuation ij. It is straightforward to see that $\boldsymbol{Sq} \odot \boldsymbol{IS} \odot \boldsymbol{IS}$ is even isomorphic to $\boldsymbol{Sq} \otimes \boldsymbol{U}$.

We conclude that there is no communication pattern that is update equivalent to this action model $\boldsymbol{U}$. Therefore, communication pattern logic is not at least as update expressive as action model logic. □

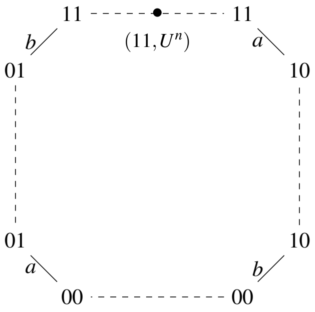
We continue by showing that action model logic is not at least as update expressive as communication pattern logic. If multi-pointed action models had not been allowed, a trivial way to show that, would have been to observe that single-pointed action models unlike communication patterns may not always be executable. Although true, that is not of interest. We prove this in a more meaningful way in the following Prop. 14. Its proof assumes towards a contradiction that an action model $\boldsymbol{U}$ exists that is update equivalent to the communication pattern $\boldsymbol{IS}$, where we identify $\boldsymbol{U}$ with the multi-pointed action model $(\boldsymbol{U}, \mathscr{D}(\boldsymbol{U}))$. We then compare the updates $\boldsymbol{IS}$ and $\boldsymbol{U}$ in epistemic model $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$ for n exceeding a function of the modal depth of any precondition of $\boldsymbol{U}$, and derive a contradiction. It may assist the reader to know that Ex. 11 above replays this proof for $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{IS})$ of which the action preconditions are booleans, such that $md(\boldsymbol{U}) = 0$ and we can choose $n = 1$.

Proposition 14 *Action model logic is not at least as update expressive as communication pattern logic.*

Proof Suppose towards a contradiction that communication pattern $\boldsymbol{IS}$ is update equivalent to an action model $\boldsymbol{U} = (E, \sim, \mathsf{pre})$.

What do we know about $\boldsymbol{U}$? As $\boldsymbol{IS}$ is always executable, we may assume that the disjunction ψ of all preconditions of actions e in the domain E of $\boldsymbol{U}$ is the triviality. Otherwise, given some model with $M,w \models \neg\psi$, we could update with $\boldsymbol{IS}$ but not with $\boldsymbol{U}$. Similarly, for any action e in the domain E of $\boldsymbol{U}$, there must be $f \in E$ such that $e \sim_a f$ and $\mathsf{pre}(e) = \mathsf{pre}(f)$ (and for agent b there must be a $g \in E$ such that $g \sim_b f$ and $\mathsf{pre}(g) = \mathsf{pre}(f)$). Otherwise, consider a model (M,w) that can only be updated with $(\boldsymbol{U},e)$ (for which $M,w \models \mathsf{pre}(e)$). It can be updated with $(\boldsymbol{IS},U)$ and also with $(\boldsymbol{IS},R^{ba})$ resulting in states (w,U) and (w,R^{ba}) satisfying different properties, as $(w,U) \sim_a (w,R^{ba})$ (because $Ua = R^{ba} = \{a,b\}$), so that one or the other but not both can be bisimilar to (w,e). Therefore, $\boldsymbol{U}$ must be a refinement of $\boldsymbol{IS}$ seen as a structure R^{ab}—b—U—a—R^{ba}. Its actions can therefore be assumed to have shape (R,φ) where R is one of R^{ab},U,R^{ba} and where $\varphi \in \mathscr{L}^{\times}$ is the precondition of that action, that is, $\mathsf{pre}(R,\varphi) = \varphi$.[1]

The modality $[\boldsymbol{U}]$ is an operator in the language $\mathscr{L}^{\times}$ and $|E|$ is finite, so that $md(\boldsymbol{U}) = \max\{md(\mathsf{pre}(e)) \mid e \in E\}$ is defined. Choose $n \in \mathbb{N}$ with $n > \log_3 2(md(\boldsymbol{U})+1)$ and consider $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$, schematically depicted as:



$\boldsymbol{Sq} \odot \boldsymbol{IS}^n$:

and concretely its three-action fragment:

$$(*): \quad (11,U^{n-1}R^{ba}) \stackrel{a}{\rule{3em}{0.4pt}} (11,U^n) \stackrel{b}{\rule{3em}{0.4pt}} (11,U^{n-1}R^{ab})$$

where world $(11,U^n)$ of $(*)$ is the same as the depicted world $(11,U^n)$ of $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$.

We can now justify the bound $n > \log_3 2(md(\boldsymbol{U})+1)$. We need in the proof that the three worlds of $(*)$ satisfy the same actions of $\boldsymbol{U}$, and we guarantee that because they are bounded collectively bisimilar for an appropriate bound. Given $(11,U^n)$, the bound should exceed the modal depth of any possible precondition of any action in $\boldsymbol{U}$. That explains $md(\boldsymbol{U})$. Plus one, as we need this to hold for the surrounding worlds too. That explains $md(\boldsymbol{U})+1$. Twice that, $2\cdot(md(\boldsymbol{U})+1)$, is the required length of one side of the squarish model $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$ with therefore $8\cdot(md(\boldsymbol{U})+1)$ worlds. Starting with four worlds, every iteration of $\boldsymbol{IS}$ multiplies the number of worlds by 3. So we therefore want to iterate $\boldsymbol{IS}$ by some n such that $4\cdot 3^n > 8\cdot(md(\boldsymbol{U})+1)$, that is, $n > \log_3 2(md(\boldsymbol{U})+1)$.

Consider $\boldsymbol{Sq} \odot \boldsymbol{IS}^n \otimes \boldsymbol{U}$. Recalling what is known about $\boldsymbol{U}$, there must be an $e \in E$ such that $\boldsymbol{Sq} \odot \boldsymbol{IS}^n,(11,U^n) \models \mathsf{pre}(e)$. Also, there must be $f,g \in E$ with $e \sim_a f$ and $f \sim_b g$ and $\mathsf{pre}(e) = \mathsf{pre}(f) = \mathsf{pre}(e)$. Let $\mathsf{pre}(e)$ be θ. These actions e,f,g therefore have shape (R^{ab},θ), (U,θ), (R^{ba},θ) respectively.

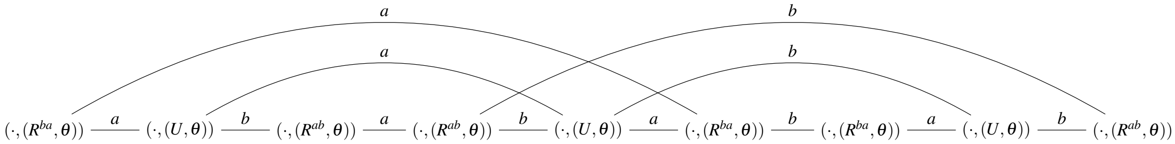
As $n > \log_3 2(md(\boldsymbol{U})+1)$, the three worlds in $(*)$ are bounded collectively bisimilar:

$$(\boldsymbol{Sq} \odot \boldsymbol{IS}^n,(11,U^{n-1},R^{ba})) \underline{\leftrightarrow}^{md(\boldsymbol{U})+1} (\boldsymbol{Sq} \odot \boldsymbol{IS}^n,(11,U^n)) \underline{\leftrightarrow}^{md(\boldsymbol{U})+1} (\boldsymbol{Sq} \odot \boldsymbol{IS}^n,(11,U^{n-1},R^{ab}))$$

As $md(\theta) \leq md(\boldsymbol{U})$, all three worlds in $(*)$ satisfy θ, so actions e,f,g can be executed in all these worlds.

The model $\boldsymbol{Sq} \odot \boldsymbol{IS}^n \otimes \boldsymbol{U}$ therefore contains the submodel

[1] By *refinement* we mean that R^{ab} can be seen as an equivalence class $\{(R^{ab},\varphi) \mid (R^{ab},\varphi) \in \mathscr{D}(\boldsymbol{U})\}$, and similarly for U and R^{ba}, where two such equivalence classes are indistinguishable for a if there are $(R,\varphi),(R',\varphi')$ such that $(R,\varphi) \sim_a (R',\varphi')$, and similarly for b.



wherein only some additional pairs for $\sim_a$ and $\sim_b$ are shown, and where from those shown we merely justify one as an example: for the leftmost and the middle worlds, we have that $(11,U^{n-1},R^{ba},(R^{ba},\theta)) \sim_a (11,U^n,(U,\theta))$, because by the semantics of action model execution, $(11,U^{n-1},R^{ba}) \sim_a (11,U^n)$ in $\boldsymbol{Sq} \odot \boldsymbol{IS}^n$ and $(R^{ba},\theta) \sim_a (U,\theta)$ in $\boldsymbol{U}(\boldsymbol{IS})$. Furthermore (unlike in Example 11), worlds $(\dots,(R,\theta))$ shown, may be indistinguishable for a or b from worlds $(\dots,(R,\xi))$ not shown, for actions (R,ξ) with ξ non-equivalent to θ.

Consequently, $\boldsymbol{Sq} \odot \boldsymbol{IS}^n \otimes \boldsymbol{U}$ is not a circular ab-chain like $\boldsymbol{Sq} \odot \boldsymbol{IS}^{n+1}$ that locally looks like:

$$(\cdot,R^{ba}) \overset{a}{\rule{3em}{0.4pt}} (\cdot,U) \overset{b}{\rule{3em}{0.4pt}} (\cdot,R^{ab}) \overset{a}{\rule{3em}{0.4pt}} (\cdot,R^{ab}) \overset{b}{\rule{3em}{0.4pt}} (\cdot,U) \overset{a}{\rule{3em}{0.4pt}} (\cdot,R^{ba}) \overset{b}{\rule{3em}{0.4pt}} (\cdot,R^{ba}) \overset{a}{\rule{3em}{0.4pt}} (\cdot,U) \overset{b}{\rule{3em}{0.4pt}} (\cdot,R^{ab})$$

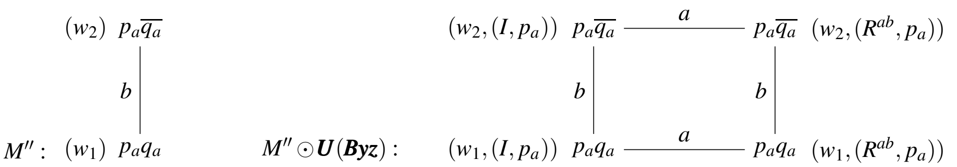
Now the assumption of update equivalence implies that $\boldsymbol{Sq} \odot \boldsymbol{IS}^{n+1}$ is collectively bisimilar to $\boldsymbol{Sq} \odot \boldsymbol{IS}^n \otimes \boldsymbol{U}$. The supposed bisimulation relation Z linking $\boldsymbol{Sq} \odot \boldsymbol{IS}^{n+1}$ and $\boldsymbol{Sq} \odot \boldsymbol{IS}^n \otimes \boldsymbol{U}$ should therefore such that $Z : (w,\sigma,R) \mapsto (w,\sigma,(R,\mathsf{pre}(e))$ for all $w \in W$, $\sigma \in \boldsymbol{IS}^n$, and $e \in E$ with $\boldsymbol{Sq} \odot \boldsymbol{IS}^n,(w,\sigma) \models \mathsf{pre}(e)$, in particular the three worlds in $(*)$ and the e,f,g above with preconditions θ. On the other hand, clearly, a pair of worlds in this relation cannot be bisimilar, as the additional a-links and b-links allow shorter paths to a 01-world. Differently said, as bounded bisimilarity implies the same truth value for formulas of at most that modal depth, the worlds in such a pair satisfy different formulas. (See Ex. 11 for $n = 1$.)

This contradicts our assumption that $\boldsymbol{U}$ is update equivalent to $\boldsymbol{IS}$ and thus concludes the proof. □

Prop. 14 holds for any countable set of local variables P. In the proof of Prop. 14 we only need two: $P = \{p_a, p_b\}$. When P is countably infinite there is a shorter proof of Prop. 14, given below.

Proof Let P be countably infinite. Suppose towards a contradiction that there is an action model $\boldsymbol{U}$ with $[\boldsymbol{U}]$ (or $[\boldsymbol{U},e]$) in the logical language (so that the domain of $\boldsymbol{U}$ is necessarily finite) that is update equivalent to $\boldsymbol{Byz}$. As $\boldsymbol{U}$ is finite and P is countably infinite, there exists a $q_a \in P$ not occurring in any of the preconditions of the actions in the domain of $\boldsymbol{U}$. Now consider epistemic model M'' as in Example 10 but with q_a true in w_1 and false in w_2 and with p_a true in both worlds. When executing $\boldsymbol{U}$ in M'', the update $M'' \otimes \boldsymbol{U}$ will never get the required asymmetry of $M'' \odot \boldsymbol{Byz}$, because any action (point) e that is executable in w_1 is also executable in w_2, as for any $p \in P \setminus \{q_a\}$, $p \in L(w_1)$ iff $p \in L(w_2)$. In particular we therefore will have that $(w_1,I,\dots) \sim_b (w_2,I,\dots)$ iff $(w_1,R^{ab},\dots) \sim_b (w_2,R^{ab},\dots)$. (An argument involving $\sim_a$ and $\sim_b$ similar to the one in the proof of Prop. 14 is omitted for brevity.)

More simply said, if we were to execute $\boldsymbol{U}(\boldsymbol{Byz})$ of Example 10 in that model M'', the following model would result (as p_a is true in w_1 and w_2, the alternatives with precondition $\neg p_a$ never execute):



□

Corollary 15 *Communication pattern logic and action model logic are incomparable in update expressivity.*

5 Communication patterns for history-based structures

Example 11 demonstrated that interpreted systems are not closed under update with communication patterns. We therefore could not obtain a result for update expressivity for the class of interpreted systems. In this section we show that this is after all possible if we adjust the structures in which we execute updates and simultaneously adjust the definition of the update. In order to store the sequence of past events we generalize our epistemic models to history-based epistemic models [4, 8]. Simultaneously, we change the semantics of the update with a communication pattern namely by having this depend on the number of previous updates that already took place, what is known as the number of previous *rounds* in an oblivious protocol arbitrarily often executing that communication pattern. The change consists in recording the information of previous rounds in designated history variables, that store the *view* for each agent on all previous rounds. These variables are also local.

Example 16 *When updating epistemic model $(\boldsymbol{Sq},11)$ with communication pattern $(\boldsymbol{IS},R^{ab})$, we record that $R^{ab}a = \{a\}$ and $R^{ab}b = \{a,b\}$ in the resulting world $(11,R^{ab})$ by indexing these sets with the names of the agents, so as $\{a\}_a$ and $\{a,b\}_b$, that we write as a_a and ab_b. These are local variables. Then, when updating $(\boldsymbol{Sq}\odot\boldsymbol{IS},(11,R^{ab}))$ with $(\boldsymbol{IS},R^{ba})$, we record the entire history so far for a and b, where a but not b also receives b's history of the previous round, as $((a,ab).ab)_a$ for agent a and $(ab.b)_b$ for agent b.*

We explain the first. As a receives information from b, and by default from itself, $R^{ba}a = \{a,b\}$, written as ab, is preceded by the list $(\{a\},\{a,b\})$ containing $R^{ab}a = \{a\}$ and $R^{ab}b = \{a,b\}$ of the previous round, which is written as (a,ab). The expression $(a,ab).ab$ is the view *of agent a on the history, which is a tree. This view is indexed with the name a of the agent, such that $((a,ab).ab)_a$ is a local variable for agent a, wherein the views of a and of b in the previous round are lexicographically ordered.*

And so on for every next round. Such history variables are designated local variables, initially false.

We adapt the semantics of update $\odot$ by making history variables for a given round of communication true after the update representing that round. We name this semantics $\dot\odot$.

- *in $(\boldsymbol{Sq},11)$, variables p_a and p_b are true and all others false;*
- *in $(\boldsymbol{Sq}\dot\odot\boldsymbol{IS},(11,R^{ab}))$, variables p_a,p_b,a_a,ab_b are true and all others false;*
- *in $(\boldsymbol{Sq}\dot\odot\boldsymbol{IS}\dot\odot\boldsymbol{IS},(11,R^{ab},R^{ba}))$, variables $p_a,p_b,a_a,ab_b,((a,ab).ab)_a,(ab.b)_b$ are true and ...*

The $\dot\odot$ semantics is then closed for the class of interpreted systems. We proceed with formalities.

Definition 17 (View, history variable) *Let a communication pattern $\boldsymbol{R}$ be given. A* history *is a member $\sigma \in \boldsymbol{R}^*$ (a finite sequence of communication graphs in $\boldsymbol{R}$). The* view *of $a \in A$ on history σ is defined as:*

$$\mathsf{view}_a(\varepsilon) \;:=\; \emptyset \qquad\qquad \mathsf{view}_a(\sigma.R) \;:=\; \mathsf{view}_{Ra}(\sigma).Ra$$

where $\mathsf{view}_{Ra}(\sigma)$ is the ordered list of views $\mathsf{view}_b(\sigma)$ for $b \in Ra$. The set of history variables *is $\boldsymbol{\Sigma} := \{(\mathsf{view}_a(\sigma))_a \mid \sigma \text{ a history}, a \in A\}$. Also, $\Sigma^n := \{(\mathsf{view}_a(\sigma))_a \in \boldsymbol{\Sigma} \mid |\sigma| = n, a \in A\}$, and $\Sigma^{<n} = \bigcup_{m<n}\Sigma^m$.*

The view of agent a on history σ defines a **tree** with root Ra where R is the last element of σ. A history variable for a is nothing but the view of a of a history σ, subscripted with a, denoting a local variable. The set of views is known as the *full-information protocol* [14]. We now model the arbitrary iteration of a communication pattern in an epistemic model, while keeping track of the previous rounds by way of history variables. The definition is for agents A and variables $P \cup \boldsymbol{\Sigma}$ (and not, as before, for A and P).

Definition 18 (History epistemic model) *Given an epistemic model $M = (W,\sim,L)$, a communication pattern $\boldsymbol{R}$, and $n \in \mathbb{N}$, a* history epistemic model *$M\dot\odot\boldsymbol{R}^n$ is defined as follows. For $n = 0$, $M\dot\odot\boldsymbol{R}^0 = M$. For $n \geq 0$, given $M\dot\odot\boldsymbol{R}^n = (W \times \boldsymbol{R}^n,\sim,L)$, we define $M\dot\odot\boldsymbol{R}^{n+1} := (W \times \boldsymbol{R}^{n+1},\sim',L')^2$ such that:*

[2] Allowing slight abuse of the notation $\boldsymbol{R}^n$.

- $(w,\sigma.R) \sim'_a (w',\sigma'.R')$ *iff* $(w,\sigma) \sim_{Ra} (w',\sigma')$ *and* $Ra = R'a$*;*
- $L'(w,\sigma.R) := L(w,\sigma) \cup \{(\mathsf{view}_b(\sigma.R))_b \mid b \in A\}$.

The domain of $M\dot{\odot}\boldsymbol{R}^n$ is $W \times \boldsymbol{R}^n$, so that domain elements have shape (w,σ). As $M\dot{\odot}\boldsymbol{R}^0 = M$, all history variables in $M\dot{\odot}\boldsymbol{R}^0$ are false. This means that no round of communication has taken place.

The difference between the $\dot{\odot}$ update and the $\odot$ update is therefore *only* in the labeling of local variables: we now require a countably infinite set of local history variables such that in each round for each agent the entire history is again recorded by making such a variable true. We will see that this guarantees that interpreted systems are closed under update.

Given the $\dot{\odot}$ update, the history-based semantics is now as expected, and unlike the previous semantics it has the property that the update of an interpreted system remains an interpreted system.

Definition 19 (History-based semantics) *Given* $M\dot{\odot}\boldsymbol{R}^n = (W \times \boldsymbol{R}^n, \sim, L)$ *and* $(w,\sigma) \in W$*, define* satisfaction relation $\models$ *by induction on* $\varphi \in \mathscr{L}$ *(where* $p \in P$*,* $a \in A$*,* $B \subseteq A$*,* $\boldsymbol{R}$ *a communication pattern,* $R \in \boldsymbol{R}$*,* $\sigma \in \boldsymbol{R}^n$*, and* $\tau \in \boldsymbol{R}^*$ *— that is,* τ *is an arbitrary history).*

$$\begin{array}{lll}
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models p_a & \textit{iff} & p_a \in L(w) \\
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models (\mathsf{view}_a(\tau))_a & \textit{iff} & (\mathsf{view}_a(\tau))_a \in L(w,\sigma) \\
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models \neg\varphi & \textit{iff} & M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \not\models \varphi \\
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models \varphi \wedge \psi & \textit{iff} & M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models \varphi \textit{ and } M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models \psi \\
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models D_B\varphi & \textit{iff} & M\dot{\odot}\boldsymbol{R}^n,(v,\tau) \models \varphi \textit{ for all } (v,\tau) \sim_B (w,\sigma) \\
M\dot{\odot}\boldsymbol{R}^n,(w,\sigma) \models [\boldsymbol{R},R]\varphi & \textit{iff} & M\dot{\odot}\boldsymbol{R}^{n+1},(w,\sigma.R) \models \varphi
\end{array}$$

Proposition 20 *Let interpreted system* M *and* $\boldsymbol{R}$ *be given. Then* $M\dot{\odot}\boldsymbol{R}^n$ *is an interpreted system.*

Proof Let $M\dot{\odot}\boldsymbol{R}^n = (W \times \boldsymbol{R}^n, \sim, L)$. We are required to show that $(w,\sigma) \sim_a (w',\sigma')$ iff $L(w,\sigma)_a = L(w',\sigma')_a$.

For $n = 0$ this is because M is an interpreted system.

Let us now assume $M\dot{\odot}\boldsymbol{R}^n$ is an interpreted system and consider $M\dot{\odot}\boldsymbol{R}^{n+1}$, and $R, R' \in \boldsymbol{R}$. We then have that (where $|\sigma| = |\sigma'| = n$):

$$\begin{array}{lr}
(w,\sigma.R) \sim_a (w',\sigma'.R') & \\
\Leftrightarrow & \text{by definition of } \sim_a \\
Ra = R'a \text{ and } (w,\sigma) \sim_{Ra} (w',\sigma') & \\
\Leftrightarrow & \\
Ra = R'a, \text{ and for all } b \in Ra : (w,\sigma) \sim_b (w',\sigma') & \\
\Leftrightarrow & \text{inductive hypothesis} \\
Ra = R'a, \text{ and for all } b \in Ra : L(w,\sigma)_b = L(w',\sigma')_b & \\
\Leftrightarrow & (*) \\
L(w,\sigma.R)_a = L(w',\sigma'.R')_a &
\end{array}$$

$(*)$: By definition, we have that $L(w,\sigma.R) = L(w,\sigma) \cup \{(\mathsf{view}_b(\sigma.R))_b \mid b \in A\}$. Therefore, for agent a, we have that $L(w,\sigma.R)_a = L(w,\sigma)_a \cup \{(\mathsf{view}_a(\sigma.R))_a\}$. As $a \in Ra$, we may assume by induction that $L(w,\sigma)_a = L(w',\sigma')_a$. It therefore remains to show that $(\mathsf{view}_a(\sigma.R))_a = (\mathsf{view}_a(\sigma'.R'))_a$. By the definition of view, this is equivalent to requiring that $Ra = R'a$, and that $(\mathsf{view}_b(\sigma))_b = (\mathsf{view}_b(\sigma'))_b$ for all $b \in Ra$. The latter is given above. Concerning the former: from the inductive assumption that $L(w,\sigma)_b = L(w',\sigma')_b$ for all $b \in Ra$ and the definition of view for these agents b it follows that $(\mathsf{view}_b(\sigma))_b = (\mathsf{view}_b(\sigma'))_b$ for all $b \in Ra$. □

In order to compare the update expressivity of action models and communication patterns in this semantics, we must also change Def. 8 of induced action model $\boldsymbol{U}(\boldsymbol{R})$. There are now infinitely many local variables, so that the description of a valuation is no longer a formula but an infinite conjunction. However, for every round of communication a description of the valuation of a finite subset is sufficient.

Definition 21 (Induced action model for round n) *The induced action model* $\boldsymbol{U}^n(\boldsymbol{R}) = (E, \sim, \mathsf{pre})$ *for round n of iterated execution of* $\boldsymbol{R}$ *is defined as follows, where* $R, R' \in \boldsymbol{R}$*,* $Q, Q' \subseteq P \cup \Sigma^{<n}$*, and* $a \in A$*:*

$$\begin{array}{lll} E & = & \boldsymbol{R} \times \mathscr{P}(P \cup \Sigma^{<n}) \\ (R,Q) \sim_a (R',Q') & \textit{iff} & Ra = R'a \textit{ and } Q_{Ra} = Q'_{R'a} \\ \mathsf{pre}(R,Q) & = & \delta_{Q,P\cup\Sigma^{<n}} \end{array}$$

Although $P \cup \boldsymbol{\Sigma}$ infinite, $P \cup \Sigma^{<n}$ is finite. Note that $\boldsymbol{U}(\boldsymbol{R})$ is $\boldsymbol{U}^1(\boldsymbol{R})$, where δ_Q is now $\delta_{Q,P}$, as $\Sigma^{<1} = \Sigma^0 = \emptyset$. We recall the definition of $\delta_{Q,P\cup\Sigma^{<n}}$ from Sect. 2.1. From Prop. 20 and Prop. 9 we directly obtain:

Proposition 22 *Let interpreted system M and communication pattern* $\boldsymbol{R}$ *be given. Then* $M \dot{\odot} \boldsymbol{R}^n$ *is bisimilar to* $M \otimes \boldsymbol{U}^1(\boldsymbol{R}) \otimes \cdots \otimes \boldsymbol{U}^n(\boldsymbol{R})$*.*

By abbreviation inductively define $(\boldsymbol{R}^0, \varepsilon) := \varepsilon$ and $(\boldsymbol{R}^{n+1}, \sigma.R) := (\boldsymbol{R}^n, \sigma).(\boldsymbol{R}, R)$, where $\sigma \in \boldsymbol{R}^n$. Recalling the definition of $[\boldsymbol{R}]\varphi$ as $\bigwedge_{R\in\boldsymbol{R}}[\boldsymbol{R},R]\varphi$, we let $[\boldsymbol{R}^n]\varphi$ stand for $\bigwedge_{\sigma\in\boldsymbol{R}^n}[\boldsymbol{R}^n,\sigma]\varphi$. Just as $[\boldsymbol{R}^n]\varphi$ is equivalent to $[\boldsymbol{R}]^n\varphi$, $[\boldsymbol{R},\sigma]\varphi$ is equivalent to $[\boldsymbol{R},R_1]\dots[\boldsymbol{R},R_n]\varphi$, where $\sigma = R_1 \dots R_n$.

In the $\odot$ semantics, the answer to the question whether a communication pattern $(\boldsymbol{R}, R)$ is update equivalent to an action model $(\boldsymbol{U}(\boldsymbol{R}), T)$ where $T = \{(R,Q) \mid Q \subseteq P\}$, on the class of epistemic models, was 'no' (Example 11). This now becomes the question whether in the history-based $\dot{\odot}$ semantics an iterated communication pattern $(\boldsymbol{R}^n, \sigma)$ is update equivalent to a multi-pointed action model on the class of interpreted systems with empty histories. The answer to that is 'yes'. However, communication pattern modalities occurring in a formula may not be interpreted in the empty history. For example, given $[\boldsymbol{R},R](p_a \to D_B[\boldsymbol{R},R']p_b)$, subformula $[\boldsymbol{R},R']p_b$ will be interpreted in some world (w,R), not in some world (w,ε). We want it equivalent to some formula of shape $[\boldsymbol{R},RR']\psi$. We therefore show that any formula $\varphi \in \mathscr{L}^\circ$ is equivalent to one wherein all subformulas $[\boldsymbol{R}^n,\sigma]\psi$ have that $\psi \in \mathscr{L}^-$ (without dynamic modalities). All dynamic modalities are then interpreted in an empty history epistemic model.

Define the *iterated update normal form* (IUNF), the language $\mathscr{L}^\circ_{\mathsf{iunf}}$ (with members φ) by BNF as:

$$\begin{array}{lll} \varphi & := & p_a \mid \neg\varphi \mid \varphi \wedge \varphi \mid D_B\varphi \mid [\boldsymbol{R}^n,\sigma]\psi \\ \psi & := & p_a \mid \neg\psi \mid \psi \wedge \psi \mid D_B\psi \end{array}$$

Lemma 23 *Every formula in* $\mathscr{L}^\circ$ *is equivalent to one in* $\mathscr{L}^\circ_{\mathsf{iunf}}$*, in iterated update normal form.*

Proof We define a translation $t : \mathscr{L}^\circ \to \mathscr{L}^\circ_{\mathsf{iunf}}$. We prove by induction that any φ is equivalent to $t(\varphi)$. All clauses are trivial, and the one for the dynamic modality has a subinduction. The subinduction uses the reduction axioms for communication patterns found in [6].

$$\begin{array}{lll} t([\boldsymbol{R},R]p_a) & := & [\boldsymbol{R},R]p_a \\ t([\boldsymbol{R},R](\varphi\wedge\psi)) & := & t([\boldsymbol{R},R]\varphi) \wedge t([\boldsymbol{R},R]\psi) \\ t([\boldsymbol{R},R]\neg\varphi) & := & \neg t([\boldsymbol{R},R]\varphi) \\ t([\boldsymbol{R},R]D_B\varphi) & := & \bigwedge_{R'B\equiv RB} D_{RB}t([\boldsymbol{R},R']\varphi) \\ t([\boldsymbol{R},R][\boldsymbol{R}',R']\varphi) & := & t([\boldsymbol{R},R]t([\boldsymbol{R}',R']\varphi)) \end{array}$$

In particular, we have that

$$\begin{array}{lclclcl} t([\boldsymbol{R},R](\varphi\vee\psi)) & = & t([\boldsymbol{R},R]\neg(\neg\varphi\wedge\neg\psi)) & = & \neg t([\boldsymbol{R},R](\neg\varphi\wedge\neg\psi)) & = \\ \neg(t([\boldsymbol{R},R]\neg\varphi)\wedge t([\boldsymbol{R},R]\neg\psi)) & = & \neg(\neg t([\boldsymbol{R},R]\varphi)\wedge\neg t([\boldsymbol{R},R]\psi)) & = & t([\boldsymbol{R},R]\varphi)\vee t([\boldsymbol{R},R]\psi) \end{array}$$

We recall that notation $R'B \equiv RB$ was defined in Sect. 2.2. Further proof details are omitted. □

Proposition 24 *Action model logic is at least as update expressive as communication pattern logic on the class of interpreted systems, in the history-based semantics.*

Proof Let an interpreted system M and a communication pattern $\boldsymbol{R}$ be given. Then $M\dot{\odot}\boldsymbol{R}^n$ is bisimilar to $M\otimes\boldsymbol{U}^1(\boldsymbol{R})\otimes\cdots\otimes\boldsymbol{U}^n(\boldsymbol{R})$ (Prop. 22). Consider the action model $\boldsymbol{U}$ that is the *composition* of $\boldsymbol{U}^1(\boldsymbol{R})$, $\ldots$, $\boldsymbol{U}^n(\boldsymbol{R})$, where we note that, unlike communication patterns, action models are indeed closed under composition (see [2] for the definition of action model composition).

Let us now consider what action model some $(\boldsymbol{R}^n,\sigma)$ is update equivalent to. We can assume that modalities $[\boldsymbol{R}^n,\sigma]$ are only interpreted in M, a history epistemic model for an empty history (Lemma 23). Iterated communication pattern $\boldsymbol{R}^n$ is clearly update equivalent to $\boldsymbol{U}$. It is almost worded as such in Prop. 24. Also, any $(\boldsymbol{R}^n,\sigma)$ is update equivalent to $(\boldsymbol{U},T)$, where, if $\sigma = R^1R^2\ldots R^n$,

$$T = \{(R^1,Q^1)(R^2,Q^2)\ldots(R^n,Q^n) \mid Q^1\subseteq P, Q^2\subseteq\Sigma^1,\ldots,Q^n\subseteq\Sigma^{n-1}\}.$$

Details are omitted. Note that $P\cup\Sigma^1\cup\ldots\Sigma^{n-1} = P\cup\Sigma^{<n}$, the set of all atoms considered at round n. □

It is easy to see that Prop. 13 still holds for the history-based semantics. Therefore:

Corollary 25 *Action model logic is more update expressive than communication pattern logic on the class of interpreted systems, in the history-based semantics.*

This story on history-based semantics could just as well have been told for sequences $\boldsymbol{R}_1,\ldots,\boldsymbol{R}_n$ of possibly different communication patterns, instead of for n iterations of a given communication pattern $\boldsymbol{R}$. We would then get models $M\dot{\odot}\boldsymbol{R}_1\dot{\odot}\ldots\dot{\odot}\boldsymbol{R}_n$ instead of models $M\dot{\odot}\boldsymbol{R}^n$, and we would get induced action models $M\otimes\boldsymbol{U}^1(\boldsymbol{R}_1)\otimes\cdots\otimes\boldsymbol{U}^n(\boldsymbol{R}_n)$, etcetera. However, in distributed computing it is common to consider arbitrary iteration of the same communication pattern (the mentioned oblivious model).

Although in such a generalization we can continue to view histories as sequences of communication graphs, it is important to realize that the same communication graph can then be the point of a different communication pattern, which may give their execution a different meaning. For example, recall $R^{ab}b = \{a,b\}\cup I$. Given $R^{ab}\in\boldsymbol{IS}$, agent b is uncertain whether a has received his message. But given $R^{ab}\in\{R^{ab}\}$, the singleton communication pattern, agent b knows that agent a has not received his message.

6 Conclusions and further research

We have shown that action model logic and communication pattern logic are incomparable in update expressivity on epistemic models, and that action model logic is more update expressive than communication pattern logic on interpreted systems. It seems promising to investigate communication patterns further, also on epistemic models that are not local (clearly, incomparability does not depend on that). Induced action models are exponentially larger than communication patterns. Communication patterns intuitively specify system dynamics that abstracts from message content. Results in temporal epistemics on synchronous and asynchronous computation should carry over to dynamic epistemics.

Acknowledgements We thank the TARK reviewers for their comments. This work was partially supported by Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), project grants IN108720 and IN108723. Diego A. Velázquez is the recipient of a fellowship from CONACyT.

References

[1] T. Ågotnes & Y.N. Wáng (2017): *Resolving distributed knowledge*. Artif. Intell. 252, pp. 1–21, doi:10.1016/j.artint.2017.07.002.

[2] A. Baltag, L.S. Moss & S. Solecki (1998): *The Logic of Public Announcements, Common Knowledge, and Private Suspicions*. In: Proc. of 7th TARK, pp. 43–56.

[3] A. Baltag & S. Smets (2020): *Learning What Others Know*. In: Proc. of 23rd LPAR, EPiC Series in Computing 73, pp. 90–119, doi:10.29007/plm4.

[4] J. van Benthem, J. van Eijck & B. Kooi (2006): *Logics of Communication and Change*. Information and Computation 204(11), pp. 1620–1662, doi:10.1016/j.ic.2006.04.006.

[5] P. Blackburn, M. de Rijke & Y. Venema (2001): *Modal Logic*. Cambridge University Press, doi:10.1017/CBO9781107050884.

[6] A. Castañeda, H. van Ditmarsch, D.A. Rosenblueth & D.A. Velázquez (2022): *Communication Pattern Logic: Epistemic and Topological Views*. CoRR abs/2207.00823, doi:10.48550/arXiv.2207.00823. To appear in *Journal of Philosophical Logic*.

[7] H. van Ditmarsch, W. van der Hoek, B. Kooi & L.B. Kuijer (2020): *Arrow Update Synthesis*. Information and Computation, p. 104544, doi:10.1016/j.ic.2020.104544.

[8] H. van Ditmarsch, J. Ruan & W. van der Hoek (2013): *Connecting Dynamic Epistemic and Temporal Epistemic Logics*. Logic journal of the IGPL 21(3), pp. 380–403, doi:10.1093/jigpal/jzr038.

[9] C. Dwork & Y. Moses (1990): *Knowledge and Common Knowledge in a Byzantine Environment: Crash Failures*. Inf. Comput. 88(2), pp. 156–186, doi:10.1016/0890-5401(90)90014-9.

[10] J. van Eijck, J. Ruan & T. Sadzik (2012): *Action emulation*. Synthese 185(1), pp. 131–151, doi:10.1007/s11229-012-0083-1.

[11] M. Herlihy, D. Kozlov & S. Rajsbaum (2013): *Distributed Computing Through Combinatorial Topology*. Morgan Kaufmann, doi:10.1016/C2011-0-07032-1.

[12] B. Kooi & B. Renne (2011): *Generalized Arrow Update Logic*. In: Proc. of 13th TARK, pp. 205–211, doi:10.1145/2000378.2000403.

[13] L. Lamport, R. Shostak & M. Pease (1982): *The Byzantine Generals Problem*. ACM Trans. Program. Lang. Syst. 4(3), pp. 382–401, doi:10.1145/357172.357176.

[14] Y. Moses & M.R. Tuttle (1988): *Programming Simultaneous Actions Using Common Knowledge*. Algorithmica 3, pp. 121–169, doi:10.1007/BF01762112.

[15] J.A. Plaza (1989): *Logics of Public Communications*. In: Proc. of the 4th ISMIS, Oak Ridge National Laboratory, pp. 201–216.

[16] F. Roelofsen (2007): *Distributed knowledge*. Journal of Applied Non-Classical Logics 17(2), pp. 255–273, doi:10.3166/jancl.17.255-273.

[17] D.A. Velázquez, A. Castañeda & D.A. Rosenblueth (2021): *Communication Pattern Models: an Extension of Action Models for Dynamic-Network Distributed Systems*. In: Proc. of TARK XVIII, EPTCS 335, pp. 307–321, doi:10.4204/EPTCS.335.29.

[18] Y.N. Wáng & T. Ågotnes (2015): *Relativized common knowledge for dynamic epistemic logic*. J. Appl. Log. 13(3), pp. 370–393, doi:10.1016/j.jal.2015.06.004.

Complete Conditional Type Structures (Extended Abstract)

Nicodemo De Vito
Department of Decision Sciences
Bocconi University
Milan, Italy
nicodemo.devito@unibocconi.it

Hierarchies of conditional beliefs (Battigalli and Siniscalchi 1999) play a central role for the epistemic analysis of solution concepts in sequential games. They are practically modelled by type structures, which allow the analyst to represent the players' hierarchies without specifying an infinite sequence of conditional beliefs. Here, we study type structures that satisfy a "richness" property, called *completeness*. This property is defined on the type structure alone, without explicit reference to hierarchies of beliefs or other type structures. We provide sufficient conditions under which a complete type structure represents all hierarchies of conditional beliefs. In particular, we present an extension of the main result in Friedenberg (2010) to type structures with conditional beliefs.

Keywords: Conditional Probability Systems, Hierarchies of Conditional Beliefs, Type Structures, Completeness, Terminality.

1 Introduction

Hierarchies of conditional beliefs (Battigalli and Siniscalchi 1999) play a central role for the epistemic analysis of solution concepts in sequential games. Conditional beliefs generalize ordinary probabilistic beliefs in that every player is endowed with a collection of conditioning events, and forms conditional beliefs given each hypothesis in a way that updating is satisfied whenever possible. Such a collection of measures is called conditional probability system (CPS, hereafter). A player's first-order conditional beliefs are described by a CPS over the space of primitive uncertainty; her second-order conditional beliefs are described by a CPS over the spaces of primitive uncertainty and of the co-players' first-order conditional beliefs; and so on.

Battigalli and Siniscalchi (1999) show that hierarchies of CPSs can be practically represented by *conditional type structures*, i.e., compact models which mimic Harsanyi's representation of hierarchies of probabilistic beliefs. Namely, for each agent there is a set of types. Each type is associated with a CPS over the set of primitive uncertainty and the set of the co-players' types. Such structure induces a(n infinite) hierarchy of CPSs for each type.

Here, we study conditional type structures that satisfy a "richness" property, called *completeness*. This property—which plays a crucial role for epistemic foundations[1] of some solution concepts—is defined on the type structure alone, without explicit reference to hierarchies of CPSs or other type structures. Loosely speaking, a type structure is complete if it induces all possible conditional beliefs about types.

We ask: When does a complete type structure represent all hierarchies of conditional beliefs? The main result of the paper (Theorem 1) can be briefly summarized as follows. Suppose that a (conditional) type structure is complete. Then:

[1] See Dekel and Siniscalchi (2015) for a survey.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 173–184, doi:10.4204/EPTCS.379.15

(i) if the structure is Souslin, then it is *finitely terminal*, i.e., it induces all finite order conditional beliefs;

(ii) if the structure is compact and continuous, then it is *terminal*, i.e., it induces all hierarchies of conditional beliefs.

Precise definitions are given in the main text. Here we point out that Theorem 1 is an extension of the main result in Friedenberg (2010) to conditional type structures. Specifically, Friedenberg studies complete type structures with beliefs represented by ordinary probabilities; her main result shows that (i) and (ii) are sufficient conditions for finite terminality and terminality, respectively. Friedenberg (2010, Section 5) leaves open the question whether her result still holds when beliefs are represented by CPSs: our result provides an affirmative answer.

To prove the main result of the paper (Theorem 1), we adopt an approach that is different from the one in Friedenberg (2010). Specifically, we provide a construction—based on the set-up in Heifetz (1993)—of the canonical space of hierarchies of CPSs which allows us to characterize the notion of (finite) terminality in a convenient way (Proposition 2). With this, the crucial step of the proof relies on Lemma 3, an "extension" result for CPSs whose proof makes use of a selection argument. The details are spelled out in the paper.[2]

2 Preliminaries

A measurable space is a pair (X,Σ_X), where X is a set and Σ_X is a σ-algebra, the elements of which are called **events**. Throughout this paper, when it is clear from the context which σ-algebra on X we are considering, we suppress reference to Σ_X and simply write X to denote a measurable space. Furthermore, given a function $f: X \to Y$ and a family $\mathscr{F}_Y$ of subsets of Y, we let

$$f^{-1}(\mathscr{F}_Y) := \left\{E \subseteq X : \exists F \in \mathscr{F}_Y, E = f^{-1}(F)\right\}.$$

So, if Y is a measurable space, then $f^{-1}(\Sigma_Y)$ is the σ-algebra on X generated by f.

We write $\Delta(X)$ for the set of probability measures on Σ_X. Fix measurable spaces X and Y. Given a measurable function $f: X \to Y$, we let $\mathscr{L}_f : \Delta(X) \to \Delta(Y)$ denote the pushforward-measure map induced by f; that is, for each $\mu \in \Delta(X)$, $\mathscr{L}_f(\mu)$ is the image measure of μ under f, and is defined by $\mathscr{L}_f(\mu)(E) := \mu\left(f^{-1}(E)\right)$ for every $E \in \Sigma_Y$.

If X is a topological space, we keep using Σ_X to denote the Borel σ-algebra on X. All the topological spaces considered in this paper are assumed to be metrizable. We consider any product, finite or countable, of metrizable spaces as a metrizable space with the product topology. Moreover, we endow each subset of a metrizable space with the subspace topology. A **Souslin** (resp. **Lusin**) **space** is a topological space that is the image of a complete, separable metric space under a continuous surjection (resp. bijection). Clearly, a Lusin space is also Souslin. Examples of Souslin (resp. Lusin) spaces include analytic (resp. Borel) subsets of a complete separable metric space. In particular, a Polish space (i.e., a topological space which is homeomorphic to a complete, separable metric space) is a Lusin space. Furthermore, if X is a Lusin space, then (X,Σ_X) is a **standard Borel** space, i.e., there is a Polish space Y such that (X,Σ_X) is isomorphic to (Y,Σ_Y). If X is a Souslin space, then (X,Σ_X) is an **analytic measurable** space, i.e., there is a Polish space Y and an analytic subset $A \subseteq Y$ such that (X,Σ_X) is isomorphic to (A,Σ_A); see Cohn (2013, Chapter 8).

For a metrizable space X, the set $\Delta(X)$ of (Borel) probability measures is endowed with the topology of weak convergence. With this topology, $\Delta(X)$ becomes a metrizable space.

[2] The paper can be found at https://arxiv.org/abs/2305.08940.

3 Conditional probability systems

We represent the players' beliefs as conditional probability systems (cf. Rényi 1955). Fix a measurable space (X,Σ_X). A family of **conditioning events** of X is a non-empty family $\mathscr{B}\subseteq\Sigma_X$ that does not include the empty set. A possible interpretation is that an individual is uncertain about the realization of the "state" $x\in X$, and $\mathscr{B}$ represents a family of observable events or "relevant hypotheses." If X is a metrizable space, then each conditioning event $B\in\mathscr{B}$ is a Borel subset of X. In this case, we say that $\mathscr{B}$ is **clopen** if each element $B\in\mathscr{B}$ is both closed and open. For instance, $\mathscr{B}$ is clopen if X is a (finite) set endowed with the discrete topology; if $\mathscr{B}=\{X\}$, then $\mathscr{B}$ is trivially clopen.

Definition 1 *Let (X,Σ_X) be a measurable space and $\mathscr{B}\subseteq\Sigma_X$ be a family of conditioning events. A* ***conditional probability system (CPS)*** *on $(X,\Sigma_X,\mathscr{B})$ is an array of probability measures $\mu:=(\mu(\cdot|B))_{B\in\mathscr{B}}$ such that:*
(i) for all $B\in\mathscr{B}$, $\mu(B|B)=1$;
(ii) for all $A\in\Sigma_X$ and $B,C\in\mathscr{B}$, if $A\subseteq B\subseteq C$ then $\mu(A|B)\,\mu(B|C)=\mu(A|C)$.

Definition 1 says that a CPS μ is an element of the set $\Delta(X)^{\mathscr{B}}$, i.e., μ is a function from $\mathscr{B}$ to $\Delta(X)$.[3] We write $\mu(\cdot|B)$ to stress the interpretation as a conditional probability given event $B\in\mathscr{B}$. Condition (ii) is the **chain rule** of conditional probabilities and it can be written as follows: if $A\subseteq B\subseteq C$, then

$$\mu(B|C)>0\Rightarrow\mu(A|B)=\frac{\mu(A|C)}{\mu(B|C)}.$$

We let $\Delta^{\mathscr{B}}(X)$ denote the set of CPSs on $(X,\Sigma_X,\mathscr{B})$. The following result (whose proof can be found in Appendix A) records some topological properties of $\Delta^{\mathscr{B}}(X)$ when X is a metrizable space and $\mathscr{B}$ is countable.[4]

Lemma 1 *Fix a metrizable space X and a countable family $\mathscr{B}\subseteq\Sigma_X$ of conditioning events.*
(i) The space $\Delta^{\mathscr{B}}(X)$ is metrizable.
(ii) If X is Souslin or Lusin, so is $\Delta^{\mathscr{B}}(X)$.
(iii) Suppose that $\mathscr{B}$ is clopen. Then $\Delta^{\mathscr{B}}(X)$ is compact if and only if X is compact.

Note that if X is a Polish space, then $\Delta^{\mathscr{B}}(X)$ may fail to be Polish. But, by Lemma 1.(ii), $\Delta^{\mathscr{B}}(X)$ is a Lusin space. We can conclude that $\Delta^{\mathscr{B}}(X)$ is a Polish space provided that $\mathscr{B}$ is clopen (cf. Battigalli and Siniscalchi 1999, Lemma 1).

Fix measurable spaces (X,Σ_X) and (Y,Σ_Y), and families $\mathscr{B}_X\subseteq\Sigma_X$ and $\mathscr{B}_Y\subseteq\Sigma_Y$ of conditioning events. Suppose that $f:X\to Y$ is a measurable function such that

$$f^{-1}(\mathscr{B}_Y)=\mathscr{B}_X.$$

The function $\overline{\mathscr{L}}_f:\Delta^{\mathscr{B}_X}(X)\to\Delta^{\mathscr{B}_Y}(Y)$ defined by

$$\overline{\mathscr{L}}_f(\mu)(E|B):=\mu\left(f^{-1}(E)\,|f^{-1}(B)\right),$$

where $E\in\Sigma_Y$ and $B\in\mathscr{B}_Y$, is the **pushforward-CPS map** induced by f. Note that, for any $\mu\in\Delta^{\mathscr{B}_X}(X)$, we can write $\overline{\mathscr{L}}_f(\mu)$ as

$$\overline{\mathscr{L}}_f(\mu)=\left(\mathscr{L}_f\left(\mu\left(\cdot|f^{-1}(B)\right)\right)\right)_{B\in\mathscr{B}_Y},$$

[3] For every pair of sets X and Y, we let Y^X denote the set of functions with domain X and codomain Y.

[4] Lemma 1 is a generalization of analogous results (for the case when X is Polish) in Battigalli and Siniscalchi (1999, Lemma 1).

where $\mathscr{L}_f : \Delta(X) \to \Delta(Y)$ is the pushforward-measure map induced by f.

We record some basic results on the pushforward-CPS map that will be used extensively throughout the paper. In particular, Lemma 2.(i) ensures that $\overline{\mathscr{L}}_f$ is well-defined and justifies the terminology: if $\mu \in \Delta^{\mathscr{B}_X}(X)$, then $\overline{\mathscr{L}}_f(\mu)$ is a CPS on $(Y, \Sigma_Y, \mathscr{B}_Y)$.

Lemma 2 *Fix measurable spaces (X, Σ_X) and (Y, Σ_Y), and families $\mathscr{B}_X \subseteq \Sigma_X$ and $\mathscr{B}_Y \subseteq \Sigma_Y$ of conditioning events. Suppose that $f : X \to Y$ is a measurable function such that $f^{-1}(\mathscr{B}_Y) = \mathscr{B}_X$. The following statements hold.*
(i) The map $\overline{\mathscr{L}}_f : \Delta^{\mathscr{B}_X}(X) \to \Delta^{\mathscr{B}_Y}(Y)$ is well-defined.
(ii) Suppose that $\mathscr{B}_X$ and $\mathscr{B}_Y$ are countable, X is a metrizable space and Y is a Souslin space. If f is Borel measurable (resp. continuous), then $\overline{\mathscr{L}}_f$ is Borel measurable (resp. continuous).

A special case of image CPS induced by a function is of particular interest—namely, the marginalization of a CPS on a product space. Consider measurable spaces X and Y, and denote by π_X the coordinate projection from $X \times Y$ onto X. Fix a family $\mathscr{B} \subseteq \Sigma_X$ of conditioning events, and define $\mathscr{B}_{X\times Y}$ as

$$\mathscr{B}_{X\times Y} := (\pi_X)^{-1}(\mathscr{B}) = \{C \subseteq X \times Y : \exists B \in \mathscr{B}, C = B \times Y\}, \tag{3.1}$$

that is, $\mathscr{B}_{X\times Y}$ is the set of all cylinders $B \times Y$ with $B \in \mathscr{B}$. The function $\overline{\mathscr{L}}_{\pi_X} : \Delta^{\mathscr{B}_{X\times Y}}(X \times Y) \to \Delta^{\mathscr{B}}(X)$ defined by

$$\overline{\mathscr{L}}_{\pi_X}(\mu) := \left(\mathscr{L}_{\pi_X}(\mu(\cdot|B))\right)_{B\in\mathscr{B}}$$

is called **marginal-CPS map**, and $\overline{\mathscr{L}}_{\pi_X}(\mu)$ is called the **marginal** on X of $\mu \in \Delta^{\mathscr{B}_{X\times Y}}(X \times Y)$.

4 Type structures and hierarchies of conditional beliefs

Throughout, we fix a two-player set I;[5] given a player $i \in I$, we let j denote the other player in I. We assume that both players share a common measurable space (S, Σ_S), called **space of primitive uncertainty**. For each $i \in I$, there is a family $\mathscr{B}_i \subseteq \Sigma_S$ of conditioning events. One interpretation (which is borrowed from Battigalli and De Vito 2021) is the following: S is a product set, viz. $S := \times_{i\in I} S_i$, and each element $s := (s_i)_{i\in I}$ is an objective description of players' behavior in a game with complete information and without chance moves—technically, $(s_i)_{i\in I}$ is a strategy profile. Each player is uncertain about the "true" behavior $s \in S$, including his own. If the game has sequential moves, then each $\mathscr{B}_i$ is a collection of *observable events*; that is, each $B \in \mathscr{B}_i$ is the set of strategy profiles inducing an information set of player i. Other interpretations of S and $(\mathscr{B}_i)_{i\in I}$ are also possible; a more thorough discussion can be found in Battigalli and Siniscalchi (1999, pp. 191-192). The results in this paper do not hinge on a specific interpretation.

From now on, we maintain the following technical assumptions on S and $(\mathscr{B}_i)_{i\in I}$:

- S is a Souslin space, and
- $\mathscr{B}_i \subseteq \Sigma_S$ is countable for every $i \in I$.

Following Battigalli and Siniscalchi (1999), we adopt the following notational convention.

Convention 1. Given a product space $X \times Y$ and a family $\mathscr{B} \subseteq \Sigma_X$ of conditioning events of X, the family of conditioning events of $X \times Y$ is $\mathscr{B}_{X\times Y}$ as defined in (3.1). Accordingly, we let $\Delta^{\mathscr{B}}(X \times Y)$ denote the set of CPSs on $(X \times Y, \Sigma_{X\times Y}, \mathscr{B}_{X\times Y})$.

[5] The assumption of a two-player set is merely for notational convenience. The analysis can be equivalently carried out with any finite set I with cardinality greater than two.

4.1 Type structures

We use the framework of type structures (or "type spaces") to model players' hierarchies of conditional beliefs. We adopt the following definition of type structure (cf. Battigalli and Siniscalchi 1999).

Definition 2 *An* $\left(S,(\mathscr{B}_i)_{i\in I}\right)$*-**based type structure** is a tuple*

$$\mathscr{T} := \left(S,(\mathscr{B}_i,T_i,\beta_i)_{i\in I}\right)$$

such that, for every $i \in I$,
*(i) the **type set*** T_i *is a metrizable space;*
*(ii) the **belief map*** $\beta_i : T_i \to \Delta^{\mathscr{B}_i}(S\times T_j)$ *is Borel measurable.*
Each element of T_i*, viz.* t_i*, is called (player i's)* ***type****.*

Definition 2 says that, for any $i \in I$, T_i represents the set of player i's possible "ways to think." Each type $t_i \in T_i$ is associated with a CPS on the set of primitive uncertainty as well as on the possible "ways to think" (types) of player j. Each conditioning event for $\beta_i(t_i)$ has the form $B \times T_j$ with $B \in \mathscr{B}_i$.

If $\mathscr{B}_i = \{S\}$ for every player $i \in I$, then each set $\Delta^{\mathscr{B}_i}(S\times T_j)$ can be naturally identified with $\Delta(S\times T_j)$. In this case, Definition 2 coincides essentially with the definition in Friedenberg (2010),[6] and we say that $\mathscr{T}$ is an **ordinary type structure**. Moreover, we will sometimes refer to type structures via Definition 2 as **conditional type structures**.

Definition 3 *An* $\left(S,(\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T} := \left(S,(\mathscr{B}_i,T_i,\beta_i)_{i\in I}\right)$ *is*
(i) ***Souslin*** *(resp.* ***Lusin****,* ***compact****) if, for every* $i \in I$*, the type set* T_i *is a Souslin (resp. Lusin, compact) space;*[7]
(ii) ***continuous*** *if, for every* $i \in I$*, the belief map* β_i *is continuous.*

Next, we introduce the notion of completeness for a type structure.

Definition 4 *An* $\left(S,(\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T} := \left(S,(\mathscr{B}_i,T_i,\beta_i)_{i\in I}\right)$ *is* ***complete*** *if, for every* $i \in I$*, the belief map* β_i *is surjective.*

In words, completeness says that, for each player i, and for each conditional belief $\mu \in \Delta^{\mathscr{B}_i}(S\times T_j)$ that player i can hold, there is a type of player i which induces that belief. Thus, it is a "richness" requirement which may not be satisfied by some type structures. For instance, suppose that S is not a singleton. Then a type structure where the type set of some player has finite cardinality is not complete.

A type structure provides an implicit representation of the hierarchies of beliefs. To address the question whether a complete type structure represents all hierarchies of beliefs, we need to formally clarify *how* type structures generate a collection of hierarchies of beliefs for each player. This is illustrated in the following section.

4.2 The canonical space of hierarchies

In this section we first offer a construction of the set of all hierarchies of conditional beliefs satisfying a *coherence* condition. Loosely speaking, coherence means that lower-order beliefs are the marginals of

[6] The only difference is that S is assumed to be a Polish space in Friedenberg (2010). Such difference is immaterial for the remainder of the analysis.

[7] In Friedenberg (2010), a(n ordinary) type structure is called analytic if each type set is an analytic subset of a Polish space—hence, a metrizable Souslin (sub)space. We adopt the definition of Souslin type structure because we can extend our analysis (as we do in the Supplementary Appendix of the paper) without assuming metrizability of the topological spaces.

higher-order beliefs. The construction—which is based on the set-up in Heifetz (1993)—shows that this set of hierarchies identifies in a natural way a type structure, which we call it "canonical." Next, we show how each profile of types in a type structure can be associated with an element of the constructed set of hierarchies. This part is standard (cf. Heifetz and Samet 1998).

4.2.1 From hierarchies to types

To construct the set of hierarchies of conditional beliefs, we define recursively, for each player, two sequences of sets as well as a sequence of conditioning events. The first sequence, $(\Theta_i^n)_{n\geq 0}$, represents player i's $(n+1)$-order domain of uncertainty, for each $n\geq 0$. The second sequence, $(H_i^n)_{n\geq 1}$, consists of player i's n-tuples of *coherent* conditional beliefs over these space. The notion of coherence, formally defined below, says that, conditional on any relevant hypothesis, beliefs at different order do not contradict one another.

Formally, for each player $i\in I$, let

$$\begin{aligned} \Theta_i^0 &: = S, \\ \mathscr{B}_i^0 &: = \mathscr{B}_i, \\ H_i^1 &: = \Delta^{\mathscr{B}_i^0}\left(\Theta_i^0\right). \end{aligned}$$

The set Θ_i^0 is player i's 1-order (primitive) domain of uncertainty, and a first-order belief, viz. μ_i^1, is an element of the set H_i^1.

For $n\geq 1$, assume that $(\Theta_i^m)_{m=0,\ldots,n-1}$, $(\mathscr{B}_i^m)_{m=0,\ldots,n-1}$ and $(H_i^m)_{m=1,\ldots,n}$ have been defined for each player $i\in I$. Then, for each $i\in I$, let

$$\Theta_i^n := \Theta_i^0\times H_j^n.$$

That is, Θ_i^n is player i's $(n+1)$-order domain of uncertainty: it consists of the space of primitive uncertainty and what player $j\neq i$ believes about the space of primitive uncertainty, what player j believes about what player i believes about the space of primitive uncertainty,..., and so on, up to level n. For each $i\in I$ and $n\geq 1$, let $\pi_i^{n,n+1}: H_i^{n+1}\to H_i^n$ and $\rho_i^{n-1,n}:\Theta_i^n\to\Theta_i^{n-1}$ denote the coordinate projections. By construction, these maps satisfy the following property:

$$\forall i\in I,\forall n\geq 2,\rho_i^{n-1,n}=\left(\mathrm{Id}_{\Theta_i^0},\pi_j^{n-1,n}\right),$$

where $\mathrm{Id}_{\Theta_i^0}$ is the identity on Θ_i^0.

To define players' conditional beliefs on the $(n+1)$-th order domain of uncertainty, for each player $i\in I$, let

$$\begin{aligned} \mathscr{B}_i^n &: = \left(\rho_i^{n-1,n}\right)^{-1}\left(\mathscr{B}_i^{n-1}\right)=\left\{C\subseteq\Theta_i^n:\exists B\in\mathscr{B}_i^{n-1},C=\left(\rho_i^{n-1,n}\right)^{-1}(B)\right\}, \\ H_i^{n+1} &: = \left\{\left(\left(\mu_i^1,\ldots,\mu_i^n\right),\mu_i^{n+1}\right)\in H_i^n\times\Delta^{\mathscr{B}_i^n}\left(\Theta_i^n\right):\overline{\mathscr{L}}_{\rho_i^{n-1,n}}\left(\mu_i^{n+1}\right)=\mu_i^n\right\}. \end{aligned}$$

Specifically, $\mathscr{B}_i^n$ represents the set of relevant hypotheses upon which player i's $(n+1)$-th order conditional beliefs are defined. That is, $\mu_i^{n+1}\in\Delta^{\mathscr{B}_i^n}\left(\Theta_i^n\right)$ is player i's $(n+1)$-th order CPS with $\mu_i^{n+1}(\cdot|B)\in\Delta\left(\Theta_i^n\right)$, $B\in\mathscr{B}_i^n$. Recursively, it can be checked that, for all $n\geq 1$,

$$\mathscr{B}_i^n=\left\{C\subseteq\Theta_i^n:\exists B\in\mathscr{B}_i,C=B\times H_j^n\right\},$$

i.e., $\mathscr{B}_i^n$ is a set of cylinders in Θ_i^n generated by $\mathscr{B}_i$. If $\mathscr{B}_i$ is clopen, then every $B \in \mathscr{B}_i^n$ is clopen in Θ_i^n, since each coordinate projection $\rho_i^{n-1,n}$ is a continuous function. By definition of each Θ_i^n, we write, according to Convention 1,

$$\Delta^{\mathscr{B}_i^n}\left(\Theta_i^n\right) = \Delta^{\mathscr{B}_i}\left(\Theta_i^n\right).$$

The set H_i^{n+1} is the set of player i's $(n+1)$-tuples of CPSs on Θ_i^0, Θ_i^1,..., Θ_i^n. The condition on μ_i^{n+1} in the definition of H_i^{n+1} is the **coherence** condition mentioned above. Given the recursive construction of the sets, CPSs μ_i^{n+1} and μ_i^n both specify a (countable) array of conditional beliefs on the domain of uncertainty Θ_i^{n-1}, and those beliefs cannot be contradictory. Formally, for all $B \in \mathscr{B}_i^{n-1}$ and for every event $E \subseteq \Theta_i^{n-1}$,

$$\mu_i^{n+1}\left(\left(\rho_i^{n-1,n}\right)^{-1}(E)\middle|\left(\rho_i^{n-1,n}\right)^{-1}(B)\right) = \mu_i^n\left(E|B\right).$$

That is, the conditional belief $\mu_i^{n+1}(\cdot|(\rho_i^{n-1,n})^{-1}(B))$ must assign to event $\left(\rho_i^{n-1,n}\right)^{-1}(E)$ the same number as $\mu_i^n(\cdot|B)$ assigns to event E.

Remark 1 *For each $i \in I$ and $n \geq 1$, the set H_i^{n+1} is a closed subset of $H_i^n \times \Delta^{\mathscr{B}_i}(\Theta_i^n)$. So H_i^{n+1} is a Souslin (resp. Lusin) space provided S is a Souslin (resp. Lusin) space. If $\mathscr{B}_i$ is clopen for each $i \in I$, then H_i^{n+1} is compact if and only if S is compact.*

In the limit, for each $i \in I$, let

$$\begin{aligned} H_i &:= \left\{\left(\mu_i^1, \mu_i^2, ...\right) \in \times_{n=0}^{\infty} \Delta^{\mathscr{B}_i}\left(\Theta_i^n\right) : \forall n \geq 1, \left(\mu_i^1, ..., \mu_i^n\right) \in H_i^n\right\}, \\ \Theta_i &:= S \times H_j. \end{aligned}$$

Remark 2 *The set H_i is a closed subset of $\times_{n=0}^{\infty} \Delta^{\mathscr{B}_i}(\Theta_i^n)$. So H_i is a Souslin (resp. Lusin) space provided S is a Souslin (resp. Lusin) space. If $\mathscr{B}_i$ is clopen for each $i \in I$, then H_i is compact if and only if S is compact.*

The following result corresponds to Proposition 2 in Battigalli and Siniscalchi (1999).

Proposition 1 *For each $i \in I$, the spaces H_i and $\Delta^{\mathscr{B}_i}(S \times H_j)$ are homeomorphic.*

The set $H := \times_{i \in I} H_i$ is the set of all pairs of *collectively coherent* hierarchies of conditional beliefs; that is, H is the set of pairs of coherent hierarchies satisfying common full belief of coherence.[8]

The homeomorphisms in Proposition 1 are "canonical" in the following sense: every coherent hierarchy $\left(\mu_i^1, \mu_i^2, ...\right)$ of player i is associated with a unique CPS μ_i on the space of primitive uncertainty and the coherent hierarchies of the co-player, i.e., $S \times H_j$. Then, for all $n \geq 0$, the marginal of μ_i on player i's $(n+1)$-order domain of uncertainty, viz. Θ_i^n, is precisely what it should be, namely μ_i^{n+1}. A formal definition of such homeomorphisms is not needed for the statements and proofs of the results in this paper. Instead, we will make use of the following implication of Proposition 1: we can define an $\left(S, (\mathscr{B}_i)_{i \in I}\right)$-based type structure $\mathscr{T}^{\mathrm{c}} := \left(S, (\mathscr{B}_i, T_i^{\mathrm{c}}, \beta_i^{\mathrm{c}})_{i \in I}\right)$ by letting, for each $i \in I$,

$$T_i^{\mathrm{c}} := H_i,$$

[8] An event E is fully believed under a CPS $(\mu(\cdot|B))_{B \in \mathscr{B}}$ if $\mu(E|B) = 1$ for every $B \in \mathscr{B}$. The notion of "common full belief of coherence" is made explicit in the alternative construction of the canonical space *à la* Battigalli and Siniscalchi (1999). A note on terminology: in Battigalli and Siniscalchi (1999) the expression "certainty" is used in place of "full belief."

and $\beta_i^{\mathrm{c}} : T_i^{\mathrm{c}} \to \Delta^{\mathscr{B}_i}\left(S \times T_j^{\mathrm{c}}\right)$ is the "canonical" homeomorphism. Following the terminology in the literature, we call $\mathscr{T}^{\mathrm{c}}$ the **canonical type structure**.[9]

Remark 3 *Structure $\mathscr{T}^{\mathrm{c}}$ is Souslin, continuous and complete. If S is Lusin, then $\mathscr{T}^{\mathrm{c}}$ is Lusin. If $\mathscr{B}_i$ is clopen for each $i \in I$, then $\mathscr{T}^{\mathrm{c}}$ is a compact type structure if and only if S is a compact space.*

4.2.2 From types to hierarchies

The next step is to consider the relationship between the set of hierarchies constructed in the previous section and any other type structure. In so doing, we specify how types generate (collectively) coherent hierarchies of conditional beliefs. As we did in the previous section, given a set X, we let Id_X denote the identity map.

Fix an $\left(S, (\mathscr{B}_i)_{i\in I}\right)$-based type structure $\mathscr{T} := \left(S, (\mathscr{B}_i, T_i, \beta_i)_{i\in I}\right)$. We construct a natural (Borel) measurable map, called **hierarchy map**, which unfolds the higher-order beliefs of each player $i \in I$. This map assigns to each $t_i \in T_i$ a hierarchy of beliefs in H_i.

For each $i \in I$, let $h_{-i}^0 : \Theta_i^0 \times T_j \to \Theta_i^0$ be the projection map (recall that $\Theta_i^0 := S$ for each $i \in I$). The "first-order map" for each player i, viz. $h_i^1 : T_i \to H_i^1$, is defined by

$$h_i^1(t_i) := \overline{\mathscr{L}}_{h_{-i}^0}(\beta_i(t_i)).$$

In words, $h_i^1(t_i)$ is the marginal on S of CPS $\beta_i(t_i)$. Measurability of each map h_i^1 holds by Lemma 2 and measurability of belief maps.

With this, for each $i \in I$, let $h_{-i}^1 : \Theta_i^0 \times T_j \to \Theta_i^0 \times H_j^1 = \Theta_i^1$ be the map defined as $h_{-i}^1 := \left(\mathrm{Id}_S, h_j^1\right)$; i.e., for each pair $(s, t_j) \in \Theta_i^0 \times T_j$, the expression $h_{-i}^1(s, t_j) = \left(s, h_j^1(t_j)\right) \in \Theta_i^1$ describes the profile s and the first-order beliefs for type $t_j \in T_j$. Standard arguments show that, for each $i \in I$, the map h_{-i}^1 is measurable; furthermore, it can be checked that h_{-i}^1 satisfies

$$h_{-i}^0 = \rho_i^{0,1} \circ h_{-i}^1 \text{ and } \mathscr{B}_{\Theta_i^0 \times T_j} = \left(h_{-i}^1\right)^{-1}\left(\mathscr{B}_i^1\right).$$

Recursively, we define the "$(n+1)$th-orders" maps. For $n \geq 1$, assume that measurable maps $h_i^n : T_i \to H_i^n$ have been defined for each player $i \in I$. Moreover, for each $i \in I$, assume that $h_{-i}^n : \Theta_i^0 \times T_j \to \Theta_i^0 \times H_j^n = \Theta_i^n$ is the unique measurable function, defined as $h_{-i}^n := \left(\mathrm{Id}_S, h_j^n\right)$, which satisfies

$$\mathscr{B}_{\Theta_i^0 \times T_j} = \left(h_{-i}^n\right)^{-1}\left(\mathscr{B}_i^n\right) \tag{4.1}$$

and

$$h_{-i}^{n-1} = \rho_i^{n-1,n} \circ h_{-i}^n. \tag{4.2}$$

Fix a player $i \in I$. Note that, since (4.1) holds, $\overline{\mathscr{L}}_{h_{-i}^n} : \Delta^{\mathscr{B}_i}(S \times T_j) \to \Delta^{\mathscr{B}_i^n}(\Theta_i^n)$ is a well-defined measurable map by Lemma 2. With this, define $h_i^{n+1} : T_i \to H_i^n \times \Delta^{\mathscr{B}_i^n}(\Theta_i^n)$ by

$$h_i^{n+1}(t_i) := \left(h_i^n(t_i), \overline{\mathscr{L}}_{h_{-i}^n}(\beta_i(t_i))\right).$$

Using the same arguments as above, it is easily verified that the map h_i^{n+1} is measurable. By (4.2), it follows that:

[9] There are differences, in terms of technical assumptions, between our construction of $\mathscr{T}^{\mathrm{c}}$ and the one in Battigalli and Siniscalchi (1999). Actually, our construction is a generalization of theirs since we allow for weaker conditions on (*i*) the family of conditioning events, and (*ii*) the topological property of the primitive uncertainty space S.

Remark 4 $h_i^{n+1}(T_i) \subseteq H_i^{n+1}$ *for each* $i \in I$.

It is easily seen that, for each $t_i \in T_i$,

$$h_i^{n+1}(t_i) = \left(\overline{\mathscr{L}}_{h_{-i}^0}(\beta_i(t_i)), ..., \overline{\mathscr{L}}_{h_{-i}^{n-1}}(\beta_i(t_i)), \overline{\mathscr{L}}_{h_{-i}^n}(\beta_i(t_i))\right).$$

Finally, for each $i \in I$, the map $h_i : T_i \to \times_{n=0}^{\infty} \Delta^{\mathscr{B}_i}\left(S \times H_j^n\right)$ is defined by

$$h_i(t_i) := \left(\overline{\mathscr{L}}_{h_{-i}^n}(\beta_i(t_i))\right)_{n\geq 0}.$$

Thus, $h_i(t_i)$ is the hierarchy generated by type $t_i \in T_i$. Each type generates a (collectively) coherent hierarchy of beliefs, i.e., $h_i(T_i) \subseteq H_i$.

Remark 5 *For each* $i \in I$*, the map* $h_i : T_i \to H_i$ *is well-defined and Borel measurable. Furthermore, if* $\mathscr{T}$ *is continuous, then, for each* $i \in I$*, the map* h_i *is continuous.*

5 Terminal type structures

The following definitions are extensions to conditional type structures of the definitions put forward by Friedenberg (2010, Section 2) for ordinary type structures.

Definition 5 *An* $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T} := \left(S, (\mathscr{B}_i, T_i, \beta_i)_{i\in I}\right)$ *is* ***finitely terminal*** *if, for each type structure* $\mathscr{T}^* := \left(S, (\mathscr{B}_i, T_i^*, \beta_i^*)_{i\in I}\right)$*, each type* $t_i^* \in T_i^*$ *and each* $n \in \mathbb{N}$*, there is a type* $t_i \in T_i$ *such that* $h_i^{*,n}(t_i^*) = h_i^n(t_i)$.

Definition 6 *An* $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T} := \left(S, (\mathscr{B}_i, T_i, \beta_i)_{i\in I}\right)$ *is* ***terminal*** *if, for each type structure* $\mathscr{T}^* := \left(S, (\mathscr{B}_i, T_i^*, \beta_i^*)_{i\in I}\right)$ *and each type* $t_i^* \in T_i^*$*, there is a type* $t_i \in T_i$ *such that* $h_i^*(t_i^*) = h_i(t_i)$.

Definition 5 says that $\mathscr{T}$ is finitely terminal if, for every type t_i^* that occurs in some structure $\mathscr{T}^*$ and every $n \in \mathbb{N}$, there exists a type t_i in $\mathscr{T}$ whose hierarchy agrees with the hierarchy generated by t_i^* up to level n. Definition 6 says that $\mathscr{T}$ is terminal if, for every type t_i^* that occurs in some structure $\mathscr{T}^*$, there exists a type t_i in $\mathscr{T}$ which generates the same hierarchy as t_i^*.

The notion of terminality in Definition 6 can be equivalently expressed as follows: $\mathscr{T}$ is terminal if, for every structure $\mathscr{T}^*$, there exists a **hierarchy morphism** from $\mathscr{T}^*$ to $\mathscr{T}$, i.e., a map that preserves the hierarchies of beliefs. Here we show that (a) Definition 5 is equivalent to the requirement that a type structure generates all finite-order beliefs consistent with coherence and common full belief of coherence; and (b) Definition 6 is equivalent to the requirement that a type structure generates all collectively coherent hierarchies of beliefs.

Remark 6 *An* $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T}$ *is finitely terminal if and only if, for each* $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T}^*$*, each player* $i \in I$ *and each* $n \in \mathbb{N}$*,*

$$h_i^{*,n}(T_i^*) \subseteq h_i^n(T_i).$$

An $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T}$ *is terminal if and only if, for each* $\left(S, (\mathscr{B}_i)_{i\in I}\right)$*-based type structure* $\mathscr{T}^*$*, and for each player* $i \in I$*,*

$$h_i^*(T_i^*) \subseteq h_i(T_i).$$

The following result establishes the relationship between any (finitely) terminal type structure and the canonical space of hierarchies.

Proposition 2 *Fix an $\left(S,(\mathscr{B}_i)_{i\in I}\right)$-based type structure $\mathscr{T} := \left(S,(\mathscr{B}_i,T_i,\beta_i)_{i\in I}\right)$.*
(i) $\mathscr{T}$ is finitely terminal if and only if $h_i^n(T_i) = H_i^n$ for each $i\in I$ and each $n\in\mathbb{N}$.
(ii) $\mathscr{T}$ is terminal if and only if $h_i(T_i) = H_i$ for each $i\in I$.

Proposition 2 provides a characterization of (finite) terminality which turns out to be useful for the proof of the main result. It is basically a version of Result 2.1 (and Proposition B1.(ii)) in Friedenberg (2010).

6 Main result

The main result of this paper is the following theorem.

Theorem 1 *Fix an $\left(S,(\mathscr{B}_i)_{i\in I}\right)$-based type structure $\mathscr{T} := \left(S,(\mathscr{B}_i,T_i,\beta_i)_{i\in I}\right)$.*
(i) If $\mathscr{T}$ is Souslin and complete, then $\mathscr{T}$ is finitely terminal.
(ii) If $\mathscr{T}$ is complete, compact and continuous, then $\mathscr{T}$ is terminal.

If $\mathscr{B}_i = \{S\}$ for every $i\in I$, then Theorem 1 corresponds to Theorem 3.1 in Friedenberg (2010). The proof of Theorem 1 relies on the following result, whose proof makes use of Von Neumann Selection Theorem.

Lemma 3 *Fix Souslin spaces X, Y and Z, and a countable family $\mathscr{B}\subseteq\Sigma_X$ of conditioning events. Let $f_1 : Y\to Z$ be Borel measurable, and define $f_2 : X\times Y\to X\times Z$ as $f_2 := (\mathrm{Id}_X, f_1)$. Then:*
(i) f_2 is Borel measurable, and the map $\overline{\mathscr{L}}_{f_2} : \Delta^{\mathscr{B}}(X\times Y)\to\Delta^{\mathscr{B}}(X\times Z)$ is well-defined;
(ii) if f_1 is surjective, then $\overline{\mathscr{L}}_{f_2}$ is surjective.

The proof of part (i) of Theorem 1 is by induction on $n\in\mathbb{N}$. The proof of the base step does not rely on the hypothesis that $\mathscr{T}$ is Souslin. Lemma 3 is used *only* in the inductive (and crucial) step. The proof of part (ii) of Theorem 1 uses the same arguments as in Friedenberg (2010).

Some comments on Theorem 1 are in order. First, complete type structures that are finitely terminal can be easily constructed. A simple example—which uses the ideas in Brandenburger et al. (2008, proof of Proposition 7.2)—is the following. For each $i\in I$, let T_i be the Baire space, i.e., the (non-compact) Polish space $\mathbb{N}^{\mathbb{N}}$. Every Souslin space is the image of $\mathbb{N}^{\mathbb{N}}$ under a continuous function.[10] Since $\Delta^{\mathscr{B}_i}(S\times T_j)$ is a Souslin space by Lemma 1, there exists a continuous surjection $\beta_i : T_i\to\Delta^{\mathscr{B}_i}(S\times T_j)$. These maps give us a Souslin and complete type structure $\mathscr{T}$ that is finitely terminal, but not necessarily terminal. A more complex example of a(n ordinary) complete, finitely terminal type structure which is *not* terminal can be found in Friedenberg and Keisler (2021, Section 6).

Second, we point out that complete, compact and continuous type structures may not exist. This is so because the structural hypothesis on the families of conditioning events $\mathscr{B}_i$ ($i\in I$) are quite weak—each element of $\mathscr{B}_i$ is a Borel subset of S. To elaborate, suppose that $\mathscr{T}$ is complete and compact. Completeness yields $\beta_i(T_i) = \Delta^{\mathscr{B}_i}(S\times T_j)$ for each player $i\in I$. If $\mathscr{T}$ were continuous, then compactness of T_i would imply compactness of $\Delta^{\mathscr{B}_i}(S\times T_j)$ as well, because the continuous image of a compact set is compact. But, in general, $\Delta^{\mathscr{B}_i}(S\times T_j)$ is not a compact space, even if the underlying space $S\times T_j$

[10] It is well-known that every non-empty Polish space is the image of $\mathbb{N}^{\mathbb{N}}$ under a continuous function. Using this result, it is easy to check—by inspection of definitions—that an analogous conclusion also holds for Souslin spaces (cf. Cohn 2003, Corollary 8.2.8).

is compact (cf. Lemma 1).[11] In other words, unless each family $\mathscr{B}_i$ ($i \in I$) satisfies some specific assumptions, there is no guarantee that a complete, compact and continuous type structure exists—in particular, complete and continuous structures may fail the compactness requirement.[12] If this is the case, Theorem 1.(ii) still holds, but *vacuously* because the antecedent of the conditional is false. An immediate implication of this fact is that the completeness test—as formalized by completeness, compactness and continuity—cannot be applied.

With this in mind, suppose now that each family $\mathscr{B}_i$ ($i \in I$) is clopen. If S is a compact space, then complete, compact and continuous type structures do exist. The canonical structure is a prominent example, but there are also complete, compact and continuous structures which can be distinct from the canonical one. For instance, one can take each T_i to be the Cantor space $\{0,1\}^{\mathbb{N}}$, a compact metrizable space. Lemma 1.(iii) yields that each $\Delta^{\mathscr{B}_i}(S \times T_j)$ is compact metrizable, so there exists a continuous surjection $\beta_i : T_i \to \Delta^{\mathscr{B}_i}(S \times T_j)$ (Aliprantis and Border 2006, Theorem 3.60). The resulting structure $\mathscr{T}$ is complete, compact and continuous, hence terminal.

Finally, note that if each $\mathscr{B}_i$ is clopen, then compactness of S is a necessary condition for the existence of a complete, compact and continuous structure $\mathscr{T}$. Indeed, continuity and surjectivity of the belief maps entail that each set $\Delta^{\mathscr{B}_i}(S \times T_j)$ is compact; by Lemma 1.(iii) and Tychonoff's theorem, S is a compact space.

Acknowledgements. I thank three anonymous reviewers from TARK 2023, Pierpaolo Battigalli and Emiliano Catonini for helpful feedback. Financial support from the Italian Ministry of Education, PRIN 2017, Grant Number 2017K8ANN4, is gratefully acknowledged.

References

[1] C. Aliprantis & K. Border (2006): *Infinite Dimensional Analysis*. Springer-Verlag, doi:10.1007/3-540-29587-9.

[2] P. Battigalli & M. Siniscalchi (1999): *Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games*. Journal of Economic Theory 88, pp. 188–230, doi:10.1006/jeth.1999.2555.

[3] P. Battigalli & N. De Vito (2021): *Beliefs, Plans, and Perceived Intentions in Dynamic Games*. Journal of Economic Theory 105:105283, doi:10.1016/j.jet.2021.105283.

[4] A. Brandenburger, A. Friedenberg & H.J. Keisler (2008): *Admissibility in Games*. Econometrica 76, pp. 307–352, doi:10.1111/j.1468-0262.2008.00835.x.

[5] D.L. Cohn (2013): *Measure Theory*. Birkhauser, doi:10.1007/978-1-4614-6956-8.

[6] E. Dekel & M. Siniscalchi (2015): *Epistemic Game Theory*. In P. Young & S. Zamir, editors: Handbook of Game Theory with Economic Applications, 4, North-Holland, pp. 619–702, doi:10.1016/B978-0-444-53766-9.00012-4.

[7] A. Friedenberg (2010): *When Do Type Structures Contain All Hierarchies of Beliefs?* Games and Economic Behavior 68, pp. 108–129, doi:10.1016/j.geb.2009.05.005.

[8] A. Friedenberg & H.J. Keisler (2021): *Iterated Dominance Revisited*. Economic Theory 68, pp. 377–421, doi:10.1007/s00199-020-01275-z.

[9] A. Heifetz (1993): *The Bayesian Formulation of Incomplete Infomation—the Non-Compact Case*. International Journal of Game Theory 21, pp. 329–338, doi:10.1007/BF01240148.

[11] Let X be a compact space. The set $\Delta^{\mathscr{B}}(X)$ may fail to be a closed subset of $\Delta(X)^{\mathscr{B}}$, which is a compact metrizable space. Yet, by Lemma 1.(ii), $\Delta^{\mathscr{B}}(X)$ is a Lusin subspace of $\Delta(X)^{\mathscr{B}}$. In particular, $\Delta^{\mathscr{B}}(X)$ is a Borel subset of $\Delta(X)^{\mathscr{B}}$.

[12] Analogously, complete and compact type structures may fail the continuity requirement.

[10] A. Heifetz & D. Samet (1998): *Topology-Free Typology of Beliefs*. Journal of Economic Theory 82, pp. 324–341, doi:10.1006/jeth.1998.2435.

[11] A. Renyi (1955): *On a New Axiomatic Theory of Probability*. Acta Mathematica Academiae Scientiarum Hungaricae 6, pp. 285–335, doi:10.1007/BF02024393.

Causal Kripke Models*

Yiwen Ding^{2} † Krishna Manoorkar^{2} ‡ Apostolos Tzimoulis^{2} Ruoding Wang^{2,3}† Xiaolong Wang^{1,2}†

Shandong University^{1} Vrije Universiteit, Amsterdam^{2} Xiamen University^{3}

This work extends Halpern and Pearl's causal models for actual causality to a possible world semantics environment. Using this framework we introduce a logic of actual causality with modal operators, which allows for reasoning about causality in scenarios involving multiple possibilities, temporality, knowledge and uncertainty. We illustrate this with a number of examples, and conclude by discussing some future directions for research.

1 Introduction

Causality is crucial in human reasoning and knowledge. Defining and formalizing causality has been a significant area of research in philosophy and formal methods [12, 21, 24, 11]. In recent years, with the rise of machine learning and AI, there has been growing interest in formalizing causal reasoning. One of the key areas of AI research is designing algorithms capable of comprehending causal information and performing causal reasoning [5, 29, 30]. Causal reasoning can be instrumental in formally modeling notions such as responsibility, blame, harm, and explanation, which are important aspects in designing ethical and responsible AI systems [3].

In this article we focus on the kind of causality known as "actual causality" (a.k.a. token causality) [10, 20, 19, 31]. Actual causality refers to the causality of a specific event which has actually happened (e.g. "John died because Alice shot him") rather than general causes (e.g. "smoking causes cancer"). Several formal approaches have been used for modelling actual causality [24, 25, 13, 14]. One of the most prominent formalizations of actual causation was developed by Halpern and Pearl [28, 17, 18]. This model describes dependencies between *endogenous variables* and *exogenous variables* using *structural equations*. Based on causal models Halpern and Pearl have given three different definitions of actual causality known as *original, updated and modified* definitions [17, 18, 15] of actual causality using counterfactual reasoning. The formal language developed to describe actual causality in this model is used to define several notions like normality, blame, accountability and responsibility. This model has been used in several applications in law [27], database theory [26], model checking [8, 9, 4], and AI [22, 11, 3].

Notions like knowledge, temporality, possibility, normality (or typicality) and uncertainty play important role in causal reasoning and related applications. In the past, attempts have been made to incorporate some of these notions into the causal models of Halpern and Pearl. In [4], Beer et.al. define causality in linear temporal logic to explain counterexamples. This line of research has been carried forward in model checking and program verification [1, 23]. The Halpern and Pearl formalism has also been extended to define causality in frameworks such as transition systems and Hennessy-Milner logic [6, 7, 1].

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007627.

†Yiwen Ding, Ruoding Wang and Xiaolong Wang are supported by the China Scholarship Council

‡Krishna Manoorkar is supported by the NWO grant KIVI.2019.001 awarded to Alessandra Palmigiano.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 185–200, doi:10.4204/EPTCS.379.16

In [2], Barbero et.al. define causality with epistemic operators. However, to the best of our knowledge a general Kripke model for actual causality based on Halpern and Pearl framework has not been studied yet.

In this work, we develop the notion of causal Kripke models and introduce a modal language for causal reasoning with uncertainty, temporality, possibility, and epistemic knowledge. We show that our model can formalize notions like sufficient causality, blame, responsibility, normality, and explanations. Our framework provides a more natural definition of sufficient causality [16, Section 2.6] by considering nearby contexts, which Halpern's causal model does not support (c.f. 5.1). The developed causal Kripke models offer a straightforward way to describe nearby contexts and define sufficient causality as intended by Halpern. In order to stay as close as possible to Halpern's original framework, where formally only atomic events can be causes, we utilize a hybrid language not only contains modalities but also names for the possible worlds.

The structure of the paper is as follows. In Section 2, we provide preliminaries on causal models and logic of causality. In Section 3, we give several examples to motivate the development of causal Kripke semantics. In Section 4, we define causal Kripke models, and develop a modal logic of actual causality to reason about them. We generalize the Halpern-Pearl definitions of actual causality to this framework and provide a sound and complete axiomatization of the modal logic of actual causality. In Section 5, we model the examples discussed in Section 3 using our framework and also show how this model can be used to provide an intuitive definition of sufficient causality. Finally, in Section 6 we conclude and provide some directions for future research.

2 Preliminaries

2.1 Causal models

In this section we briefly recall key concepts and ideas of the standard logic of causal reasoning as presented in [16]. A causal model describes the world in terms of variables which take values over certain sets. The variables and their ranges are given by a *signature* $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$ where $\mathscr{U}$ is a finite set of *exogenous* variables (i.e., variables whose value is independent of other variables in the model), $\mathscr{V}$, which is disjoint with $\mathscr{U}$, is a finite set of *endogenous* variables (i.e., variables whose value is determined by other variables in the model), and $\mathscr{R}(X)$ for any $X \in \mathscr{U} \cup \mathscr{V}$, is the (finite) range of X. These variables may have dependencies between them described by *structural equations* defined as follows.

Definition 2.1. *A causal model is a pair $(\mathscr{S}, \mathscr{F})$, where $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$ is the model's signature and $\mathscr{F} = (f_{V_i} \mid V_i \in \mathscr{V})$ assigns to each endogenous variable V_i a map such that*

$$f_{V_i} : \mathscr{R}(\mathscr{U} \cup \mathscr{V} - \{V_i\}) \to \mathscr{R}(V_i).$$

Definition 2.2. *For any variables $V \in \mathscr{V}$ and $X \in \mathscr{U} \cup \mathscr{V}$, we say "X is a direct cause, or a parent, of V" if there exist $x, x' \in \mathscr{R}(X)$ and $\overline{z} \in \mathscr{R}(\mathscr{U} \cup \mathscr{V} - \{X, V\})$ such that $f_V(\overline{z}, x) \neq f_V(\overline{z}, x')$. A causal model is said to be* ***recursive*** *if it contains no cyclic dependencies.*

Definition 2.3. *For a causal model $M = (\mathscr{S}, \mathscr{F})$, a* ***context*** *$\overline{t}$ assigns every variable $U \in \mathscr{U}$ a value in $\mathscr{R}(U)$. A* ***causal setting*** *is a pair $(M, \overline{t})$, where M is a causal model and $\overline{t}$ is a context for it.*

In recursive models, as there are no cyclic dependencies the values of all endogenous variables are determined by the context. Throughout this paper we only consider recursive causal models.

Definition 2.4. *Let $M = (\mathscr{S}, \mathscr{F})$ be some causal model and $\mathscr{Y} \subseteq \mathscr{V}$ be a set of endogenous variables. Let $\overline{Y}$ be the injective listing of the variables of $\mathscr{Y}$. Let $\overline{Y} = \overline{y}$ be an assignment such that $y_i \in \mathscr{R}(Y_i)$ for* *every $Y_i \in \mathscr{Y}$. The causal model obtained from* ***intervention*** *setting values of variables of $\overline{Y}$ to $\overline{y}$ is given by $M_{\overline{Y}\leftarrow\overline{y}} = (\mathscr{S}, \mathscr{F}_{\overline{Y}\leftarrow\overline{y}})$, where $\mathscr{F}_{\overline{Y}\leftarrow\overline{y}}$ is obtained by replacing for every variable $Y_i \in \mathscr{Y}$, the structural equation f_{Y_i} with $Y_i = y_i$.*

Here, we consider the exogenous variables as given. Thus, we do not allow interventions on them.

2.2 Basic language for describing causality

The basic language, L_C, for describing causality is an extension of propositional logic where *primitive events* are of the form $X = x$, where $X \in \mathscr{V}$ is an endogenous variable and $x \in \mathscr{R}(X)$. Given the signature $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$, the formulas $\phi \in \mathrm{L}_C$ are defined by the following recursion:

$$\alpha ::= X = x \mid \neg\alpha \mid \alpha \wedge \alpha \quad \text{where}, X \in \mathscr{V}, x \in \mathscr{R}(X)$$
$$\phi ::= X = x \mid \neg\phi \mid \phi \wedge \phi \mid [\overline{Y} \leftarrow \overline{y}]\alpha \quad \text{where}, \overline{Y} \leftarrow \overline{y} \text{is an intervention}$$

For any causal setting $(M,\overline{t})$ and formula $\phi \in \mathrm{L}_C$, the satisfaction relation $(M,\overline{t}) \Vdash \phi$ is defined as follows. For any formula $X = x$, $(M,\overline{t}) \Vdash X = x$ if the value of endogenous variable X is set to x in context $\overline{t}$. Satisfaction for the Boolean connectives is defined in a standard manner. Satisfaction of intervention formulas is defined as follows: for any event α, $(M,\overline{t}) \Vdash [\overline{Y} \leftarrow \overline{y}]\alpha$ iff $(M_{\overline{Y}\leftarrow\overline{y}},\overline{t}) \Vdash \alpha$. This language is used by Halpern and Pearl to provide three different definitions of causality referred as *original* , *updated* and *modified* definitions of causality [16, Section 2.2] (for details see Appendix A.1).

3 Motivation for possible world semantics of causal models

The basic language for causal reasoning described above uses propositional logic as the language of events and for reasoning with causal formulas. However, we are interested in describing causal reasoning in scenarios that involve notions like possibility, knowledge or belief, temporality, uncertainty and accessibility. Here we provide several such examples.

Example 3.1 (Umbrella)**.** *Alice is going on a trip to London. She thinks that it may rain when she is there. Thus, she decides to take her umbrella with her for the trip. In this example, the possibility of raining in London in the future seems to be the cause for Alice taking her umbrella with her.*

Example 3.2 (Chess)**.** *In a chess game, if knight and the king are only pieces that can move but every king move leads to king getting in check, then the player is forced to move the knight. Suppose that the king can not move to a certain square because it is covered by a bishop. Then it seems reasonable that the fact that bishop covers this square to be a cause for player being forced to move the knight. This example shows that reasoning with causality naturally involves considering possibilities.*

Example 3.3 (Police)**.** *Suppose John is a criminal who is currently absconding. Inspectors Alice and Bob are trying to catch John. John is currently in Amsterdam. He has a train ticket to Brussels. Thus, his (only) options are to stay in Amsterdam or to take the train to Brussels. Bob decides to go to Brussels to catch John in case he takes the train. John learns this information and decides to stay in Amsterdam where Alice catches him. In this case, John's belief of Bob's presence in Brussels leads to him staying in Amsterdam. It seems reasonable that this should be part of the cause of John getting caught, even though he was caught by Alice in Amsterdam. This shows that the knowledge of the agents is crucially involved in causal reasoning.*

Example 3.4 (Robot)**.** *Consider a scenario in which a robot is being commanded by a scientific team. Upon receiving command c, the robot completes task t or malfunctions. In this case the possibility of* *causing malfunction may become the cause of not sending command c. i.e. , causal reasoning involves scenarios in which the dependencies between different events may be ”indeterministic” or ”underdetermined”. Halpern considers such scenarios in [16, Section 2.5], using the notion of probabilities over causal models. However, in certain cases qualitative reasoning in terms of possibilities may be more appropriate.*

Example 3.5 (Navigation)**.** *Suppose Alice is trying to reach village A. She reaches a marker which indicates that she is at location B, C or D. She does not know in which of these locations she is at. However, she knows that A is to the east of all of these locations. Thus, she decides to go east. Suppose Alice was actually at point B. The fact that A is to the east of locations C and D is still part of Alice's considerations and seems to be part of the cause for her going east.*

These examples highlight that notions such as possibility, knowledge and uncertainty play an important role in causal reasoning. Possible world semantics, formally described by Kripke models, are the natural logical framework for modeling such notions. In the next section we develop a framework for causal reasoning, based on Kripke frames, which allows for modeling such scenarios in a clear, intuitive and efficient way.

4 Possible world semantics for causal reasoning

In this section, we define the causal Kripke model, introduce the modal language for causality and give the corresponding three HP definitions of causality in causal Kripke models. In our framework, we allow the same variable to possibly take different values in different worlds. Moreover, the structural equations treat the same endogenous variable separately for each different possible world.

Definition 4.1. *A **causal Kripke model** is a tuple $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$, where W is a finite set of possible worlds, $R \subseteq W \times W$ is an accessibility relation, and $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$ is the signature such that $\mathscr{U}$ and $\mathscr{V}$ are the disjoint sets of exogenous and endogenous variables, and $\mathscr{R}$ is a function assigning each $\Gamma \in \mathscr{U} \cup \mathscr{V}$ and a world $w \in W$ a set of possible values that Γ can take at w, and $\mathscr{F} = (f_{(X_i, w_j)} \mid X_i \in \mathscr{V}, w_j \in W)$ assigns to each endogenous variable X_i and each world w_j a map such that*

$$f_{(X_i,w_j)} : \mathscr{R}((\mathscr{U} \cup \mathscr{V}) \times W) - \{(X_i, w_j)\}) \to \mathscr{R}(X_i, w_j).$$

For any causal Kripke model $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$ we refer to $\mathscr{S}$ as its *signature*. For any variable Γ and world w we use (Γ, w) to denote the restriction of variable Γ to the world w. That is, (Γ, w) is a variable which takes a value c iff the propositional variable Γ takes the value c at the world w. For any $\Gamma \in \mathscr{U}$ (resp. $\Gamma \in \mathscr{V}$) and any world $w \in W$, we say (Γ, w) is an exogenous (resp. endogenous) variable. Note that we allow the same endogenous variable to have different structural equations associated with it in different worlds.

Definition 4.2. *A **context** over a causal Kripke model $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$ is a function $\bar{t}$ such that for any $w \in W$, and $U \in \mathscr{U}$, assigns a value in $\mathscr{R}(U, w)$. A **causal Kripke setting** is a pair $(\mathscr{K}, \bar{t})$, where $\mathscr{K}$ is a causal Kripke model and $\bar{t}$ is a context for it.*

Definition 4.3. *For any variables $X \in \mathscr{V}$, and $\Gamma \in \mathscr{U} \cup \mathscr{V}$, and any $w, w' \in W$, we say ”(X, w) is a direct cause, or a parent, of (Γ, w')” if there exist $\gamma, \gamma' \in \mathscr{R}(\Gamma, w')$, and $\bar{z} \in \mathscr{R}((\mathscr{U} \cup \mathscr{V}) \times W - \{(\Gamma, w')\})$ such that $f_{(X,w)}(\bar{z}, \gamma) \neq f_{(X,w)}(\bar{z}, \gamma')$. A causal Kripke model is said to be **recursive** if it contains no cyclic dependencies.*

In recursive models, as there are no cyclic dependencies the values of all endogenous variables at all the worlds are completely determined by the context. If $\mathscr{V}$ only contains binary variables (i.e. the variable which take values either 0 or 1), then for any context $\bar{t}$, and any world w, we use $\bar{t}(w)$ to denote the set of endogenous variables set to value 1 at w by $\bar{t}$. In this paper, we only consider recursive causal Kripke models.

Definition 4.4. *Given a causal Kripke model $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$, as* assignment *over $\mathscr{K}$ is a function on some subset $\mathscr{Y} \subseteq \mathscr{V} \times W$ such that, for every $Y = (X, w) \in \mathscr{Y}$, it assigns some value in $\mathscr{R}(X, w)$.*

Definition 4.5. *Let $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$ be some causal Kripke model and $\mathscr{Y}$ be a finite subset of $\mathscr{V} \times W$. Let $\overline{Y}$ be an injective (possibly empty) listing of all the variables in $\mathscr{Y}$. Let $\overline{Y} = \overline{y}$ be an assignment such that for any $Y_i \in \mathscr{Y}$, $y_i \in \mathscr{R}(Y_i)$. The causal Kripke model obtained from* ***intervention*** *setting values of variables of $\overline{Y}$ to $\overline{y}$ is given by $\mathscr{K}_{\overline{Y} \leftarrow \overline{y}} = (\mathscr{S}, W, R, \mathscr{F}_{\overline{Y} \leftarrow \overline{y}})$, where $\mathscr{F}_{\overline{Y} \leftarrow \overline{y}}$ is obtained by replacing for every variable $Y_i \in \mathscr{Y}$, the structural equation f_{Y_i} with $Y_i = y_i$.*

4.1 Modal logic language for describing causality

In this section we define the formal logical framework we introduce for describing causality. Since we want to talk about variables whose values depend on the possible world of a Kripke model, our language will be hybrid in character, augmenting the standard language (as presented e.g. in Section 2.2) not only with modal operators but also with a countable set of names, denoted with W. In principle each model M comes with an assignment from W to points of M, however in practice we will often conflate names with elements of Kripke models. The reason we require a countable number of names, even though the models are always finite, is because there is no bound on the size of the models. We will denote the language with $\mathrm{L}_M(W)$. We often omit W and write L_M when W is clear from the context. $\mathscr{S}$ is a given signature, and all X, Y, x, y come from $\mathscr{S}$. In what follows we will consistently use Y to denote a variable parametrized with a name for a world (i.e. $Y = (X, w)$). It is important to notice that in our language interventions involve only such variables. Any event α and formula ϕ of the language L_M is defined by the following recursion.

$$\alpha ::= X = x \mid (X, w) = x \mid \neg\alpha \mid \alpha \wedge \alpha \mid \Box\alpha \quad \text{where, } X \in \mathscr{V}, w \in W$$
$$\phi ::= X = x \mid (X, w) = x \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi \mid [\overline{Y} \leftarrow \overline{y}]\alpha \quad \text{where, } \overline{Y} \leftarrow \overline{y} \text{is an intervention}$$

In particular, the language L_M has two types of atomic propositions, using variables of the form X and of the form (X, w). The second, the hybrid aspect of our language, provides global information regarding the Kripke model. For any causal Kripke setting $(\mathscr{K}, \bar{t})$ with $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$, any causal formula ϕ, and any world $w \in W$, we define satisfaction relation $\Vdash$ in the following way. For any primitive event $X = x$ (resp. $(X, w') = x$), $(\mathscr{K}, \bar{t}, w) \Vdash X = x$ (resp. $(\mathscr{K}, \bar{t}, w) \Vdash (X, w') = x$) iff the value of X is set to be x at w (resp. at w') by the context $\bar{t}$. Note that the satisfaction of $(X, w') = x$ is independent of the world it is evaluated at. The satisfaction relation for Boolean connectives is defined by standard recursion. For the $\Box$ operator,

$$(\mathscr{K}, \bar{t}, w) \Vdash \Box\alpha \quad \text{iff} \quad \text{for all } w',\ wRw' \text{ implies } (\mathscr{K}, \bar{t}, w') \Vdash \alpha.$$

Let $\mathscr{Y} \subseteq \mathscr{V} \times W$ be a set of endogenous variables. Satisfaction of intervention formulas is defined as for any event α, $(\mathscr{K}, \bar{t}, w) \Vdash [\overline{Y} \leftarrow \overline{y}]\alpha$ iff $(\mathscr{K}_{[\overline{Y} \leftarrow \overline{y}]}, \bar{t}, w) \Vdash \alpha$. Satisfaction for Boolean combinations of causal formulas is defined in a standard manner. For the $\Box$ operator,

$$(\mathscr{K}, \bar{t}, w) \Vdash \Box\phi \quad \text{iff} \quad \text{for all } w',\ wRw' \text{ implies } (\mathscr{K}, \bar{t}, w') \Vdash \phi.$$

We now extend the HP definition(s) of causality to the setting of causal Kripke models.

Definition 4.6. *Let α be any event. For $\overline{Y} \subseteq \mathscr{V} \times W, \overline{Y} = \overline{y}$ is an* actual cause *of α in a causal Kripke setting $(\mathscr{K}, \bar{t})$ at a world w if the following conditions hold.*

AC1. $(\mathscr{K},\overline{t},w)\Vdash\alpha$ *and for every* $w_j\in W$, $(\mathscr{K},\overline{t},w_j)\Vdash(X_i,w_j)=y_{ij}$, *for every* $(X_i,w_j)=y_{ij}\in\overline{Y}=\overline{y}$.

AC2a. There exists a partition of $\mathscr{V}\times W$ *into two disjoint subsets* $\overline{Z}$ *and* $\overline{N}$ *with* $\overline{Y}\subseteq\overline{Z}$ *and settings* $\overline{y'}$ *and* $\overline{n}$ *of variables in* $\overline{Y}$ *and* $\overline{N}$, *such that*

$$(\mathscr{K},\overline{t},w)\Vdash[\overline{Y}\leftarrow\overline{y'},\overline{N}\leftarrow\overline{n}]\neg\alpha.$$

AC2b^o. Let $\overline{z}^*$ *be the unique setting of the variables in* $\overline{Z}$ *such that* $(\mathscr{K},\overline{t},w)\Vdash\overline{Z}=\overline{z}^*$. *If* $(\mathscr{K},\overline{t},w')\Vdash X=z^*$, *for every* $(X,w')=z^*\in\overline{Z}=\overline{z}^*$, *then for all subsets* $\overline{Z}'$ *of* $\overline{Z}\setminus\overline{Y}$ *we have*

$$(\mathscr{K},\overline{t},w)\Vdash[\overline{Y}\leftarrow\overline{y},\overline{N}\leftarrow\overline{n},\overline{Z}'\leftarrow\overline{z}'^*]\alpha.\ ^1$$

AC3. $\overline{Y}$ *is a minimal set of variables that satisfy AC1 and AC2.*

We say that $\overline{Y}=\overline{y}$ *is an* ***actual cause*** *of* α *in a causal Kripke setting* $(\mathscr{K},\overline{t})$ *at a world* w *by* ***updated definition*** *iff AC1, AC2a, AC3 hold and AC2b^o is replaced by the following condition.*

AC2b^u. Let $\overline{z}^*$ *be the unique setting of the variables in* $\overline{Z}$ *such that* $(\mathscr{K},\overline{t},w)\Vdash\overline{Z}=\overline{z}^*$. *If* $(\mathscr{K},\overline{t},w')\Vdash X=z^*$, *for every* $(X,w')=z^*\in\overline{Z}=\overline{z}^*$, *then for all subsets* $\overline{Z}'$ *of* $\overline{Z}\setminus\overline{Y}$ *and* $\overline{N}'$ *of* $\overline{N}$ *we have*

$$(\mathscr{K},\overline{t},w)\Vdash[\overline{Y}\leftarrow\overline{y},\overline{N}'\leftarrow\overline{n},\overline{Z}'\leftarrow\overline{z}'^*]\alpha.$$

We say that $\overline{Y}=\overline{y}$ *is an* ***actual cause*** *of* α *in a causal Kripke setting* $(\mathscr{K},\overline{t})$ *at a world* w *by* ***modified definition*** *iff AC1, AC3 hold and AC2 is replaced by the following condition.*

AC2a^m. If there exists a set of variables $\overline{N}\subseteq\mathscr{V}\times W$, *and a setting* $\overline{y}'$ *of the variables in* $\overline{Y}$ *such that if* $\overline{n}^*$ *is such that* $(\mathscr{K},\overline{t},w')\Vdash X=n^*$, *for every* $(X,w')=n^*\in\overline{N}=\overline{n}^*$, *then*

$$(\mathscr{K},\overline{t},w)\Vdash[\overline{Y}\leftarrow\overline{y}',\overline{N}\leftarrow\overline{n}^*]\neg\alpha.$$

We will refer to these definitions as original, updated, and modified definitions henceforth. Example B.1 shows that these definitions do not in general coincide. Theorem B.3 which relates these definitions in causal models can be generalized to causal Kripke models in a straightforward manner (see Theorem, B.5).

For any set of variables $\mathscr{Y}\subseteq\mathscr{V}\times W$, we use $cause^o(\overline{Y}=\overline{y},\alpha)$ (resp. $cause^u(\overline{Y}=\overline{y},\alpha)$, $cause^m(\overline{Y}=\overline{y},\alpha)$) as an abbreviation for stating $\overline{Y}=\overline{y}$ is a cause of α by the original (resp. updated, modified) definition. We write $(\mathscr{K},\overline{t},w)\Vdash cause^o(\overline{Y}=\overline{y},\alpha)$ (resp. $(\mathscr{K},\overline{t},w)\Vdash cause^u(\overline{Y}=\overline{y},\alpha)$, $(\mathscr{K},\overline{t},w)\Vdash cause^m(\overline{Y}=\overline{y},\alpha)$) as an abbreviation for stating $\overline{Y}=\overline{y}$ is a cause of α in causal Kripke setting $(\mathscr{K},\overline{t})$ at a world w by the original (resp. updated, modified) definition. Moreover, For $x=o,u,m$, we write $(\mathscr{K},\overline{t},w)\Vdash\Box cause^x(\overline{Y}=\overline{y},\alpha)$ if for all w' such that wRw', $(\mathscr{K},\overline{t},w')\Vdash cause^x(\overline{Y}=\overline{y},\alpha)$ and $(\mathscr{K},\overline{t},w)\Vdash\Diamond cause^x(\overline{Y}=\overline{y},\alpha)$ if there exists w' such that wRw' and $(\mathscr{K},\overline{t},w')\Vdash cause^x(\overline{Y}=\overline{y},\alpha)$.

4.2 Axiomatization

In [16], Halpern provides a sound and complete axiomatization for the logic of causality. This axiomatization can be extended to the modal logic of causality by adding the following axioms to the axiomatization in [16, Section 5.4]:

- All substitution instances of axioms of basic modal logic K.
- Necessitation rule: from ϕ infer $\Box\phi$
- $\Diamond$-axiom : $[\overline{Y}\leftarrow\overline{y}]\Diamond\phi\Leftrightarrow\Diamond[\overline{Y}\leftarrow\overline{y}]\phi$ and $\Box$-axiom : $[\overline{Y}\leftarrow\overline{y}]\Box\phi\Leftrightarrow\Box[\overline{Y}\leftarrow\overline{y}]\phi$
- G-axiom: $([\overline{Y}\leftarrow\overline{y}](X,w)=x)\Rightarrow\Box([\overline{Y}\leftarrow\overline{y}](X,w)=x)$

[1] Here we use the abuse of notation that if $\overline{Z}'\subseteq\overline{Z}$ and $\overline{Z}=\overline{z}^*$, then $\overline{z}'^*$ in $\overline{Z}'\leftarrow\overline{z}'^*$ refers to the restriction of $\overline{z}^*$ to $\overline{Z}'$.

Notice that, similar to the axiomatization in [16, Section 5.4], the schemes $\Diamond$-axiom, $\Box$-axiom, and G-axiom include empty interventions. When importing the axioms from [16, Section 5.4] axiom scheme C5 involves only variables of the form (X,w). For the axiom schemes C1-4 and C6, the axioms involve atoms both of the form $X = x$ and of the form $(X,w) = x$. Notice also that G-axiom is similar to axioms in Hybrid modal logic.

Since the language in [16] is finite (modulo classical tautologies), weak and strong completeness coincide. However our language is countable (given that W is countable). Since there is no upper bound on the size of the models, we cannot hope to have strong completeness w.r.t. finite models. However the axioms presented in this section are sound and weakly complete. In Appendix C we provide the proofs of **soundness** and weak **completeness** w.r.t. the modal logic of causality.

5 Examples and applications

In this section, we analyze the examples discussed in Section 3 using causal Kripke models. For any endogenous variable X, and any world w, we use $Eq(X,w)$ to denote the structural equation for X at w. Throughout this section, we use U to denote exogenous variables only.

Example 5.1 (Umbrella)**.** *Let $\mathscr{S}$ be the signature with endogenous variables p, q, and r standing for 'it rains in London' and 'Alice adds umbrella to the luggage' and 'Alice is in London'. Let w_0 be the current world and w_1, w_2, w_3 be the future possible worlds considered by Alice. Let $W = \{w_0, w_1, w_2, w_3\}$ and $R = \{(w_0,w_1),(w_0,w_2),(w_0,w_3)\}$. Let $U = (U_1,U_2) \in \{0,1\}^2$ be such that $(p,w) = (U_1,w)$ and $(r,w) = (U_2,w)$ for any $w \in W$. Let $Eq(q,w) = \Diamond(p \wedge r)$ for all w, i.e., Alice puts her umbrella in her luggage if she thinks it is possible that in the future she will be in London and it rains there.*

Let $\bar{t}$ be a context such that U is set to be $(0,0), (0,1)$, $(1,0)$ and $(1,1)$ at the worlds w_0, w_1, w_2 and w_3 respectively. We have $\bar{t}(w_0) = \{q\}$, $\bar{t}(w_1) = \{r\}$, $\bar{t}(w_2) = \{p\}$, $\bar{t}(w_3) = \{p,r\}$. Here, $(p,w_3) = 1$ and $(r,w_3) = 1$ are both causes of $q = 1$ at w_0 by all three definitions.

We show that $(p,w_3) = 1$ is a cause by the original and updated definitions. The proof for $(r,w_3) = 1$ is analogous. Indeed, as $(\mathscr{K},\bar{t},w_3) \Vdash (p,w_3) = 1$, $(\mathscr{K},\bar{t},w_3) \Vdash (r,w_3) = 1$ and $(\mathscr{K},\bar{t},w_0) \Vdash q$, AC1 is satisfied. Let $\overline{Z} = \{(r,w_3),(p,w_3)\}$ and $\overline{N} = \varnothing$, $\overline{Y} = \{(p,w_3)\}$, and $\overline{y'} = 0$. Then from the structural equation as no world related to w_0 satisfies $p \wedge r$ under this intervention, we have

$$(\mathscr{K},\bar{t},w_0) \Vdash [\overline{Y} \leftarrow 0]\neg(q = 1) \quad \textit{and} \quad (\mathscr{K},\bar{t},w_0) \Vdash [\overline{Y} \leftarrow 1, \overline{Z'} \leftarrow \overline{z}^*]q = 1.$$

where $\overline{Z'} = (r,w_3)$, $\overline{z}^* = (1,1)$ as described by the context. Thus, AC2 is satisfied and AC3 is trivial as we are considering a single variable. The updated definition in this case is equivalent to the original definition as $\overline{N} = \emptyset$. The modified definition is satisfied for the same setting $\overline{y'} = 0$ and $\overline{N} = \{(r,w_3)\}$. Thus, the fact that it rains in the world w_3 is a cause of Alice carrying her umbrella by all three definitions.

Now consider a slight variation of the above example in which we are sure Alice will be in the London and do not include r as a variable in our analysis. In this case, the structural equation for q is given by $q = \Diamond p$. Note that, in this new model, we can argue in the way similar to the above example that any world w' accessible from w_0, $(p,w') = 1$ would be a part of the cause of Alice carrying her umbrella at w_0 by all three definitions. This can be interpreted as the fact that 'Alice considers the possibility of a future world in which it rains in London and she will be in London' is a part of cause of her adding umbrella to luggage.

In general, for any event α and endogenous variable X we say that "the possibility of $X = x$" is a cause of α at w iff $\bigwedge\{(X,w') = x \mid wRw' \& (\mathscr{K},\bar{t},w') \Vdash X = x\}$ is a cause of α at w by the modified definition. Here, we use the modified definition because as mentioned by Halpern [16, Example 2.3.1], the conjunction being a cause by modified definition can in fact be interpreted as a cause being disjunctive, i.e. , the disjunction of the conjuncts can be interpreted as the cause of the event. Hence under this interpretation the existence of some w' which is accessible from w_0 and where $X = x$ is a cause of α here. This can be interpreted as "the possibility of $X = x$" being a cause of α. Thus, in the variation of the example discussed in the above paragraph, we can say that the possibility of the world where it rains in London is a cause of Alice adding umbrella to her luggage.

Example 5.2 (stalemate)**.** *Let $\mathscr{S}$ be the signature with endogenous variables p, q and r standing for 'The king is in check', 'The king and the knight are the only pieces that can move in the current position' and 'The player is forced to move the knight'. Let w_0 be the current position. Let w_1 and w_2 be the positions obtained from the (only) available moves by the king. Let $W = \{w_0, w_1, w_2\}$ and $R = \{(w_0, w_1), (w_0, w_2)\}$. Let $U = (U_1, U_2) \in \{0,1\}^2$ be such that $(p, w) = (U_1, w)$ and $(q, w) = (U_2, w)$ for any $w \in W$. Let $Eq(r, w) = \neg p \wedge q \wedge \Box p$ at any w, i.e., the player is forced to move the knight if the king and the knight are the only pieces that can move, the king is not in check, and the king's every available move leads to the king being in check.*

Let $\bar{t}$ be a context such that U is set to be $(0,1), (1,1)$, and $(1,0)$ at the worlds w_0, w_1, and w_2 respectively. We have $\bar{t}(w_0) = \{q, r\}$, $\bar{t}(w_1) = \{p, q\}$, $\bar{t}(w_2) = \{p\}$. In the same way as in the last example, we can show that $(p, w_0) = 0$, $(q, w_0) = 1$, $(p, w_1) = 1$ and $(p, w_2) = 1$ are all the causes of $r = 1$ at w_0 by all three definitions. This can intuitively be seen as the certainty of the king ending up in a check regardless of the king move (while not currently being in check) is a part of cause of being forced to move the knight.

In general, for any variable X, and any event α we say that the certainty of $X = x$ is a cause of α at w iff $(X, w') = x$ is a cause of α at w for all wRw' by the modified definition.

Example 5.3 (Police)**.** *Let $\mathscr{S}$ be the signature with endogenous variables p, q, r, and s standing for 'Inspector Bob is in Brussels', 'Inspector Alice is in Amsterdam', 'John takes the train', and 'John is caught in Amsterdam'. Let w_0 be the current world and w_1, w_2 be the possible future worlds considered by John. Let $W = \{w_0, w_1, w_2\}$ and $R = \{(w_0, w_1), (w_0, w_2)\}$.*

Let $U \in \{0,1\}^2$ be such that $(p, w) = (U_1, w)$ and $(q, w) = (U_2, w)$ for any $w \in W$. Let $Eq(r, w_0) = \neg\Box p$ and $Eq(r, w_1) = Eq(r, w_2) = 1$. i.e. , John takes the train if there is a possible future scenario in which inspector Bob is not in Brussels (John considers future scenarios when he takes the train). Let $Eq(s, w_0) = q \wedge \neg r$ (John gets caught in Amsterdam if Alice is there and he does not take the train) and $Eq(s, w_1) = Eq(s, w_2) = 0$ (John considers future scenarios in which he is not caught).

Let $\bar{t}$ be a context such that U is set to be $(1,1)$ at all the worlds. We have $\bar{t}(w_0) = \{p, q, s\}$ and $\bar{t}(w_1) = \bar{t}(w_2) = \{p, q, r\}$. It is easy to check that $(p, w_1) = 1$ and $(p, w_2) = 1$ are both causes of $s = 1$ at w_0 by all three definitions. Thus, the fact that Bob is present in Brussels in all possible worlds considered by John is a cause of John getting caught in Amsterdam. If we assume that John knows Bob is in Brussels iff he is actually in Brussels, then we can say that Bob's presence in Brussels is a cause of John getting caught in Amsterdam.

Here we have assumed that John's knowledge of Bob's presence in Brussels is the same as Bob being actually present in Brussels. However, this need not be the case always. Now consider a slightly more complicated version of the same story in which we have another endogenous variable o standing for 'John lost his ticket'. Suppose John does not take the train if he loses the ticket or he knows inspector Bob is in Brussels. i.e. ,$Eq(r, w_0) = \neg(\Box p \vee o)$. Other structural equations remain the same. Let $\bar{t}'$ be a context that sets variables p, q to the same values as $\bar{t}$ at all the worlds and sets o to be 1 at w_0. In this case $(p, w_1) = 1$, $(p, w_2) = 1$ and $(o, w_0) = 1$ are all the causes of $s = 1$ at w_0 by all three definitions. Now consider slightly different context $\bar{t}''$ such that p is set to be true at worlds w_0 and w_1, q at worlds

w_0, w_1, and w_2 and o at world w_0. In this case, we again have $r = 0$ and $s = 1$ at w_0. However, only $(o,w_0) = 1$ is a cause for $s = 1$ at w_0 by all three definitions. Now suppose that Bob actually did go to Brussels, however John does not ***know*** *this information and thinks that there is a possibility that Bob may not be in Brussels. The only reason he does not take the train is that he lost the ticket. Thus, Bob being present in Brussels is not a cause of John getting caught in Amsterdam in this case. This shows that the knowledge John has about the presence of Bob in Brussels (and not just presence itself) is an important part of causal reasoning.*

Example 5.4 (Robot)**.** *Let $\mathscr{S}$ be the signature with endogenous variables p, q and r standing for 'The command c is sent by the scientific team', 'The task t is completed by the robot' and 'The robot malfunctions'. Let w_0 be the current world, the world in which the scientific team is reasoning. Let w_1 and w_2 be the possible worlds considered by the team. Let $W = \{w_0,w_1,w_2\}$ and $R = \{(w_0,w_1),(w_0,w_2)\}$.*

Suppose $Eq(q,w_1) = \blacklozenge p$, $Eq(r,w_1) = 0$, $Eq(r,w_2) = \blacklozenge p$, $Eq(q,w_2) = 0$, and $Eq(q,w_0) = Eq(r,w_0) = 0$, where $\blacklozenge$ is the diamond operator corresponding to the relation R^{-1}, i.e., there are two possible scenarios. In one scenario sending command c leads to the completion of task t and no malfunctioning, while in the other it leads to the malfunctioning of the robot and task t is not completed.

Let $U \in \{0,1\}$ be such that $(p,w_i) = (U,w_i)$. Let $\bar{t}$ be a context such that U is set to be 1 in all the worlds. We have $\bar{t}(w_0) = \{p\}$, $\bar{t}(w_1) = \{p,q\}$, $\bar{t}(w_2) = \{p,r\}$. It is easy to see that $(p,w_0) = 1$ is the cause of $r = 1$ at w_2 but not at w_1 (it is not even true at w_1) by all three definitions. Suppose the scientific team believes that if sending command can possibly cause malfunctioning then command shouldn't be sent. Then as $(\mathscr{K},\bar{t},w_0) \Vdash \Diamond cause((p,w_0) = 1, r = 1)$ (for all three definitions), the team will decide not to send the command. On the other hand, if the team believes that the command should be sent if it can possibly cause the completion of the task, then it must be sent as $(\mathscr{K},\bar{t},w_0) \Vdash \Diamond cause((p,w_0) = 1, q = 1)$, i.e. sending signal can cause completion of task t.

Example 5.5 (Navigation)**.** *Let $\mathscr{S}$ be the signature with endogenous variables p_x, q and r standing for 'The current location of Alice is x' for $x = B,C,D$, 'Point A is to the east of Alice's current location' and 'Alice moves to the east'. Alice does not know if she is at point B, C or D. Let w_1, w_2 and w_3 be possible worlds and $U \in \{0,1\}^4$ be such that $(p_B,w) = (U_1,w)$, $(p_C,w) = (U_2,w)$, $(p_D,w) = (U_3,w)$ and $(q,w) = (U_4,w)$ for any $w \in W$.*

Let $\bar{t}$ be a context such that U is set to be $(1,0,0,1)$, $(0,1,0,1)$, and $(0,0,1,1)$ at worlds w_1, w_2 and w_3. We have $\bar{t}(w_1) = \{p_B,q,r\}$, $\bar{t}(w_2) = \{p_C,q,r\}$, and $\bar{t}(w_3) = \{p_D,q,r\}$. Worlds w_1, w_2 and w_3 here represent the possible scenarios considered by Alice. With the current available knowledge these worlds are indistinguishable from each other for Alice. This can be represented by $R = W \times W$. At any world w $Eq(r,w) = \Box q$, i.e. Alice moves to the East iff she ***knows*** *point A is to the East of her current location. Here, in the world w_1 in which Alice is at B (the real situation), $(q,w_1) = 1$, $(q,w_2) = 1$ and $(q,w_3) = 1$ are all causes of $(r,w_1) = 1$ by all three definitions. Thus, the fact that A is to the East of point C or point D is also a cause of Alice moving to East even if she is not present there.*

These examples show that causal Kripke models can be used to model several different scenarios involving causality interacting with notions like possibility, knowledge and uncertainty.

5.1 Sufficient causality

Halpern discusses the notion of sufficient causality in [16] to model the fact that people's reasoning about causality depends on how sensitive the causality ascription is to changes in various other factors. "The key intuition behind the definition of sufficient causality is that not only does $\overline{X} = \overline{x}$ suffice to bring about ϕ in the actual context, but it also brings it about in other "nearby" contexts. Since the framework does not provide a metric on contexts, there is no obvious way to define nearby context. Thus, in the formal definition below, I start by considering all contexts.”[16, Section 2.6] Sufficient causality is thus defined in [16] using Definition A.2.

We can use the framework of causal Kripke models to define sufficient causality for a causal setting (M,u) in terms of nearby contexts instead of all the contexts (as suggested by Halpern) in the following way. We consider the causal Kripke model $\mathscr{K} = (\mathscr{S}, W, R, \mathscr{F})$, where $\mathscr{S}$ is the signature of M, W is the set of all the possible contexts on M, and $\mathscr{F}$ is the set of structural equations such that for any structural equation for the endogenous variable X, $Eq(X,w)$ is the same as the structural equation for X in M and the relation $R \subseteq W \times W$ is such that uRu' iff context u' is nearby u. Let $\bar{t}$ be the setting of exogenous variables so that for any possible world the endogenous variables are set by the context identifying that world. Let $\overline{X} = \bar{x}$ be as in Definition A.2 and let $\overline{Y} = \overline{X} \times W$. For any $Y = (X,u)$, $Y = y$ iff X is set to be x by the context u. Let $\overline{Y} \leftarrow \bar{y}$ is intervention setting X to x in all the possible contexts. In this structure we can describe sufficient causality in terms of nearby contexts by replacing the clause $SC3$ in the Definition A.2 by the condition

$$uRu' \implies (\mathscr{K}, \bar{t}, u') \models [\overline{Y} \leftarrow \bar{y}]\alpha \quad \text{or equivalently} \quad (\mathscr{K}, \bar{t}, u) \models \Box[\overline{Y} \leftarrow \bar{y}]\alpha.$$

We call this property $SC3$-local as we only require that the intervention $[\overline{Y} \leftarrow \bar{y}]$ makes α true in nearby (not all) contexts.

In this Section, we have mainly only considered the causal Kripke model with only one relation which is the ”nearby” relation. However, we can also consider causal Kripke models with multiple relations in which we have a nearby relation N on the worlds in addition to the other accessibility relations denoting relationships between world like time, indistingushibility, etc. The definition of sufficient causality discussed above can naturally be extended to this setting allowing us to describe the sufficient causality in the setting of causal Kripke models. Here, we do not go into details of this generalization but we believe this would be an interesting direction for future research.

6 Conclusions and future directions

In this paper, we have developed a possible world semantics for reasoning about actual causality. We develop a modal language and logic to formally reason in this framework. This language is used to generalize the HP definitions of actual causality for this framework. We provide a sound and complete axiomatization of the modal logic of causality developed, and give a number of examples to illustrate how our model can be used to reason about causality. Finally, we show that our framework allows us to define the intended notion of sufficient causality in a straightforward and more intuitive manner.

This work can be extended in several directions. First, results regarding the relationship of the HP definitions with but-for causality [16, Proposition 2.2.2], and transitivity of cause [16, Section 2.4], can be generalized to our modal setting. Secondly, we can allow for interventions on the relation R in causal Kripke models. Indeed, in many scenarios intuitively the cause for some event is accessibility to some possible world. Allowing interventions on R would allow us to model such scenarios. Finally, similar to sufficient causality, other notions related to actual causality like normality (or typicality) and graded causation can be described in more nuanced and possibly multiple ways using the causal modal language.

References

[1] Christel Baier, Clemens Dubslaff, Florian Funke, Simon Jantsch, Rupak Majumdar, Jakob Piribauer & Robin Ziemek (2021): *From Verification to Causality-Based Explications*. In: *48th International Colloquium on*

Automata, Languages, and Programming (ICALP 2021), 198, Schloss Dagstuhl–Leibniz-Zentrum f {\" u} r Informatik, p. 1, doi:10.4230/LIPIcs.ICALP.2021.1.

[2] Fausto Barbero, Katrin Schulz, Fernando R Velázquez-Quesada & Kaibo Xie: *Observing interventions: a logic for thinking about experiments*. *Journal of Logic and Computation*, doi:10.1093/logcom/exac011.

[3] Sander Beckers (2022): *Causal explanations and XAI*. In: *Conference on Causal Learning and Reasoning*, PMLR, pp. 90–109, doi:10.48550/arXiv.2201.13169.

[4] Ilan Beer, Shoham Ben-David, Hana Chockler, Avigail Orni & Richard Trefler (2012): *Explaining counterexamples using causality*. *Formal Methods in System Design* 40, pp. 20–40, doi:10.1007/s10703-011-0132-2.

[5] Brian Bergstein (2020): *What AI still can't do*. *MIT Technol Rev* 123(2), pp. 1–7. Available at https://www.technologyreview.com/2020/02/19/868178/what-ai-still-cant-do/.

[6] Georgiana Caltais, Stefan Leue & Mohammad Reza Mousavi (2016): *(De-) Composing Causality in Labeled Transition Systems*. In: *The 1st Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies, Eindhoven, The Netherlands, April 8, 2016*, 224, Open Publishing Association, pp. 10–24, doi:10.4204/EPTCS.224.3.

[7] Georgiana Caltais, Mohammad Reza Mousavi & Hargurbir Singh (2020): *Causal reasoning for safety in Hennessy Milner logic*. *Fundamenta Informaticae* 173(2-3), pp. 217–251, doi:10.1007/s10849-022-09357-y.

[8] Hana Chockler, Joseph Y Halpern & Orna Kupferman (2008): *What causes a system to satisfy a specification?* *ACM Transactions on Computational Logic (TOCL)* 9(3), pp. 1–26, doi:10.1145/1352582.1352588.

[9] Anupam Datta, Deepak Garg, Dilsun Kaynar, Divya Sharma & Arunesh Sinha (2015): *Program actions as actual causes: A building block for accountability*. In: *2015 IEEE 28th Computer Security Foundations Symposium*, IEEE, pp. 261–275, doi:10.1109/CSF.2015.25.

[10] A Philip Dawid et al. (2007): *Fundamentals of statistical causality*. *RSS/EPSRC Grad Train Progr* 279, pp. 1–94. Available at https://www.semanticscholar.org/paper/FUNDAMENTALS-OF-STATISTICAL-CAUSALITY-Dawid/c4bcad0bb58091ecf9204ddb5db7dce749b0d461.

[11] Didier Dubois & Henri Prade (2020): *A glance at causality theories for artificial intelligence*. *A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning*, pp. 275–305, doi:10.1007/978-3-030-06164-7_9.

[12] Andrea Falcon (2006): *Aristotle on causality*. Available at https://plato.stanford.edu/entries/aristotle-causality/.

[13] Clark Glymour & Frank Wimberly (2007): *Actual causes and thought experiments*. *Causation and explanation* 4, p. 43. Available at https://www.semanticscholar.org/paper/Actual-Causes-and-Thought-Experiments-Glymour-Wimberly/d17f5a2896cf38b06e13713042e655e33753a69b.

[14] Ned Hall (2007): *Structural equations and causation*. *Philosophical Studies* 132, pp. 109–136, doi:10.1007/s11098-006-9057-9.

[15] Joseph Y Halpern (2015): *A modification of the Halpern-Pearl definition of causality*. *arXiv preprint arXiv:1505.00162*, doi:10.48550/arXiv.1505.00162.

[16] Joseph Y Halpern (2016): *Actual causality*. MiT Press, doi:10.7551/mitpress/10809.001.0001.

[17] Joseph Y Halpern & Judea Pearl (2005): *Causes and explanations: A structural-model approach. Part I: Causes*. *The British journal for the philosophy of science*, doi:10.1093/bjps/axi147.

[18] Joseph Y Halpern & Judea Pearl (2005): *Causes and explanations: A structural-model approach. Part II: Explanations*. *The British journal for the philosophy of science*, doi:10.1093/bjps/axi148.

[19] Christopher Hitchcock (2001): *A tale of two effects*. *The Philosophical Review* 110(3), pp. 361–396, doi:10.1215/00318108-110-3-361.

[20] Christopher Hitchcock (2013): *What is the 'cause'in causal decision theory?* Erkenntnis 78, pp. 129–146, doi:`10.1007/s10670-013-9440-9`.

[21] David Hume (1748): *An enquiry concerning human understanding and other writings*.

[22] Amjad Ibrahim, Tobias Klesel, Ehsan Zibaei, Severin Kacianka & Alexander Pretschner (2020): *Actual causality canvas: a general framework for explanation-based socio-technical constructs*. In: ECAI 2020, IOS Press, pp. 2978–2985, doi:`10.3233/FAIA200472`.

[23] Florian Leitner-Fischer & Stefan Leue (2013): *Causality Checking for Complex System Models*. In Roberto Giacobazzi, Josh Berdine & Isabella Mastroeni, editors: Verification, Model Checking, and Abstract Interpretation, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 248–267, doi:`10.1007/978-3-642-35873-9_16`.

[24] David Lewis (1973): *Causation*. The journal of philosophy 70(17), pp. 556–567, doi:`10.2307/2025310`.

[25] David Lewis (2004): *Causation as influence*. The Journal of Philosophy 97(4), pp. 182–197, doi:`10.7551/mitpress/1752.003.0004`.

[26] Alexandra Meliou, Wolfgang Gatterbauer, Joseph Y Halpern, Christoph Koch, Katherine F Moore & Dan Suciu (2010): *Causality in databases*. IEEE Data Engineering Bulletin 33(ARTICLE), pp. 59–67. Available at `https://www.cs.cornell.edu/home/halpern/papers/DE_Bulletin2010.pdf`.

[27] Michael S Moore (2009): *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press on Demand, doi:`10.1093/acprof:oso/9780199256860.001.0001`.

[28] Judea Pearl (2009): *Causality*. Cambridge university press, doi:`10.1017/CBO9780511803161`.

[29] Judea Pearl & Dana Mackenzie (2018): *The book of why: the new science of cause and effect*. Basic books, doi:`10.1080/01621459.2020.1721245`.

[30] Bernhard Schölkopf (2022): *Causality for machine learning*. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 765–804, doi:`10.1145/3501714.3501755`.

[31] James Woodward (2005): *Making things happen: A theory of causal explanation*. Oxford university press, doi:`10.1111/j.1933-1592.2007.00012.x`.

A HP definitions of causality and sufficient causality

Definition A.1 ([16], Definition 2.2.1)**.** *Let α be an event obtained by Boolean combination of primitive events. Let $\mathcal{X} \subseteq \mathcal{V}$ be a set of endogenous variables. $\overline{X} = \overline{x}$ is an* ***actual cause*** *of α in a causal setting $(M,\bar{t})$ if the following conditions hold.*

AC1. $(M,\bar{t}) \Vdash \overline{X} = \overline{x}$ *and* $(M,\bar{t}) \Vdash \alpha$.

AC2a. There is a partition of $\mathcal{V}$ into two disjoint subsets $\overline{Z}$ and $\mathcal{W}$ with $\overline{X} \subseteq \overline{Z}$ and a setting $\overline{x}'$ and $\overline{w}$ of variables in $\mathcal{X}$ and $\mathcal{W}$, respectively, such that

$$(M,\bar{t}) \Vdash_p [\overline{X} \leftarrow \overline{x}', \overline{W} \leftarrow \overline{w}]\neg\alpha.$$

AC2b^o. If $\overline{z}^*$ is such that $(M,\bar{t}) \Vdash \overline{Z} = \overline{z}^*$, then for all subsets $\overline{Z}'$ of $\overline{Z} \setminus \mathcal{X}$ we have

$$(M,\bar{t}) \Vdash_p [\overline{X} \leftarrow \overline{x}, \overline{W} \leftarrow \overline{w}, \overline{Z}' \leftarrow \overline{z}^*]\alpha.$$

AC3. $\mathcal{X}$ is minimal set of variable that satisfy AC1 and AC2.

We say that $\overline{X} = \overline{x}$ is an actual cause *of α in a causal setting $(M,\bar{t})$ by* updated definition *iff AC1, AC2a, AC3 hold and AC2b^o is replaced by the following condition.*

AC2b^u. If $\overline{z}^*$ is such that $(M,\overline{t}) \Vdash \overline{Z} = \overline{z}^*$, then for all subsets $\overline{Z}'$ of $\overline{Z} \setminus \mathscr{X}$ and $\mathscr{W}'$ of $\mathscr{W}$ we have

$$(M,\overline{t}) \Vdash_p [\overline{X} \leftarrow \overline{x}, \overline{W}' \leftarrow \overline{w}, \overline{Z}' \leftarrow \overline{z}^*]\alpha.$$

We say that $\overline{X} = \overline{x}$ is an actual cause *of α in a causal setting $(M,\overline{t})$ by* modified definition *iff AC1, AC3 hold and AC2 is replaced by the following condition.*

AC2a^m. If there exists a set of variables $\mathscr{W} \subseteq \mathscr{V}$, and a setting $\overline{x}'$ of variable in $\overline{X}$ such that if $(M,\overline{t}) \Vdash \overline{W} = \overline{w}^*$, then

$$(M,\overline{t}) \Vdash_p [\overline{X} \leftarrow \overline{x}', \overline{W} \leftarrow \overline{w}^*]\neg\alpha.$$

The following theorem describes relationship between these three definitions of causality.

Definition A.2 ([16], Definition 2.6.1)**.** *$\overline{X} = \overline{x}$ is a sufficient cause of α in the causal setting $(M,\overline{u})$ if the following conditions hold:*

SC1. $(M,\overline{u}) \models \overline{X} = \overline{x}$ and $(M,\overline{u}) \models \alpha$.

SC2. Some conjunct of $\overline{X} = \overline{x}$ is part of a cause of α in $(M,\overline{u})$. More precisely, there exists a conjunct $X = x$ of $\overline{X} = \overline{x}$ and another (possibly empty) conjunction $\overline{Y} = \overline{y}$ such that $X = x \wedge \overline{Y} = \overline{y}$ is a cause of α in $(M,\overline{u})$; i.e. , AC1, AC2, and AC3 hold for (possibly empty) conjunction $\overline{Y} = \overline{y}$ such that $X = x \wedge \overline{Y} = \overline{y}$

SC3. $(M,\overline{u}') \models [\overline{X} \leftarrow \overline{x}]\alpha$ for all contexts $\overline{u}'$.

- *$\overline{X}$ is the minimal set satisfying above properties.*

B Relationship between three HP definitions of causality

In [16] Halpern gives examples to show that the three HP definitions do not coincide with each other. Here we give one example to show that the modified definition may not coincide with the original and updated definition in the causal Kripke model. Similar example can be given to show that original and updated definition do not coincide.

Example B.1 (stalemate detailed)**.** *We consider the following variation of the example 5.2. Let $\mathscr{S}$ be signature with endogenous variables p_1 and p_2 instead of p (keeping the other endogenous variables unchanged) standing for 'The king is in check by the opponent's queen' and 'The king is in check by the opponent's king'. Let $U = (U_1, U_2, U_3) \in \{0,1\}^3$ be such that $(p_1, w) = (U_1, w)$, $(p_2, w) = (U_2, w)$, and $(q, w) = (U_3, w)$ for any $w \in W$. Let the structural equation for r at w_0 be given by $r = \neg(p_1 \vee p_2) \wedge q \wedge \Box(p_1 \vee p_2)$. i.e. , the player is forced to move the knoight if if the only pieces that can move are the king and the knight, ther king is not in check and every possible king move leads to king being in the check by the king or the queen (We assume there are no other pieces on the board) of the opponenet . Let $\overline{t}$ be a context such that U is set to be $(0,0,1)$,$(1,1,1)$, and $(0,1,0)$ at the worlds w_0, w_1, and w_2 respectively. We have $\overline{t}(w_0) = \{r\}$, $\overline{t}(w_1) = \{p_1, p_2, q\}$, $\overline{t}(w_2) = \{p_2\}$. In the same way as the last example we can show that $(p_1, w_0) = 0$, $(p_2, w_0) = 0$,$(q, w_0) = 1$, $(p_1, w_1) = 1$, $(p_2, w_1) = 1$ and $(p_2, w_2) = 1$ are all the causes of $r = 1$ at w_0 by the original and updated definition. However, in the case of modified definition, neither $(p_1, w_1) = 1$ nor $(p_2, w_1) = 1$ is the causes but $(p_1, w_1) = 1 \wedge (p_2, w_1) = 1$ is a cause of $r = 1$ at w_0. To see this notice that for any choice of $\overline{N}$ we will always have*

$$(\mathscr{K}, \overline{t}, w_0) \Vdash [(p_1, w_1) \leftarrow x', \overline{N} \leftarrow \overline{n}^*]r = 1.$$

for any choice of x'. Thus, $(p_1, w_1) = 1$ is not cause of $r = 1$ at w_0 by modified definition. Similar argument holds for $(p_2, w_1) = 1$. However, $(p_1, w_1) = 1 \wedge (p_2, w_1) = 1$ is a cause is showed by setting $\overline{N} = \emptyset$ and $\overline{x}' = (0,0)$. Thus, three definitions of causality need not always match in causal Kripke models.

The following theorem describes relationship between the three HP definitions in the causal models.

Definition B.2 ([16], Section 2.2)**.** *For any event α, any variable X, and any world w, $X = x$ is a part of cause of α by original (resp. updated, modified) definition of causality if it is a conjunct in the cause of α by original (resp. updated, modified) definition.*

Theorem B.3 ([16], Theorem 2.2.3)**.** *If $X = x$ is a part of cause of α in $(M,\overline{u})$ according to*

1. *the modified HP definition then $X = x$ is a part of cause of α in $(M,\overline{u})$ according to the original HP definition .*
2. *the modified HP definition then $X = x$ is a part of cause of α in $(M,\overline{u})$ according to the updated HP definition.*
3. *the updated HP definition then $X = x$ is a part of cause of α in $(M,\overline{u})$ according to the original HP definition.*

Now, we generalize this result to our framework of causal Kripke models.

Definition B.4. *For any event α, any variable X, and any world w', $(X,w) = x$ is a part of cause of α by original (resp. updated, modified) definition of causality if it is a conjunct in the cause of α at w' by original (resp. updated, modified) definition.*

Theorem B.5. *If $(X,w) = x$ is a part of cause of ϕ in $(\mathscr{K},\bar{t})$ at w' according to*

1. *the modified HP definition then $(X,w) = x$ is a part of cause of α in $(\mathscr{K},\bar{t})$ at w' according to the original HP definition.*
2. *the modified HP definition then $(X,w) = x$ is a part of cause of α in $(\mathscr{K},\bar{t})$ at w' according to the updated HP definition.*
3. *the updated HP definition then $(X,w) = x$ is a part of cause of α in $(\mathscr{K},\bar{t})$ at w' according to the original HP definition.*

Proof. For item 1, let $(X,w) = x$ be a part of cause of α in $(\mathscr{K},\bar{t})$ at a world w' according to the modified HP definition, so that there is a cause $\overline{Y} = \overline{y}$ such that $(X,w) = x$ is one of its conjuncts. Then there must exist a value $\overline{x}' \in \mathscr{R}(\overline{Y})$ and a set $\overline{N} \subseteq \mathscr{V} \times W \setminus \overline{Y}$, such that if $(\mathscr{K},\bar{t},w) \vdash X = n^*$ for every $(X,w) = n^* \in \overline{N} = \overline{n}^*$, then $(\mathscr{K},\bar{t},w') \vdash [\overline{Y} \leftarrow \overline{y}, \overline{N} \leftarrow \overline{n}^*]\neg\alpha$. Moreover $\overline{Y}$ is minimal.

We will show that $(X,w) = x$ is a cause of α. If $\overline{Y} = \{(X,w)\}$, then the original HP definition is satisfied by $(\overline{N},\overline{n}^*,x')$ given by the condition $AC2a^m$. If $|\overline{Y}| > 1$, then without loss of generality let $\overline{Y} = ((X_1,w_1),(X_2,w_2),\cdots,(X_n,w_n))$ and $(X,w) = (X_1,w_1)$. For any vector $\overline{Y}$, we use $\overline{Y}_{-1}$ to denote all components of Y except the first.

We will show that (X_1,w_1) is a cause of α in $(\mathscr{K},\bar{t})$ at w' according to the original definition. Since $\overline{Y} = \overline{y}$ is a cause of α in $(\mathscr{K},\bar{t})$ at w' according to the modified definition, by AC1 $(\mathscr{K},\bar{t},w_1) \vdash X_1 = x_1$ and $(\mathscr{K},\bar{t},w') \vdash \alpha$. Let $\overline{N}' = (\overline{Y}_{-1},\overline{N})$, $\overline{n^*}' = (\overline{y'}_{-1},\overline{n^*})$, $y' = y'_1$, where $\overline{y}'$ is as given by the modified definition. It is easy to see that $(\mathscr{K},\bar{t},w') \vdash [(X_1,w_1) \leftarrow x'_1, \overline{Y}_{-1} \leftarrow \overline{y}_{-1}, \overline{N} \leftarrow \overline{n^*}]\neg\alpha$ satisfying condition $AC2a$. Since (X_1,w_1) is single variable, $AC3$ holds trivially. Thus, to complete the proof of (a) we need to show that $AC2b^o$ holds. Suppose $AC2b^o$ does not hold. Then there exists a subset $\overline{Z'} \subseteq \mathscr{V} \times W \setminus (\overline{Y}_{-1} \cup \overline{N})$ of variables and value z^* such that (i) for each $Z \in \overline{Z'}$, $(\mathscr{K},\bar{t},w) \vdash Z = z^*$ and (ii) $(\mathscr{K},\bar{t},w') \vdash [(X_1,w_1) \leftarrow x_1, \overline{Y}_{-1} \leftarrow \overline{y}_{-1}, \overline{N} \leftarrow \overline{n^*}, \overline{Z'} \leftarrow \overline{z^*}]\neg\alpha$. But then $\overline{Y} = \overline{y}$ is not a cause of α according to the modified definition. Indeed, $AC2a^m$ is satisfied for $\overline{T}' = \overline{Y}_{-1}$ by setting $\overline{N} = ((X_1,w_1),\overline{N},\overline{Z'})$ and $\overline{n^*} = (x_1,\overline{n^*},\overline{z^*})$ and $\overline{t'} = \overline{y'}_{-1}$ violating $AC3$ for $\overline{Y} = \overline{y}$. i.e. , $\overline{Y} = \overline{y}$ is not a minimal cause by the modified definition as the conjunct obtained by removing $(X_1,w_1) = x_1$ from it is still a cause by the modified definition. This is a contradiction. Therefore, $AC2b^o$ is valid.

For item 2, the proof is similar in spirit. In addition to 1, we need to show that if $\overline{Y'} \subseteq \overline{Y}_{-1}$, $\overline{N'} \subseteq \overline{N}$, and $\overline{Z'} \subseteq \overline{Z}$, then

$$(\mathscr{K}, \bar{t}, w) \vdash [(X, w_1) \leftarrow x_1, \overline{Y}_{-1} \leftarrow \overline{y}_{-1}, \overline{N'} \leftarrow \overline{n^{*\prime}}, \overline{Z'} \leftarrow \overline{z^{*\prime}}\phi]$$

If $X' = \emptyset$, then the condition holds since $(X, w) = x$ is a cause of ϕ according to the original definition by item 1. In case this condition does not hold for some non-empty $\overline{Y'} \subseteq \overline{Y}_{-1}$, then $\overline{Y} = \overline{y}$ does not satisfy the minimimality condition AC3 of the modified HP definition (in causal Kripke models).

For item 3, the proof is same as item 1, upto the point where we have to prove $AC2^o$. Suppose there exists $\overline{Z'} \subseteq \overline{Z}$ such that

$$(\mathscr{K}, \bar{t}, w) \vdash [(X, w_1) \leftarrow x_1, \overline{Y}_{-1} \leftarrow \overline{y}_{-1}, \overline{N'} \leftarrow \overline{n^{*\prime}}, \overline{Z'} \leftarrow \overline{z^{*\prime}}\neg\phi]$$

then $\overline{Y}_{-1} \leftarrow \overline{y}_{-1}$ satisfies AC2a and $AC2b^u$. Thus, $\overline{Y} = \overline{y}$ does not satisfy the minimimality condition AC3 for the updated definition. Hence proved. □

C Soundness and completeness

In this section we provide the proof of soundness and (weak) completeness of the axiomatization given in Section 4.2. The proof is a modification of the proof provided in [16] to include the modal operators.

Proof. Showing that the axiomatization is sound is routine. It is straightforward to verify that all the axioms except G-axiom are valid and that modus ponens and necessitation preserve validity. For the G-axiom, note that the truth of the formula $(X, w) = x$ is independent of the world w' at which it is evaluated. Thus, if it is true at some world, then it is true at all the worlds, in particular true at all the world related to w.

To prove that the axiomatization is weakly complete, we show contrapositively that if $\nvdash \psi$ then there exists a model satisfying $\neg\psi$. As usual, starting with a consistent formula φ we obtain a maximal consistent set Σ containing all axioms such that $\varphi \in \Sigma$, is closed under $\wedge$ and consequence, and enjoys the disjunction property (see also the proof of Theorem 5.4.1 in [16, Section 5.5]).

Before moving to the details of the proof, we provide a high-level presentation of the argument, to help the reader follow: Given a consistent set of formulas, we can extract the formulas that do not contain modalities. Treating the variables (X, w) and (X, w) (where $w \neq w'$) as simply distinct variables, this set can be seen as a consistent set of formulas for the standard logic of causality presented in [16], because the axioms in Section 4.2 strictly extend the axioms of the logic in [16]. Then, by the completeness presented in [16], we get a set of structural equations, which readily provides a set of structural equations over a Kripke model with the empty relation (where (X, w) and (X, w') are now interpreted as the same variable at different points in the Kripke model). By the soundness of the axioms in Section 4.2, the set of non-modal formulas at each such state is consistent. These consistent sets guarantee that the “canonical” model we construct has enough points to interpret all the names that appear in our finite set of formulas. The proof then follows a standard filtration argument to show in the usual way the truth lemma for modal formulas.

Let φ be such that $\nvdash \neg\varphi$. Let us define $W_\varphi := \{w \in W \mid w \text{ appears in } \varphi\}$ and $S_\varphi := \{\psi, \Box\psi \mid \psi \text{ is a subformula of } \varphi\}$. Now let Σ be a maximal consistent set of formulas of $\mathbf{L}(W_\varphi)$ that contains φ. Given axiom C6 in [16, Section 5.4] and $\Diamond$-axiom and $\Box$-axiom, we can assume without loss of generality that all formulas are generated from $[\overline{Y} \leftarrow \overline{y}]X = x$ and $[\overline{Y} \leftarrow \overline{y}](X, w) = x$ using the connectives $\Diamond, \Box, \wedge, \vee$ and $\neg$. Notice that Σ “decides” the value of variables (X, w) for every $w \in W_\varphi$ (that is to say, $(X, w) = x \in \Sigma$ for some $x \in$ for some $x \in \mathscr{R}(X, w)$). Consider the set $B = \{[\overline{Y} \leftarrow \overline{y}](X, w) = x \in \mathbf{L}(W_\varphi) \mid$

$[\overline{Y} \leftarrow \overline{y}](X,w) = x \in \Sigma\}$. By the completeness in [16, Section 5.5], it follows that there exists a system of structural equations satisfying the non-modal formulas of Σ. Using this system we can define a causal Kripke model with domain W_φ, and empty Kripke relation. By the soundness of this system it follows that for every $w \in W_\varphi$ the set $\Sigma_w := \{[\overline{Y} \leftarrow \overline{y}]X = x \mid [\overline{Y} \leftarrow \overline{y}](X,w) = x \in \Sigma\} \cup B$ (the set Σ_w includes the) is consistent and hence can be extended to an maximal consistent set Σ'_w.

Let $S = S_\varphi \cup \{[\overline{Y} \leftarrow \overline{y}](X,w) = x, [\overline{Y} \leftarrow \overline{y}]X = x \in \mathbf{L}(W_\varphi)\}$. Notice that S is finite. Define an equivalence relation on maximal consistent sets extending B of $\mathbf{L}(W_\varphi)$, $T_1 \sim T_2$ if and only if $T_1 \cap S = T_2 \cap S$. Given that S is finite, there exist finite many equivalence classes. Let $\mathbb{W}$ be the set of equivalence classes, and let $\mathfrak{R} \subseteq \mathbb{W} \times \mathbb{W}$ be defined as $C_1 \mathfrak{R} C_2$ if and only if there exists $T_1 \in C_1$ and $T_2 \in C_2$, such that for all $\psi \in T_2$, $\Diamond\psi \in T_1$. Define a name assignment i such that $i(w) = [\Sigma'_w]$ for $w \in W_\varphi$, and arbitrarily otherwise. Finally define the structural equations, depending only on variables of W_φ, exactly as defined in [16, Section 5.4]. In particular the equations are independent of variables in $W \setminus W_\varphi$, and $f_{(X,w)}(\overline{y}) = x$, if and only if $[\overline{Y} \leftarrow \overline{y}]X = x \in T$ for any $T \in i(w)$ (given that $[\overline{Y} \leftarrow \overline{y}]X = x \in S$, this is well defined).

We claim that $\bar{t}, [T] \Vdash \psi$ if and only if $\psi \in T$, for every $\psi \in S$, and maximal consistent set T.

The proof proceeds via induction on the complexity of the formulas. For formulas of the form $[\overline{Y} \leftarrow \overline{y}]X = x$, $[\overline{Y} \leftarrow \overline{y}](X,w) = x$, and for logical connectives the proof is verbatim the same as that of [16, Section 5.4].

Finally, let's show this for the case when ψ is $\Diamond\sigma$.

First, let's assume that $\bar{t}, [T] \Vdash \Diamond\sigma$. Then there exists $C \in \mathbb{W}$ such that $[T]\mathfrak{R}C$ and $\bar{t}, C \Vdash \sigma$. By induction hypothesis $\sigma \in T'$, for every $T' \in C$. Since $\sigma \in S$, by the definition of $\mathfrak{R}$, it follows that $\Diamond\sigma \in T$.

For the converse direction, assume that $\Diamond\sigma \in T$. Notice preliminarily, that since $\Diamond\top \wedge \Box p \Rightarrow \Diamond p$ is a theorem of classical normal modal logic, then $(\Diamond\top \wedge \Box([\overline{Y} \leftarrow \overline{y}](X,w) = x)) \Rightarrow \Diamond[\overline{Y} \leftarrow \overline{y}](X,w) = x$ is provable in our system. From the G-axiom, this implies that also

$$(\Diamond\top \wedge ([\overline{Y} \leftarrow \overline{y}](X,w) = x)) \Rightarrow \Diamond[\overline{Y} \leftarrow \overline{y}](X,w) = x \tag{1}$$

is provable. Consider the set $Z_T = \{\tau \in \mathbf{L}(W_\varphi) \mid \Diamond\tau \notin T\}$. Clearly Z_T is an ideal of the free Boolean algebra of the logic. Given that $B \subseteq T$ and $\Diamond\sigma \in T$, it follows that $\Diamond\top \in T$, and by (1) it follows that $\Diamond B \subseteq T$ and so $B \cap Z_T = \varnothing$. Hence there exists a maximal consistent set T', extending $B \cap \{\sigma\}$ such that $T' \cap Z_T = \varnothing$. By definition $[T]\mathfrak{R}[T']$, and hence $\bar{t}, [T] \Vdash \Diamond\sigma$, as required.

The proof, that the model is recursive, follows again the proof of [16, Section 5.4], using the fact that our Kripke frame is finite. □

On Imperfect Recall in Multi-Agent Influence Diagrams

James Fox
University of Oxford
james.fox@cs.ox.ac.uk

Matt MacDermott
Imperial College London
m.macdermott21@imperial.ac.uk

Lewis Hammond
University of Oxford
lewis.hammond@cs.ox.ac.uk

Paul Harrenstein
University of Oxford
paul.harrenstein@cs.ox.ac.uk

Alessandro Abate
University of Oxford
aabate@cs.ox.ac.uk

Michael Wooldridge
University of Oxford
mjw@cs.ox.ac.uk

Multi-agent influence diagrams (MAIDs) are a popular game-theoretic model based on Bayesian networks. In some settings, MAIDs offer significant advantages over extensive-form game representations. Previous work on MAIDs has assumed that agents employ behavioural policies, which set independent conditional probability distributions over actions for each of their decisions. In settings with imperfect recall, however, a Nash equilibrium in behavioural policies may not exist. We overcome this by showing how to solve MAIDs with forgetful and absent-minded agents using mixed policies and two types of correlated equilibrium. We also analyse the computational complexity of key decision problems in MAIDs, and explore tractable cases. Finally, we describe applications of MAIDs to Markov games and team situations, where imperfect recall is often unavoidable.

1 Introduction

Multi-agent influence diagrams (MAIDs) are a graphical representation for dynamic non-cooperative games, which can be more compact and expressive than extensive-form games (EFGs) [25]. Like Bayesian networks (BNs), MAIDs use a directed acyclic graph (DAG) to represent conditional probabilistic dependencies between random variables, but they also specify decision and utility variables for each agent. Each agent selects a behavioural policy – independent conditional probability distributions (CPDs) over actions for each of their decision variables – to maximise their expected utility. A MAID's mechanised graph extends this DAG by explicitly representing each variable's distribution and showing which other variables' distributions matter to an agent optimising a particular decision rule [18, 25, 10].

MAIDs, and their causal variants [18], have been used in the design of safe and fair AI systems [14, 1, 15, 16, 7], to explore reasoning patterns and deception [40, 48], and to identify agents from data [22]. However, to date, agents in MAIDs are usually assumed to have perfect (or, at least, 'sufficient') recall [25]. This assumption is often unreasonable. For example, MAIDs must allow imperfect recall to handle bounded rationality, teams with imperfect communication [13], or memoryless policies in Markov games. However, forgetfulness (of previous observations) or absent-mindedness (about whether previous decisions have even been made) can prevent the existence of a Nash Equilibrium (NE) in behavioural policies. To overcome this, one can consider other solution concepts, such as mixed or correlated equilibria.

In this work, we focus on imperfect recall in MAIDs. Imperfect recall has already been extensively studied in EFGs [41, 26, 49], but a MAID's mechanised graph makes graphically explicit the semantic difference between behavioural and mixed policies (hidden in EFGs) and readily identifies forgetful or absent-minded agents (or teams). Our insights inspire two definitions of *correlated equilibrium* in MAIDs. The first follows from the normal-form game definition [2]. The second, based on von Stengel

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 201–220, doi:10.4204/EPTCS.379.17

and Forges' extensive-form correlated equilibrium [47], is more natural for dynamic settings, can yield greater social welfare, and is easier to compute. Again, mechanised graphs clearly depict the assumptions made in both. Next, we examine MAIDs from a computational complexity perspective by studying the decision problems of finding a best response, checking whether a policy profile is an NE, and checking whether each type of NE exists. These provide an insight into what makes particular instances hard, when computations can be made tractable, and rigorously identify which problems are suitable for analysis as MAIDs. Our results also apply to refinements of MAIDs, such as *causal games* [18]. We assume familiarity with EFGs [31], BNs [24], and the complexity classes P, NP, and PP [38]. Proof sketches are provided, but details are deferred to the appendices.

Related Work. There is a rich literature on influence diagrams [23] and imperfect recall has been studied in single-agent influence diagrams [33, 34, 29, 6, 35] as well as in EFGs [3, 21, 26, 41, 49]. However, to our knowledge, we are the first to focus on imperfect recall in influence diagrams with multiple agents.

A full policy profile in a MAID induces a BN, so many of our results inherit from that setting, where the decision problem variant of marginal inference is, in general, PP-complete [30]. However, we care about the cases we encounter in practice, not just the worst case. Marginal inference in a BN can be performed in time exponential in the treewidth of the underlying graph [24], which entails a poly-time algorithm when the treewidth is small. Similarly, we will see that tractable results for computations in MAIDs can be found when problems are restricted to certain settings. We also sometimes reduce from partial order games [50], which can be interpreted as MAIDs without chance nodes, with deterministic decision rules, and where each agent has a single utility node as a child of all the decision nodes.

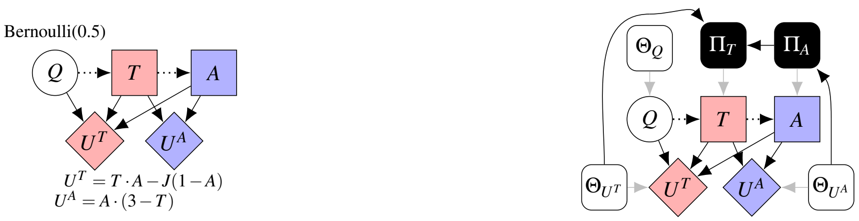
2 The Model

We use capital letters V for random variables, lowercase letters v for their instantiations, and bold letters $\boldsymbol{V}$ and $\boldsymbol{v}$, respectively, for sets of variables and their instantiations. We let $dom(V)$ denote the (finite, non-singleton) domain of V (for ease, we take this to be binary unless stated otherwise) and $dom(\boldsymbol{V}) := \times_{V \in \boldsymbol{V}} dom(V)$. Parents and children of V in a graph are denoted by $\mathbf{Pa}_V$ and $\mathbf{Ch}_V$, respectively (with $\mathbf{pa}_V$ and $\mathbf{ch}_V$ their instantiations) and $\Delta(X)$ denotes the set of all probability distributions over a set X.

Example 1. *An autonomous taxi decides whether to offer Alice a discount (T) depending on whether its journey count exceeds a quota (Q). Alice decides whether to accept a journey (A) depending on the price. The taxi wants to maximise profit, but if its journey count is less than the quota and Alice rejects it, the taxi pays a penalty (the municipality uses this mechanism to prevent a proliferation of unnecessary taxis). Alice's utility is a function of her decision and the price offered by the taxi.*

Figure 1a shows a MAID for this example. Chance variables (moves by nature), decision variables, and utility variables are represented by white circles, squares, and diamonds, respectively. Full edges leading into chance and utility nodes represent probabilistic dependence, as in a BN. Dotted edges leading into decision nodes identify information available to the agent when a decision D is made, so $\mathbf{pa}_D$, the values of $\mathbf{Pa}_D$, represents the decision context for D. In EFGs, imperfect information is represented using explicitly labelled information sets. In MAIDs, we can infer that Alice is unaware of the value of Q when making her decision by the lack of edge $Q \to A$. A parameterisation defines the CPDs for the chance and utility variables, whereas CPDs of decision nodes are chosen by the agents playing the game.

Definition 1 ([25]). *A* ***multi-agent influence diagram (MAID)*** *is a structure $\mathcal{M} = (\mathcal{G}, \boldsymbol{\theta})$. $\mathcal{G} = (N, \boldsymbol{V}, E)$ specifies a set of agents $N = \{1, \ldots, n\}$ and a DAG $(\boldsymbol{V}, E)$, where $\boldsymbol{V}$ is partitioned into chance variables*



(a) (b)

Figure 1: A MAID (a) and its mechanised graph (b) for Example 1, which is a perfect recall and imperfect, but sufficient, information game.

$\boldsymbol{X}$, decision variables $\boldsymbol{D} = \bigcup_{i \in N} \boldsymbol{D}^i$, and utility variables $\boldsymbol{U} = \bigcup_{i \in N} \boldsymbol{U}^i$. The parameters $\boldsymbol{\theta} = \{\theta_V\}_{V \in \boldsymbol{V} \setminus \boldsymbol{D}}$ define the CPDs $\Pr(V \mid \boldsymbol{Pa}_V)$ for each non-decision variable such that for any setting of the decision variables' CPDs, the resulting joint distribution over $\boldsymbol{V}$ is Markov compatible with the DAG, i.e., $\Pr(\boldsymbol{v}) = \prod_{V \in \boldsymbol{V}} \Pr(v \mid \boldsymbol{pa}_V)$.

Given a MAID, a **decision rule** π_D for $D \in \boldsymbol{D}$ is a CPD $\pi_D(D \mid \mathbf{Pa}_D)$. A **partial (behavioural) policy profile** $\pi_{\boldsymbol{D}'}$ is a set of decision rules for each $D \in \boldsymbol{D}' \subseteq \boldsymbol{D}$, whereas $\pi_{-\boldsymbol{D}'}$ is the set of decision rules for each $D \in \boldsymbol{D} \setminus \boldsymbol{D}'$. A **(behavioural) policy** $\boldsymbol{\pi}^i$ refers to $\boldsymbol{\pi}_{\boldsymbol{D}^i}$, and a **(full) policy profile** $\boldsymbol{\pi} = (\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^n)$ is a tuple of policies, where $\boldsymbol{\pi}^{-i} := (\boldsymbol{\pi}^1, \dots, \boldsymbol{\pi}^{i-1}, \boldsymbol{\pi}^{i+1}, \dots, \boldsymbol{\pi}^n)$. A decision rule is **pure** if $\pi_D(d \mid \mathbf{pa}_D) \in \{0, 1\}$, which holds for a policy (profile) if it holds for all decision rules in the policy (profile). For clarity, we use an overhead dot to mark this determinism, e.g., $\dot{\pi}_D, \dot{\boldsymbol{\pi}}^i$, or $\dot{\boldsymbol{\pi}}$.

By combining $\boldsymbol{\pi}$ with the partial distribution Pr over the chance and utility variables, we obtain a joint distribution:

$$\Pr^{\boldsymbol{\pi}}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{u}) := \prod_{V \in \boldsymbol{V} \setminus \boldsymbol{D}} \Pr(v \mid \mathbf{pa}_V) \cdot \prod_{D \in \boldsymbol{D}} \pi_D(d \mid \mathbf{pa}_D)$$

A full policy profile $\boldsymbol{\pi}$ therefore induces a BN with DAG given by the MAID's graph. Agent i's **expected utility** $EU^i(\boldsymbol{\pi})$ for a given policy profile $\boldsymbol{\pi}$ is defined as the expected sum of their utility variables:

$$EU^i(\boldsymbol{\pi}) := \sum_{U \in \boldsymbol{U}^i} \sum_{u \in dom(U)} \Pr^{\boldsymbol{\pi}}(U = u) \cdot u$$

Utility variables have deterministic CPDs, so can be interpreted as functions $U : dom(\mathbf{Pa}_U) \to \mathbb{R}$ to show their functional dependence on their parents (e.g., Figure 1a). An NE is defined in the usual way.

Definition 2 ([25])**.** *A (behavioural) policy profile $\boldsymbol{\pi}$ is a **Nash equilibrium (NE)** (in behavioural policies) if for every agent $i \in N$ and every alternative (behavioural) policy $\boldsymbol{\varpi}^i$: $EU^i(\boldsymbol{\pi}^{-i}, \boldsymbol{\pi}^i) \geq EU^i(\boldsymbol{\pi}^{-i}, \boldsymbol{\varpi}^i)$*

Collectively, the decision rules of decision variables and the CPDs of chance or utility nodes are known as mechanisms. A mechanism M_V for V is **strategically relevant** to a decision rule for D if the choice of the CPD at M_V can affect the optimal choice of this decision rule. Koller and Milch [25] define an associated sound and complete graphical criterion for strategic relevance, ***s*-reachability**, based on d-separation which can be checked in $\mathcal{O}(|\boldsymbol{V}| + |E|)$ time [43] (see Appendix A for formal definitions).

A MAID's regular graph $\mathscr{G}$ captures the probabilistic dependencies between **object-level** variables in the game's environment, but its **mechanised graph** $\mathsf{m}\mathscr{G}$ is an enhanced representation which adds an explicit representation of the strategically relevant dependencies between agents' decision rules and the game's parameterisation (see [18] for details). Each object-level variable $V \in \boldsymbol{V}$ has a mechanism parent M_V representing the distribution governing V: each decision D has a new *decision rule* parent $\Pi_D = \mathsf{M}_D$ and each non-decision V has a new *parameter* parent $\Theta_V = \mathsf{M}_V$, whose values parameterise the CPDs.

Agents select a decision rule π_D (i.e., the value of a decision rule variable Π_D) based on both the parameterisation of the game (i.e., the values of the parameter variables) and the selection of the other decision rules $\boldsymbol{\pi}_{-D}$ – these dependencies are captured by the edges from other mechanisms into decision rule nodes. s-reachability determines which of these edges are necessary, so $\mathsf{M}_V \rightarrow \Pi_D$ exists if and only if Π_D strategically relies on M_V. The mechanised graph for Example 1 (in Figure 1b) shows that Π_T strategically relies on Θ_{U^T} and Π_A, whereas Π_A only strategically relies on Θ_{U^A}. In contrast to a MAID's regular graph $\mathcal{G}$, which is a DAG, there may exist cycles between mechanisms (e.g., Figure 3a).

For convenience, we denote the set of agent i's behavioural policies as $\boldsymbol{P}^i := dom(\boldsymbol{\Pi}^i)$, with sets of pure policies denoted as $\dot{\boldsymbol{P}}^i$ and (pure) policy profiles denoted by $\boldsymbol{P}$ ($\dot{\boldsymbol{P}}$).

2.1 Concise Representations

A concise representation of MAIDs is needed for three reasons. First, real numbers may obscure the true complexity of the problems [5], so we assume that all probability parameters are given by a fraction of two integers, both expressed in finite binary notation. This is realistic since the probabilities are normally either assessed by domain experts or estimated by a learning algorithm and means that all CPDs can be read in poly-time. Second, even with binary variables, a joint distribution across $\boldsymbol{V}$ requires $2^{|\boldsymbol{V}|} - 1$ parameters. A MAID or BN's graphical Markov factorisation reduces this to $\sum_{V \in \boldsymbol{V}} 2^{|\mathbf{Pa}_V|}$, but this can still be exponential in $|\boldsymbol{V}|$. Therefore, it is standard [45, 42, 28, 24] to assume that the maximum in-degree in the graph is much less than $|\boldsymbol{V}|$ (or constant), so that the size of the CPDs are polynomial in $|\boldsymbol{V}|$. This means that the total representation of our MAID (including all CPDs) is polynomial in our chosen complexity parameter $|\boldsymbol{V}|$. Finally, as in BNs, our complexity results are strongly affected by the DAG's **treewidth**. The **treewidth** of a DAG measures its resemblance to a tree and is given by the number of vertices in the largest clique of the corresponding triangulated moral graph minus one [4].

3 Imperfect Recall in MAIDs

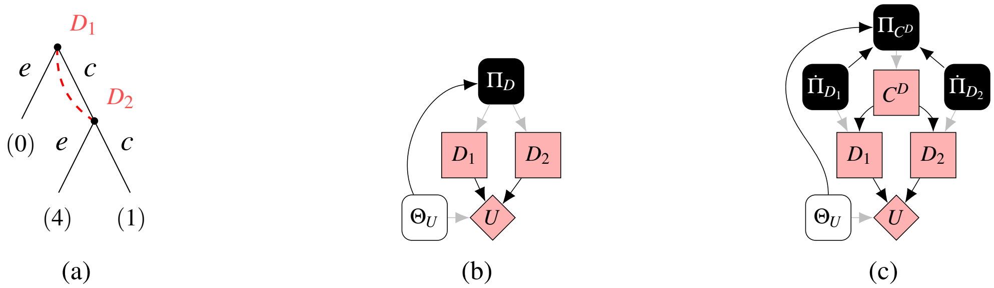
Agents may possess different degrees of information about the state of a game. A game has **perfect recall** if each agent remembers all their past decisions and observations, and it has **perfect information** if each agent is aware of *every* agent's past decisions and observations.

Definition 3 ([25])**.** *Agent i in a MAID $\mathcal{M}$ is said to have* ***perfect recall*** *if there exists a total ordering $D_1 \prec \cdots \prec D_m$ over $\boldsymbol{D}^i$ such that $(\boldsymbol{Pa}_{D_j} \cup D_j) \subseteq \boldsymbol{Pa}_{D_k}$ for any $1 \leq j < k \leq m$. $\mathcal{M}$ is a perfect recall game if all agents in $\mathcal{M}$ have perfect recall. $\mathcal{M}$ is a* ***perfect information*** *game if there exists such an ordering over $\boldsymbol{D}$.*

A MAID with perfect information (recall) can be transformed into an EFG with perfect information (recall), and vice versa [17]. Hence, these information conditions also guarantee the existence of an NE in pure (behavioural) policies in the MAID ([26] gives the equivalent results in EFGs). However, the mechanised representation of a MAID enables weaker criteria to be defined – **sufficient information** and **sufficient recall**. Later, in Proposition 3, we will see that these criteria preserve the NE existence results of perfect information and perfect recall games, respectively.

Definition 4. *Agent i in a MAID $\mathcal{M}$ has* ***sufficient recall*** *[36] if the subgraph of the mechanised graph $\mathsf{m}\mathcal{G}$ restricted to just agent i's decision rule nodes $\boldsymbol{\Pi}_{\boldsymbol{D}^i}$ is acyclic. $\mathcal{M}$ is a sufficient recall game if all agents in $\mathcal{M}$ have sufficient recall. $\mathcal{M}$ is a* ***sufficient information*** *game if the subgraph of $\mathsf{m}\mathcal{G}$ restricted to contain only and all decision rule nodes $\boldsymbol{\Pi}_{\boldsymbol{D}}$ is acyclic.*[1]

[1] Note that since previous work on influence diagrams has not modelled absent-mindedness (see our Definition 5 in Section 3.1), this definition implicitly assumes each mechanism variable has a single child.



(a) (b) (c)

Figure 2: The EFG (a) and the mechanised graphs for an absent-minded driver choosing behavioural (b) or mixed (c) policies.

3.1 Forgetfulness and Absent-Mindedness

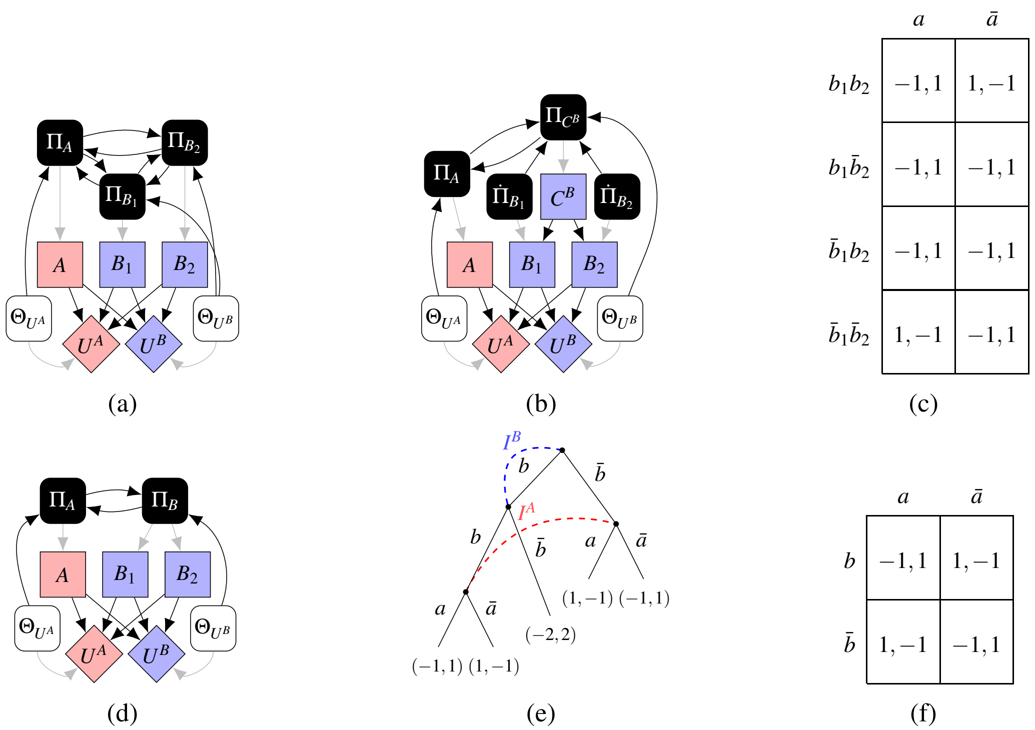
Previous work on MAIDs has assumed perfect or sufficient recall. We now begin the contributions of this paper by distinguishing between two types of imperfect recall in MAIDs. **Forgetfulness** applies when an agent forgets an observation or the *outcome* of one of their previous decisions. **Absent-mindedness** applies when an agent cannot even remember whether they have previously made a decision. To make this distinction, we leverage the following insight: *mechanism nodes represent the CPDs governing object-level variables. Every edge between a mechanism and object-level node represents an independent draw from the mechanism's distribution.* We now provide formal definitions.

Definition 5. *Agent i has **imperfect recall** in a MAID $\mathcal{M}$ if for every total ordering $D_1 \prec \cdots \prec D_m$ over $\boldsymbol{D}^i$ there exists some $j < k$ such that $(\boldsymbol{Pa}_{D_j} \cup D_j) \not\subseteq \boldsymbol{Pa}_{D_k}$ (i.e., if agent i does not have perfect recall). Agent i is **forgetful** if such a D_j and D_k have distinct decision rules and is **absent-minded** if in $\mathcal{M}$'s mechanised graph, a decision rule node has more than one outgoing edge to a decision node.*

To motivate our definition of absent-mindedness in MAIDs, we revisit Piccione and Rubinstein's absent-minded driver game [41] (its EFG is in Figure 2a). A driver on a highway may take one of two exits. Taking the first, second, or no exit yields a payoff of 0, 4, or 1, respectively. Adopting Aumann [3]'s *modified multi-selves approach* (i.e., that the driver should only be able to control her current action, not her future actions), the driver does not know which junction she is facing, so she must have the same decision rule at both junctions. We make absent-mindedness explicit with a shared decision rule node Π_D for D_1 and D_2 in the mechanised graph (Figure 2b) (note this is consistent with our mechanised graph definition). Π_D's *two outgoing edges now represent two independent draws from the same distribution*. For D_i and D_j to share a decision rule, it is necessary that $dom(D_i) = dom(D_j)$ and $dom(\boldsymbol{Pa}_{D_i}) = dom(\boldsymbol{Pa}_{D_j})$. Note that perfect recall implies that for any two decisions belonging to the same agent, one's set of parents is a strict superset of the other's, so their decision rules have a different type signature, which rules out absent-mindedness.

In the following examples, used just to explain this paper's concepts, Alice and Bob play variations of matching pennies with the usual payoffs given according to the *final* state of their two coins (where a/b and $\bar{a}/\bar{b}$ represent heads and tails, respectively). Example 2 illustrates a consequence of Bob being forgetful – meaning he cannot remember the *outcome* of his previous decision. In Example 3, Bob is absent-minded – he cannot remember whether he has made a decision at all.

Example 2 (Figures 3a-3c)**.** *Bob is told he must submit a move in advance (B_1) and then confirm it on game day (B_2). If his moves agree, payoffs correspond with normal matching pennies, but if his moves disagree, he must forfeit and always loses (these payoffs are shown in Figure 3c). Bob is forgetful, so on game day he cannot remember his advance choice (i.e., the edge $B_1 \to B_2$ is missing in Figure 3a).*



(a) (b)

	a	$\bar{a}$
b_1b_2	$-1,1$	$1,-1$
$b_1\bar{b}_2$	$-1,1$	$-1,1$
$\bar{b}_1b_2$	$-1,1$	$-1,1$
$\bar{b}_1\bar{b}_2$	$1,-1$	$-1,1$

(c)

(d) (e)

	a	$\bar{a}$
b	$-1,1$	$1,-1$
$\bar{b}$	$1,-1$	$-1,1$

(f)

Figure 3: The mechanised graphs for forgetful Bob (Example 2) using (a) behavioural or (b) mixed policies, with normal-form in (c). (d) The mechanised graph for absent-minded Bob (Example 3) using a behavioural policy, with EFG and normal-form representations in (e) and (f).

Example 3 (Figures 3d-3f)**.** *In a new game, the pennies start heads up, and Bob decides whether or not to turn the coin over (B_1). He is absent-minded, so when he sees heads he cannot remember whether he has already made his move, and he decides again (B_2). If he turns the coin having previously chosen to keep heads, Bob gets a -2 penalty and Alice a $+2$ bonus. In all other cases, the payoffs correspond with normal matching pennies (payoffs are shown at the leaves of the EFG in Figure 3e).*

Observe that the MAID's regular graph (just the object-level variables) is identical for both Figures 3a and 3d with the missing $B_1 \to B_2$ edge implying imperfect recall. The difference between forgetfulness and absent-mindedness is only revealed by the mechanised graph. Forgetful Bob has two independent decision rules Π_{B_1} and Π_{B_2} for B_1 and B_2. Absent-minded Bob only has one shared decision rule Π_B.

Examples 2 and 3 demonstrate that both types of imperfect recall can mean an NE in behavioural policies may not exist, even in zero-sum two agent MAIDs with binary decisions. The normal-form games (in Figures 3c and 3f) show that neither contains an NE in pure policies. It is also easy to prove non-existence in behavioural policies (see Appendix B). This arises due to the grand best response function being non-convex valued, which violates a condition of Kakutani's fixed point theorem.

Proposition 1. *Both forgetfulness and absent-mindedness can prevent the existence of an NE in behavioural policies.*

4 Solution Concepts for MAIDs under Imperfect Recall

To overcome the fact that a behavioural policy NE may not exist in imperfect recall MAIDs, one can use mixed or correlated policies. These ensure that the grand best response function always satisfies the conditions of Kakutani's fixed point theorem, so an equilibrium always exists. We show how the assumptions behind mixed policies, behavioural mixtures, and correlated equilibria (well-studied in EFGs [21, 47], but unexplored in MAIDs) are made graphically explicit in mechanised graphs.

4.1 Mixed Policies and Behavioural Mixtures

Behavioural policies allow agents to randomise independently at every decision node. By contrast, a **mixed policy** $\mu^i \in \Delta(\dot{\boldsymbol{P}}^i)$ is a distribution over pure policies. It allows an agent to coordinate their choice of decision rules at different decisions by randomising once at the game's outset and then committing to the assigned pure policy. More generally, **behavioural mixtures** in $\Delta(\boldsymbol{P}^i)$ are distributions over all behavioural policies. They allow agents to randomise *both* at the outset of the game and before each decision. The outcome of the first randomisation determines the distributions for the others.

A behavioural mixture changes the specification of the game because it can require correlation between different decision rules. At the object-level, a behavioural mixture for agent i requires a new (correlation) decision variable C^i with $\mathbf{Pa}_{C^i} = \varnothing$, $\mathbf{Ch}_{C^i} = \boldsymbol{D}^i$, and $dom(C^i) = \boldsymbol{P}^i$ (the set of all behavioural policies). The decision rules for each D^i become conditional on C^i, so each value of C^i determines a behavioural policy. This explains why C^i and still every $D \in \boldsymbol{D}^i$ are decision nodes – the agent chooses the CPDs for both. Even in the mixed policy case, where each D^i depends deterministically on C^i, the agent chooses the dependence independently from choosing the distribution over C^i. In the mechanised graph (see Figure 2c), C^i gets an associated mechanism variable Π_{C^i} for the distribution C^i is drawing from (its mechanism parents are again determined by s-reachability).

In EFGs, the mechanism by which agents decide on their decision rules is not explicitly shown. Mechanised graphs, however, show clearly when an agent chooses to randomise. Behavioural and mixed policies are the limiting cases of behavioural mixtures: the former where the distribution over $\boldsymbol{P}^i$ is deterministic; the latter where the decision rules $\boldsymbol{\Pi}_{\boldsymbol{D}^i}$ are deterministic. The difference between forgetful Bob in Example 2 using a behavioural or mixed policy is shown in Figures 3a and 3b. For Bob's behavioural policy, C^B and Π_{C^B} are omitted as the decision rules Π_{B_1} and Π_{B_2} are independent. This leaves a normal mechanised graph. Whereas, if Bob uses a mixed policy, he only randomises once from Π_{C^B} at the start of the game to select a pure policy at C^B. This fixes deterministic decision rules at $\dot{\Pi}_{B_1}$ and $\dot{\Pi}_{B_2}$.

Proposition 2. *Given a MAID $\mathcal{M}$ with any partial profile $\boldsymbol{\pi}^{-i}$ for agents $-i$, then if agent i is not absent-minded, for any behavioural policy $\boldsymbol{\pi}^i$ there exists a pure policy $\dot{\boldsymbol{\pi}}^i$ which yields a payoff at least as high against $\boldsymbol{\pi}^{-i}$. On the other hand, if agent i is absent-minded in $\mathcal{M}$ across a pair of decisions with descendants in $\boldsymbol{U}^i$, then there exists a parameterisation of $\mathcal{M}$ and a behavioural policy $\boldsymbol{\pi}^i$ which yields a payoff strictly higher than any payoff achievable by a pure policy.*

Proposition 2 says that a non-absent-minded agent cannot achieve more expected utility by using a behavioural rather than a pure (or mixed) policy, but an absent-minded agent often can. Consider Figure 2c, where $dom(C^D) = \dot{\boldsymbol{P}}^D$, the set of all the driver's pure policies. Π_{C^D} represents the distribution over $dom(C^D)$, so D_1 and D_2 must both be e or both be c. Therefore, $EU^D \leq 1$ under any mixed policy. Whereas, under the behavioural policy $\pi_D^1(e) = \frac{1}{3}$, $EU^D = \frac{4}{3}$. This highlights an important difference between absent-mindedness and forgetfulness. Under perfect recall, every mixed policy has an equivalent behavioural policy, in the sense of inducing the same distribution over outcomes against every opposing policy profile [18]. Under forgetfulness, whilst a mixed policy might not have an equivalent behavioural policy, a behavioural policy always has an equivalent mixed policy [26], so there must exist a pure policy which performs just as well. On the other hand, under absent-mindedness, neither mixed nor behavioural policies are guaranteed to have an equivalent of the other type, so there can be a behavioural policy which outperforms every mixed policy against a given policy profile.

We introduce mixed policies (and behavioural mixtures) to MAIDs to allow more generality in modelling when agents randomise and to guarantee an NE. However, a mixed policy can require exponentially more parameters $\mathcal{O}(2^{2^{|\boldsymbol{V}|}})$ than a behavioural policy $\mathcal{O}(2^{|\boldsymbol{V}|})$ to define. Moreover, single agents are often more naturally modelled as randomising once they meet decision points [26] (this changes for team situations described in Section 6). It is therefore important to know when existence of each type of NE is guaranteed. The sufficient recall result was proved by [18], which we adapt to get the sufficient information result (in Appendix B). The mixed policies result follows directly from Nash's theorem [37].

Proposition 3. *A MAID with sufficient information always has an NE in pure policies, a MAID with sufficient recall always has an NE in behavioural policies, and every MAID has an NE in mixed policies.*

Since both sufficient recall and sufficient information (Definition 4) can be checked in poly-time[2], they expand the class of games that have simple NEs beyond those identifiable using an EFG. For example, we can check in poly-time that the MAID in Figure 1a is an imperfect, but sufficient, information game, and hence know that there must exist an NE in pure policies.

4.2 Correlated Equilibria

We have just shown how mechanised graphs can explicitly represent the assumption behind mixed policies: a *single* agent uses a source of randomness to correlate their decision rules. We now do the same for when *multiple* agents can use the same source of randomness, so the choice of pure policy made by each agent may be correlated. An equilibrium in such a game is called a *correlated equilibrium (CE)* [2], which is a distribution κ over the set of all pure policy profiles, i.e., $\kappa \in \Delta(\dot{\boldsymbol{P}})$. A mediator samples $\dot{\boldsymbol{\pi}}$ according to κ, then recommends to each agent i the pure policy $\dot{\boldsymbol{\pi}}^i$. The distribution κ is a CE if no agent, given their information, has an incentive to unilaterally deviate from their recommended policy $\dot{\boldsymbol{\pi}}^i$.

Definition 6. *In a MAID, $\kappa \in \Delta(\dot{\boldsymbol{P}})$ is a **correlated equilibrium (CE)** if and only if $\forall i$, $\forall \dot{\boldsymbol{\pi}}^i, \dot{\boldsymbol{\varpi}}^i \in \dot{\boldsymbol{P}}^i$:*

$$\sum_{\dot{\boldsymbol{\pi}}^{-i} \in \dot{\boldsymbol{P}}^{-i}} \kappa(\dot{\boldsymbol{\pi}}^i, \dot{\boldsymbol{\pi}}^{-i}) EU^i(\dot{\boldsymbol{\pi}}^i, \dot{\boldsymbol{\pi}}^{-i}) \geq \sum_{\dot{\boldsymbol{\pi}}^{-i} \in \dot{\boldsymbol{P}}^{-i}} \kappa(\dot{\boldsymbol{\pi}}^i, \dot{\boldsymbol{\pi}}^{-i}) EU^i(\dot{\boldsymbol{\pi}}^{-i}, \dot{\boldsymbol{\varpi}}^i)$$

We illustrate how MAIDs and their mechanised graphs make explicit the assumptions used for a CE using a costless-signal variation of Spence's job market game [46].

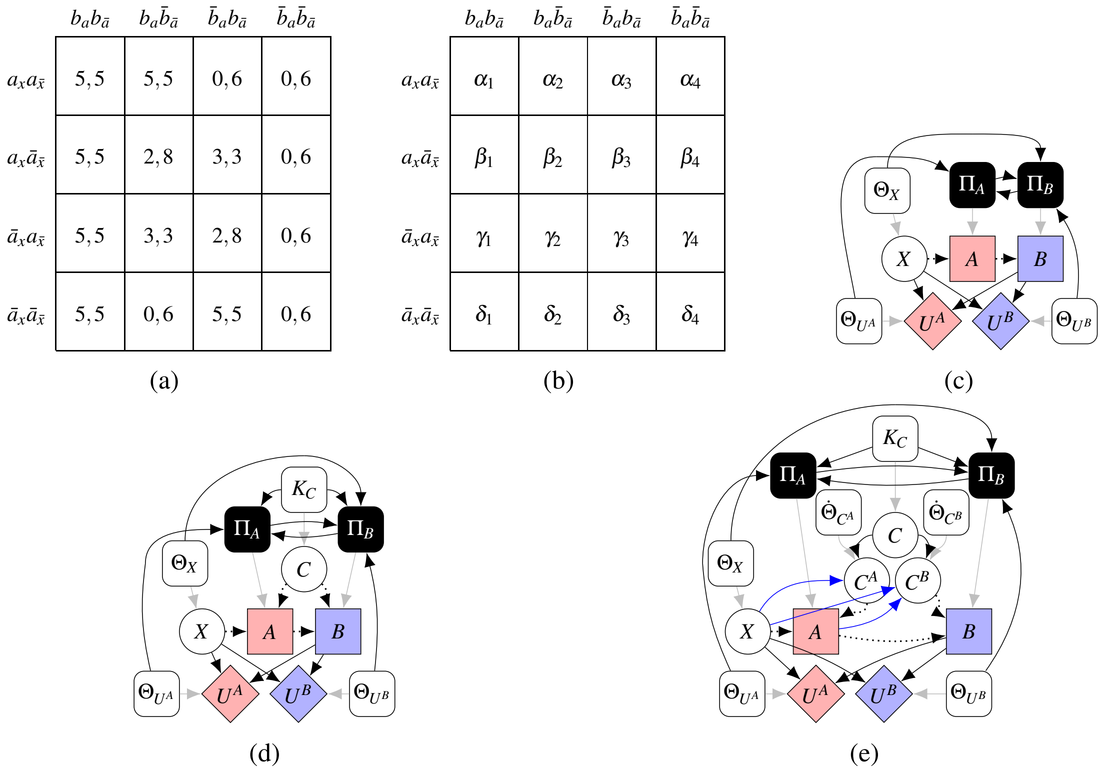
Example 4. *Alice is hardworking or lazy (X) with equal probability. She applies for a job with Bob by deciding which costless signal (A) to send. Bob can distinguish between the signals, but does not know Alice's true temperament. He decides whether to offer the job (B) to Alice. The utility functions for Alice and Bob are $U^A = (6 - 2X) \cdot B$ and $U^B = 6 + (10X - 6) \cdot B$, respectively.*

The mechanised graph for the original game's MAID is shown in Figure 4c. The cycle between Π_A and Π_B reveals that each agent's decision rule strategically relies on the other agent's decision rule.[3] Therefore, the MAID has insufficient information and no proper subgames, making it difficult to solve.

To find the CE of this game, a trusted mediator is added using a *correlation variable* C with $\mathbf{Pa}_C = \varnothing$, $\mathbf{Ch}_C = \boldsymbol{D}$, and $dom(C) = \dot{\boldsymbol{P}}$. In the mechanised graph, C's associated mechanism variable K_C represents the distribution $\kappa \in \Delta(\dot{\boldsymbol{P}})$ that the mediator draws a pure policy profile according to. This time, since

[2] The mechanised graph is constructed using s-reachability, which uses the poly-time graphical criterion d-separation [43].

[3] That Bob strategically relies on Alice's decision rule might be less obvious than the fact that Alice strategically relies on Bob's decision rule. The dependency occurs because since Bob can observe A, this unblocks an active path $\Pi_A \rightarrow A \leftarrow X \rightarrow U^B$ in the independent mechanised graph, so Π_A is s-reachable from Π_B.



	$b_a b_{\bar a}$	$b_a \bar b_{\bar a}$	$\bar b_a b_{\bar a}$	$\bar b_a \bar b_{\bar a}$
$a_x a_{\bar x}$	5,5	5,5	0,6	0,6
$a_x \bar a_{\bar x}$	5,5	2,8	3,3	0,6
$\bar a_x a_{\bar x}$	5,5	3,3	2,8	0,6
$\bar a_x \bar a_{\bar x}$	5,5	0,6	5,5	0,6

(a)

	$b_a b_{\bar a}$	$b_a \bar b_{\bar a}$	$\bar b_a b_{\bar a}$	$\bar b_a \bar b_{\bar a}$
$a_x a_{\bar x}$	α_1	α_2	α_3	α_4
$a_x \bar a_{\bar x}$	β_1	β_2	β_3	β_4
$\bar a_x a_{\bar x}$	γ_1	γ_2	γ_3	γ_4
$\bar a_x \bar a_{\bar x}$	δ_1	δ_2	δ_3	δ_4

(b)

(c)

(d)

(e)

Figure 4: The sub-figures (a) and (b) give the expected payoff for each agent under each pure policy profile and the parameterisation of the distribution κ, respectively. The mechanised graph for Example 4's original MAID is shown in (c), and the mechanised graphs for when a trusted mediator gives public or private recommendations to find a CE are shown in (d) and (e), respectively. The blue edges are added to the graph in (e) for a MAID-CE's staggered recommendations.

K_C is fixed as κ at the game outset instead of being chosen by any agent, C acts as a chance variable (in contrast to the correlation decision variable introduced for mixed policies and behavioural mixtures).

There is a well-known difference between public and private recommendations. If public, every payoff in the convex hull of the set of NE payoffs can be attained by a CE; however, if the recommendations are private, then the payoffs to each agent in a CE can lie outside this convex hull (e.g., Aumann's game of chicken [2]). This distinction is made explicit in the MAID's graph. If the recommendations are public, then the full outcome of C (the pure policy profile chosen by the mediator) is known by every agent (shown by the dotted edges between C and both A and B in Figure 4d). If the recommendations are private, then each agent only observes their decision rules (action recommendations) in C's outcome, i.e., all recommendations given to other players are hidden (at C^A and C^B in Figure 4e). In this latter case, the agent infers, using Bayes' rule, a posterior over the pure policy profile that was chosen (and also which action was recommended to the other agent(s)). If κ is a CE, then each agent picks for their decision D's decision rule the mediator's recommendation, i.e., $\dot\pi_D$ where $c = \dot{\boldsymbol{\pi}}$. The set of variables $\boldsymbol{D}$ remain as decisions because agents are free to deviate from their recommendation and pick any CPDs as decision rules for their decisions.

This mediator's distribution $\kappa \in \Delta(\dot{\boldsymbol{P}})$ can be parameterised according to that in Figure 4b. Note that $b_a \bar b_{\bar a}$ denotes the pure policy profile where Bob offers the job (b) to Alice if she selects a and Bob does not offer the job ($\bar b$) if Alice selects $\bar a$. Using the expected payoff for Alice and Bob under each pure policy profile (Figure 4a), Definition 6's incentive constraints define 24 inequalities that must be satisfied by the CE distribution. After some algebra, we find that $\alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = \gamma_1 = \gamma_2 = \gamma_3 = 0$; $\alpha_4, \beta_4, \gamma_4, \delta_4 \geq 0$; $\alpha_4 - 2\beta_4 + 3\gamma_4 \geq 0$, and $3\beta_4 - 2\gamma_4 + \delta_4 \geq 0$. Any CE, therefore, has Bob never offering a job to Alice because they play the pure policy $\bar{b}_a\bar{b}_{\bar{a}}$ with probability 1, i.e., Bob's decision rule has $\pi^B(B = \bar{b} \mid A = a) = \pi^B(B = \bar{b} \mid A = \bar{a}) = 1$. The remaining constraints require Alice not to give any incentive for Bob to offer her a job by making the conditional probability of Alice being hardworking too high relative to the conditional probability of her being lazy when he receives the signal a or $\bar{a}$. These constraints find that every CE will result in $EU^A = 0$ and $EU^B = 6$. This is unsurprising because, in a signaling game with costless signals, every CE will be a 'pooling equilibrium' [8] (an equilibrium in which Alice chooses the same action regardless of their temperament).

Whilst the CE is among the best-known solution concepts for normal-form games, and is efficiently computable in that setting (e.g., via linear programming [19]), there can be an exponential number of pure policies (so an exponential number of incentive constraints) in EFGs and even in bounded treewidth MAIDs. It is therefore currently unknown if a CE can be found in an EFG or MAID in poly-time. Motivated by these tractability concerns, Von Stengel and Forges proposed an *extensive-form correlated equilibrium (EFCE)* [47]. Along similar lines, we define a *MAID correlated equilibrium*.

Instead of revealing the entire recommendation $\dot{\boldsymbol{\pi}}^i$ to each agent i immediately, we let the mediator *stagger* their recommendations. This is made visible in the mechanised graph by adding the blue edges in Figure 4e. Importantly, if an agent deviates from any recommendation, then the mediator will *cease giving further recommendations to that agent* (but will still give recommendations to all other agents). Thus, the incentive constraints are now tied to the threat of the mediator withholding future information.

Definition 7. *Given a distribution $\kappa \in \Delta(\dot{\boldsymbol{P}})$, consider the MAID with an additional correlation variable C with $\boldsymbol{Pa}_C = \varnothing$, $\boldsymbol{Ch}_C = \{C_D\}_{D\in\boldsymbol{D}}$, and $\boldsymbol{Ch}_{C_D} = \{D\}$ for each D. Let a pure policy profile $\dot{\boldsymbol{\pi}}$ be selected at C according to κ. Then, when each decision context $\boldsymbol{pa}_D$ is reached, agent i receives a recommended move $d \in dom(D)$ specified by $\dot{\boldsymbol{\pi}}_D \in \dot{\boldsymbol{\pi}}$ (C_D hides all other recommendations $\dot{\boldsymbol{\pi}}_{-D} \in \dot{\boldsymbol{\pi}}$). A* ***MAID correlated equilibrium (MAID-CE)*** *is an NE of this game in which no agent has an incentive to deviate from their recommendations.*

The localised recommendations in a MAID-CE pose weaker incentive constraints compared to a CE, so the set of MAID-CE outcomes is larger. As such, MAID-CEs can lead to Pareto-improvements over the CEs (and NEs) in a game. We now give one such MAID-CE. The mediator chooses a signal s with equal probability for type $X = x$, i.e., $\Pr(c_A = a \mid X = x) = \Pr(c_A = \bar{a} \mid X = x) = 0.5$. Bob is recommended to offer Alice a job (b) when Alice's action matches s and to reject otherwise ($\bar{b}$). If $X = \bar{x}$, then the recommendation to Alice is arbitrary and is independent of the signal s, which is only shown to hardworking Alice. Because the mediator only gives Alice her recommendation once her decision context $\mathbf{Pa}_A$ is set, lazy Alice cannot know s. Therefore, in any situation, lazy Alice's action will match s with probability $\frac{1}{2}$. Consequently, when Bob is called to play (i.e., the decision context $\mathbf{Pa}_B$ is set), and Alice's action matches s, Alice is twice as likely to be hardworking than lazy (so $EU^B = \frac{20}{3}$ for offering Alice a job rather than $EU^B = 6$ for rejecting her). If instead, Alice's action does not match s, then he knows with certainty that Alice is lazy, so his best response is to reject. Overall, Alice's expected payoff in this MAID-CE is 3.5, and Bob's is 6.5 (higher than 0 and 6, respectively, for all CEs).

A MAID-CE can be computed in poly-time if the treewidth is bounded, via a reduction to a linear program. We follow Huang et al [20]'s method because the information sets in an EFG are in bijection with the decision contexts in a MAID, but relax beyond their conditions as MAIDs only require sufficient (rather than perfect) recall [20]. Any distribution over pure policies induced by an NE can be represented using a distribution κ, and hence any mixed NE (or equivalent behavioural NE) is also a CE and MAID-CE. As every MAID has an NE in (mixed) policies, every MAID must also have a CE and a MAID-CE.

Proposition 4. *A MAID-CE in bounded treewidth MAIDs with sufficient recall can be found in poly-time.*

5 Complexity Results in MAIDs

We now give some complexity results in MAIDs. Our first follows from the known result in normal-form games [9]. Any normal-form game $\mathscr{N}$ can be reduced to a MAID where each agent has one utility node (which copies the payoffs in $\mathscr{N}$) and one decision node. The domains of the decision variables are the set of each agent's pure strategies in $\mathscr{N}$. Edges are added from every $D \in \boldsymbol{D}$ to every $U \in \boldsymbol{U}$.

Proposition 5. *In a MAID, finding an NE in mixed policies is* PPAD*-hard.*

Problem	**Input**	**Question**
Is-Best-Response	$\mathscr{M}$, i, $\boldsymbol{\pi}^{-i}$, $q \in \mathbb{Q}$	Is there some $\hat{\boldsymbol{\pi}}^i$ such that $EU^i(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}) > q$?
Is-Nash	$\mathscr{M}$, $\boldsymbol{\pi}$	Is $\boldsymbol{\pi}$ a (behavioural) NE of $\mathscr{M}$?
Non-Emptiness:	$\mathscr{M}$	Does $\mathscr{M}$ have a (behavioural) NE?

Table 1: Three decision problems in MAIDs with behavioural policies.

In the following results, we focus on the complexity of the decision problems in Table 1.

Proposition 6. Is-Best-Response *is* $\mathrm{NP}^{\mathrm{PP}}$*-complete,* NP*-complete when restricted to MAIDs with graphs of bounded treewidth, and* PP*-complete if both* $|\boldsymbol{D}^i|$ *and the in-degrees of* $\boldsymbol{D}^i$ *are bounded.*

Proof sketch. Is-Best-Response is in $\mathrm{NP}^{\mathrm{PP}}$ because given $\hat{\boldsymbol{\pi}}^i$, we can verify that $EU^i(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi}^{-i}) > q$ in poly-time using a PP oracle for inference in a BN [30]. With bounded treewidth, verification can be done in poly-time. The final setting is in PP by analogy with Kwisthout's PARAMETER TUNING [27]. For the general case's hardness, we can reduce from E-Majsat as in [39], where MAP-nodes are replaced by agent i's decision nodes; for bounded treewidth, we can reduce from MAXSAT as in [12]; and for the final case, Is-Best-Response with $|\boldsymbol{D}^i| = 0$ is the same as inference in a BN. □

Proposition 6 suggests Is-Best-Response is, in general, only tractable if inference is easy *and* $|\boldsymbol{D}^i|$ is bounded by a constant. Proposition 7 then explains the decision problem's name.

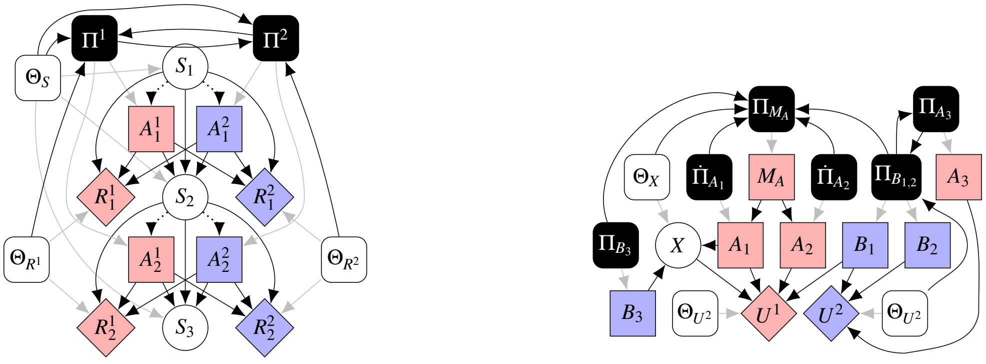
Proposition 7. *If the in-degrees of* $\boldsymbol{D}^i$ *are bounded and* Is-Best-Response *can be solved in poly-time, then a best response policy for agent i to a partial profile* $\boldsymbol{\pi}^{-i}$ *can be found in polynomial time.*

Proposition 8. Is-Nash *is* $\mathrm{coNP}^{\mathrm{PP}}$*-complete, and* coNP*-complete when restricted to MAIDs with graphs of bounded treewidth. The general problem remains* $\mathrm{coNP}^{\mathrm{PP}}$*-hard in sufficient information MAIDs. In MAIDs without chance variables, the problem remains* coNP*-hard.*

Proof sketch. For membership, we can check that $\boldsymbol{\pi}$ is *not* an NE by guessing an agent i and checking if $\boldsymbol{\pi}^{\boldsymbol{i}} \in \boldsymbol{\pi}$ is a best response in poly-time using a PP-oracle (this is unnecessary if the graph has bounded treewidth). Hardness comes from the single-agent setting where it is the complement of Is-Best-Response. In MAIDs without chance variables, we reduce from partial order games [50]. □

Proposition 3 shows when Non-Emptiness is vacuous. However, in an insufficient recall MAID, Non-Emptiness is, in general, intractable even without chance variables.

Proposition 9. Non-Emptiness *is* NEXPTIME*-hard and becomes* NEXPTIME*-complete if we restrict to MAIDs without chance variables.*



(a)

(b)

Figure 5: Mechanised graphs for a CE with (a) public and (b) private recommendations, where the blue edges are added for a MAID-CE; (c) a Markov game;(d) a team setting with imperfect communication.

Proof sketch. For hardness, we can reduce from partial order games. Without chance variables, we can determine NON-EMPTINESS using a similar algorithm to that in [50]. It exploits the setting's determinism: payoffs are poly-time computable and the number of policy profiles is reduced to $\mathscr{O}(2^{|\boldsymbol{V}|})$. □

Proposition 10. *In a MAID with sufficient information, if the in-degrees of* $\boldsymbol{D}$ *are bounded and* IS-BEST-RESPONSE *can be solved in poly-time, then a pure NE can be found in poly-time.*

This result suggests an NE can be found efficiently in certain MAIDs, but even in games without sufficient information, NEs can be found more efficiently in a MAID than in an EFG. The mechanised graph dependencies reveal more 'subgames' – parts of the MAID that can be solved independently from the rest – to which dynamic programming can be applied [25, 17]. As finding an NE in both EFGs and MAIDs depends significantly on the game's size, this can empirically lead to large compute savings [25].

6 Applications and Conclusion

We introduced forgetfulness and absent-mindedness as properties of individual agents (due to imperfect memory). However, imperfect recall also commonly arises in *team situations*; each team consists of several agents targeting a common goal with imperfect communication. Forgetfulness or absent-mindedness occurs when an agent does not know their teammates' actions (or observations) or whether they have acted at all. Mechanised graphs represent these situations where teams often employ a mix of randomisation strategies (e.g., Figure 5b). For mixed policies, the random seed is chosen at the start, before the agents set out following their distinct policies. For behavioural policies, agents pick a new random seed at every decision point. Behavioural mixtures correspond to randomising at both stages.

Another application of imperfect recall in MAIDs is to *Markov (or 'stochastic') games* [44], in which the agents move between different states over time (e.g., Figure 5a). At each time step t, each agent i selects an action A_t^i, and the game probabilistically transitions to a new state S_{t+1}, depending on the previous state S_t and the actions selected, and each agent receives a payoff R_t^i. Each S_{t+1} and R_t^i has parents $\{S_t, A_t^1, \dots, A_t^n\}$ and must be identically distributed for all t, again represented using shared mechanism variables. Often, the agent must learn a memoryless, stationary policy $\pi^i : S \to \Delta(A^i)$, where S is the set of states and $\Delta(A^i)$ the set of probability distributions over agent i's actions. Hence, the agents are absent-minded (every decision A_{t+1}^i of agent i shares the same decision rule) and use *behavioural* policies (since the action selected in each state is independently stochastic). In light of Proposition 1, it is therefore natural to ask whether a Markov game may not have an NE in memoryless stationary policies. It is known that infinite-horizon Markov games might not (for a counterexample see [11]). Although infinite games lie outside of the scope of this paper, it is nonetheless insightful to note that this possible non-existence is due to absent-mindedness: if agents can choose a different decision rule at each time step, a behavioural NE is guaranteed [32].

We have shown how to handle imperfect recall in MAIDs by overcoming the potential lack of NEs in behavioural policies using mixed and correlated equilibria. EFGs leave many assumptions about how agents play games hidden, but mechanised graphs make explicit the assumptions behind imperfect recall (both forgetfulness and absent-mindedness), mixed policies, and two types of correlated equilibria. Our complexity results highlight the importance of restricting the use of MAIDs to those with a limited number of decision variables and bounded treewidth. Finally, our applications to Markov games and team situations show that imperfect recall broadens the scope of what can be modelled using MAIDs.

Acknowledgements The authors wish to thank Ryan Carey, Tom Everitt, and Francis Rhys Ward for invaluable feedback, as well as three anonymous reviewers for their helpful comments. Fox was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (Reference: EP/S024050/1), MacDermott was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (Reference: EP/S023356/1), Hammond was supported by an EPSRC Doctoral Training Partnership studentship (Reference: 2218880), and Wooldridge was supported by a UKRI Turing AI World Leading Researcher Fellowship (Reference: EP/W002949/1).

References

[1] Carolyn Ashurst, Ryan Carey, Silvia Chiappa & Tom Everitt (2022): *Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness*. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, pp. 9494–9503, doi:10.1609/aaai.v36i9.21182.

[2] Robert J Aumann (1974): *Subjectivity and correlation in randomized strategies*. *Journal of mathematical Economics* 1(1), pp. 67–96, doi:10.1016/0304-4068(74)90037-8.

[3] Robert J Aumann, Sergiu Hart & Motty Perry (1997): *The absent-minded driver*. *Games and Economic Behavior* 20(1), pp. 102–116, doi:10.1006/game.1997.0577.

[4] Hans L Bodlaender (1993): *A linear time algorithm for finding tree-decompositions of small treewidth*. In: *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pp. 226–234, doi:10.1145/167088.167161.

[5] Hans L Bodlaender, Frank van den Eijkhof & Linda C van der Gaag (2002): *On the complexity of the MPA problem in probabilistic networks*. In: *ECAI*, pp. 675–679.

[6] Cassio P de Campos & Qiang Ji (2008): *Strategy selection in influence diagrams using imprecise probabilities*. In: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 121–128.

[7] Micah Carroll, Alan Chan, Henry Ashton & David Krueger (2023): *Characterizing Manipulation from AI Systems*. *arXiv preprint arXiv:2303.09387*.

[8] In-Koo Cho & David M Kreps (1987): *Signaling games and stable equilibria*. *The Quarterly Journal of Economics* 102(2), pp. 179–221, doi:10.2307/1885060.

[9] Constantinos Daskalakis, Paul W Goldberg & Christos H Papadimitriou (2009): *The complexity of computing a Nash equilibrium*. SIAM Journal on Computing 39(1), pp. 195–259, doi:10.1145/1132516.1132527.

[10] A. P. Dawid (2002): *Influence Diagrams for Causal Modelling and Inference*. International Statistical Review 70(2), pp. 161–189, doi:10.1111/j.1751-5823.2002.tb00354.x.

[11] Luca De Alfaro & Rupak Majumdar (2001): *Quantitative Solution of Omega-Regular Games*. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing, pp. 675–683, doi:10.1016/j.jcss.2003.07.009.

[12] Cassio Polpo De Campos & Fabio Gagliardi Cozman (2005): *The inferential complexity of Bayesian and credal networks*. In: IJCAI, 5, Citeseer, pp. 1313–1318.

[13] Apiruk Detwarasiti & Ross D Shachter (2005): *Influence diagrams for team decision analysis*. Decision Analysis 2(4), pp. 207–228, doi:10.1287/deca.1050.0047.

[14] Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega & Shane Legg (2021): *Agent incentives: A causal perspective*. In: Proceedings of the AAAI Conference on Artificial Intelligence, 35, pp. 11487–11495, doi:10.1609/aaai.v35i13.17368.

[15] Tom Everitt, Marcus Hutter, Ramana Kumar & Victoria Krakovna (2021): *Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective*. Synthese 198(Suppl 27), pp. 6435–6467, doi:10.1007/s11229-021-03141-4.

[16] Sebastian Farquhar, Ryan Carey & Tom Everitt (2022): *Path-specific objectives for safer agent incentives*. In: Proceedings of the AAAI Conference on Artificial Intelligence, 36, pp. 9529–9538, doi:10.1609/aaai.v36i9.21186.

[17] Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate & Michael Wooldridge (2021): *Equilibrium Refinements for Multi-agent Influence Diagrams: Theory and Practice*. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, pp. 574–582.

[18] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate & Michael Wooldridge (2023): *Reasoning about causality in games*. Artificial Intelligence 320, p. 103919, doi:10.1016/j.artint.2023.103919.

[19] Sergiu Hart & David Schmeidler (1989): *Existence of correlated equilibria*. Mathematics of Operations Research 14(1), pp. 18–25, doi:10.1287/moor.14.1.18.

[20] Wan Huang & Bernhard von Stengel (2008): *Computing an extensive-form correlated equilibrium in polynomial time*. In: International Workshop on Internet and Network Economics, Springer, pp. 506–513, doi:10.1007/978-3-540-92185-1_56.

[21] Mamoru Kaneko & J Jude Kline (1995): *Behavior strategies, mixed strategies and perfect recall*. International Journal of Game Theory 24(2), pp. 127–145, doi:10.1007/bf01240038.

[22] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott & Tom Everitt (2022): *Discovering Agents*. arXiv preprint arXiv:2208.08345.

[23] Uffe B Kjaerulff & Anders L Madsen (2008): *Bayesian networks and influence diagrams*. Springer Science+ Business Media 200, p. 114.

[24] Daphne Koller & Nir Friedman (2009): *Probabilistic graphical models: principles and techniques*. MIT press.

[25] Daphne Koller & Brian Milch (2003): *Multi-agent influence diagrams for representing and solving games*. Games and economic behavior 45(1), pp. 181–221, doi:10.1016/s0899-8256(02)00544-4.

[26] Harold W. Kuhn (1953): *Extensive Games and the Problem of Information*. In: Contributions to the Theory of Games (AM-28), 2, Princeton University Press, pp. 193–216, doi:10.1515/9781400881970-012.

[27] Johan Kwisthout & Linda C van der Gaag (2008): *The computational complexity of sensitivity analysis and parameter tuning*. In: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, pp. 349–356.

[28] Johan Henri Petrus Kwisthout et al. (2009): *The computational complexity of probabilistic networks*. Utrecht University.

[29] Steffen L Lauritzen & Dennis Nilsson (2001): *Representing and solving decision problems with limited information*. Management Science 47(9), pp. 1235–1251, doi:10.1287/mnsc.47.9.1235.9779.

[30] Michael L Littman, Stephen M Majercik & Toniann Pitassi (2001): *Stochastic boolean satisfiability*. Journal of Automated Reasoning 27(3), pp. 251–296.

[31] Michael Maschler, Shmuel Zamir & Eilon Solan (2020): *Game theory*. Cambridge University Press.

[32] Eric Maskin & Jean Tirole (2001): *Markov perfect equilibrium: I. Observable actions*. Journal of Economic Theory 100(2), pp. 191–219, doi:10.1006/jeth.2000.2785.

[33] Denis Deratani Mauá, Cassio P de Campos & Marco Zaffalon (2012): *Solving limited memory influence diagrams*. Journal of Artificial Intelligence Research 44, pp. 97–140, doi:10.1613/jair.3625.

[34] Denis Deratani Mauá & Fabio Gagliardi Cozman (2016): *Fast local search methods for solving limited memory influence diagrams*. International Journal of Approximate Reasoning 68, pp. 230–245, doi:10.1016/j.ijar.2015.05.003.

[35] Chris van Merwijk, Ryan Carey & Tom Everitt (2022): *A Complete Criterion for Value of Information in Soluble Influence Diagrams*. Proceedings of the AAAI Conference on Artificial Intelligence 36(9), pp. 10034–10041, doi:10.1609/aaai.v36i9.21242.

[36] Brian Milch & Daphne Koller (2008): *Ignorable Information in Multi-agent Scenarios*. Technical Report MIT-CSAIL-TR-2008-029, Computer Science and Artificial Intelligence Laboratory, MIT.

[37] J. F. Nash (1950): *Equilibrium Points in N-person Games*. Proceedings of the National Academy of Sciences 36(1), pp. 48–49.

[38] Christos Papadimitriou (1994): *Computational Complexity*. Addison Wesley.

[39] James D Park & Adnan Darwiche (2004): *Complexity results and approximation strategies for MAP explanations*. Journal of Artificial Intelligence Research 21, pp. 101–133, doi:10.1613/jair.1236.

[40] Avi Pfeffer & Ya'akov Gal (2007): *On the reasoning patterns of agents in games*. In: AAAI, pp. 102–109.

[41] Michele Piccione & Ariel Rubinstein (1997): *On the interpretation of decision problems with imperfect recall.* Games and Economic Behavior 20(1), pp. 3–24, doi:10.1016/0165-4896(96)81573-3.

[42] Dan Roth (1996): *On the hardness of approximate reasoning.* Artificial Intelligence 82(1-2), pp. 273–302, doi:10.1016/0004-3702(94)00092-1.

[43] Ross D Shachter (1998): *Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams).* In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp. 480–487.

[44] Lloyd S Shapley (1953): *Stochastic games.* Proceedings of the national academy of sciences 39(10), pp. 1095–1100.

[45] Solomon Eyal Shimony (1994): *Finding MAPs for belief networks is NP-hard.* Artificial intelligence 68(2), pp. 399–410, doi:10.1016/0004-3702(94)90072-8.

[46] Michael Spence (1978): *Job market signaling.* In: Uncertainty in economics, Elsevier, pp. 281–306.

[47] Bernhard Von Stengel & Françoise Forges (2008): *Extensive-form correlated equilibrium: Definition and computational complexity.* Mathematics of Operations Research 33(4), pp. 1002–1022, doi:10.1287/moor.1080.0340.

[48] Francis Rhys Ward, Francesca Toni & Francesco Belardinelli (2022): *On Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios.* In: AAMAS, pp. 1759–1761.

[49] Kevin Waugh, Martin Zinkevich, Michael Johanson, Morgan Kan, David Schnizlein & Michael H Bowling (2009): *A Practical Use of Imperfect Recall.* In: SARA.

[50] Valeria Zahoransky, Julian Gutierrez, Paul Harrenstein & Michael Wooldridge (2021): *Partial order games.* Games 13(1), p. 2, doi:10.3390/g13010002.

A Strategic Relevance and Subgames

Koller and Milch define **strategic relevance** to infer whether the choice of a decision rule can affect the optimality of another decision rule [25]. Hammond et al. extend strategic relevance to also consider whether the parameterisation of non-decision nodes can affect the decision rule's optimality [18]. Intuitively, a mechanism M_V is strategically relevant to the decision rule Π_D of $D \in \boldsymbol{D}^i$ if the choice of CPD at M_V can affect agent i's utility nodes that are downstream of D (i.e., those in $\boldsymbol{U}^i \cap \mathbf{Desc}_D$). Formally:

Definition 8 ([25, 18])**.** *Recall that $dom(\Pi_D)$ gives the set of possible decision rules at Π_D for decision node D. Given a MAID with $D \in \boldsymbol{D}^i$ and $V \neq D \in \boldsymbol{D}$, the mechanism M_V for V is* ***strategically relevant*** *to Π_D if there exist two joint distributions over $\boldsymbol{V}$ parameterised by mechanisms m and m' respectively such that:*

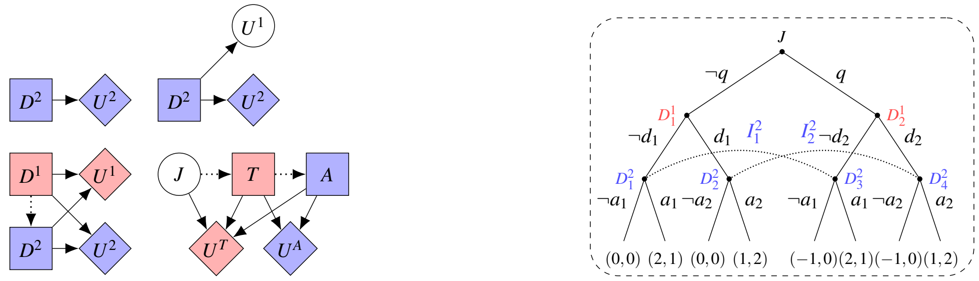
- $\pi_D \in \arg\max_{\varpi_D \in dom(\Pi_D)} EU^i((\varpi_D, \boldsymbol{\pi}_{-D}) \mid \mathsf{m})$
- m *differs from* m' *only at* M_V,
- $\pi_D \notin \arg\max_{\varpi_D \in dom(\Pi_D)} EU^i((\varpi_D, \boldsymbol{\pi}_{-D}) \mid \mathsf{m}')$*, and neither does any decision rule ϖ_D that agrees with π_D on all $\boldsymbol{pa}_D$ such that* $\Pr(\boldsymbol{pa}_D \mid \mathsf{m}') > 0$.

The first two conditions say: if the decision rule π_D is optimal for the MAID parameterisation (i.e., the setting of all mechanism variables) m, and Π_D does not strategically rely on M_V, then π_D must also be optimal for any other parameterisation m' that differs from m only at M_V. The third condition deals with sub-optimal decision rules in response to zero-probability decision contexts (i.e., non-credible threats).

Koller and Milch [25] also derive a graphical criterion for strategic relevance, called *s-reachability*, which is sound (if M_V is strategically-relevant to Π_D, then M_V is s-reachable from Π_D) and complete (if M_V is s-reachable from Π_D, then there is some parameterisation m of the MAID and some policy profile π such that M_V is strategically-relevant to Π_D). This uses the *independent mechanised graph* $\mathsf{m}_\perp\mathcal{G}$, which contains a separate mechanism parent for each variable in the original MAID graph, but no edges between the mechanism variables.

Definition 9 ([25]). *M_V is s-reachable from Π_D if $\mathsf{M}_V \not\perp_{\mathsf{m}_\perp\mathcal{G}} \boldsymbol{U}^i \cap \boldsymbol{Desc}_D \mid D, \boldsymbol{Pa}_D$.*

s-reachability determines which inter-mechanism edges are present in the MAID's mechanised graph; $\mathsf{M}_V \rightarrow \Pi_D$ exists in the mechanised graph if and only if Π_D strategically relies on M_V.



(a) (b)

Figure 6: (a) shows the four subdiagrams (three of which are 'proper') of the MAID in Figure 1a and (b) shows the corresponding EFG in which none of the MAID's proper subgames can be recognised.

We now briefly introduce subgames (see [18]) for more details) because they simplify the presentation of some of our proofs in Appendix B. Subgames in EFGs represent parts of the game that can be solved independently from the rest. In MAIDs, they fulfil the same purpose: they identify parts of the game that can be solved independently (and allow a subgame-perfect equilibrium refinement to be defined). Subgames in MAIDs are found by exploiting s-reachability to find the graphs underlying the subgames, called sub-diagrams. To then find the subgames for each subdiagram, the parameterisation of the remaining variables is updated to be consistent with the original game and graph structure.

Importantly, because MAIDs explicitly represent conditional independencies between variables, we can often find more subgames in a MAID than in a corresponding EFG. This is the case for Example 1's MAID (shown in Figure 1a) with the four subdiagrams (three proper) in Figure 6a. Each subdiagram has a set of associated subgames, one for each instantiation of the variables outside of the subdiagram. None of the proper MAID subgames can be recognised as subgames in the corresponding EFG (in Figure 6b).

Definition 10. *Given a MAID $\mathcal{M} = (\mathcal{G}, \boldsymbol{\theta})$, with $\mathcal{G} = (N, \boldsymbol{V}, E)$, the subgraph $(\boldsymbol{V}', E')$ of $\mathcal{G}$, along with the set of agents $N' \subseteq N$ possessing decision variables in that subgraph, is known as a* ***subdiagram*** *$\mathcal{G}' = (N', \boldsymbol{V}', E')$ if:*

- *$\boldsymbol{V}'$ contains every variable Z such that M_Z is s-reachable from some Π_D with $D \in \boldsymbol{V}'$,*
- *$\boldsymbol{V}'$ contains, for all $X, Y \in \boldsymbol{V}'$, every variable that lies on a directed path $X \dashrightarrow Y$ in $\mathcal{G}$.*

A ***subgame*** *of* $\mathscr{M}$ *is a new MAID* $\mathscr{M}' = (\mathscr{G}', \boldsymbol{\theta}')$ *where* $\mathscr{G}'$ *is a subdiagram of* $\mathscr{G}$ *and* $\boldsymbol{\theta}'$ *is defined by* $\Pr'(\boldsymbol{v}'; \boldsymbol{\theta}') := \Pr(\boldsymbol{v}' \mid \boldsymbol{z}; \boldsymbol{\theta})$*, where* $\boldsymbol{z}$ *is some instantiation of the variables* $\boldsymbol{Z} = \boldsymbol{V} \setminus \boldsymbol{V}'$*. A subgame is* ***feasible*** *if there exists a policy profile* $\boldsymbol{\pi}$ *where* $\Pr^{\boldsymbol{\pi}}(\boldsymbol{z}) > 0$*.*

The first condition on $\boldsymbol{V}'$ ensures that for any decision variable D in the subdiagram, any variable whose mechanism may impact the optimal decision rule for D is also included in the graph. The second condition says that additional variables may also be included in the subdiagram as long as mediators are included too. This ensures that the CPDs for all the variables in the subgame remain consistent.

B Proofs

Proposition 1. *Both forgetfulness and absent-mindedness can prevent the existence of an NE in behavioural policies.*

Proof. Example 2 (Figures 3a-3c) and Example 3 (Figures 3d-3f) are counterexamples for each case.

Proof for Example 2 (forgetfulness): The normal-form game showing the payoffs for each agent is shown in Figure 3c. First, observe that there are no NE in pure policies. Now, suppose that there does exist an NE in behavioural policies. If Alice always plays a or always $\bar{a}$ – i.e., $\pi^A(a) = 1$ or $\pi^A(a) = 0$ – then Bob's best response is always $\bar{b}_1\bar{b}_2$ or always b_1b_2, respectively. However, this does not form an NE. So, Alice must select a stochastic decision rule π_A and be indifferent (by the principle of indifference) between a and $\bar{a}$.

Letting Π_{B_1} and Π_{B_2} be parameterised by $p, q \in [0,1]$ where $\pi_{B_1}(b_1) = p$ and $\pi_{B_2}(b_2) = q$, we obtain two constraints on p and q. On the one hand, by virtue of Alice's indifference, Bob's behavioural policy $\boldsymbol{\pi}^B$ must result in $\pi^B(\neg b_1, \neg b_2) = \pi^B(b_1, b_2)$, and so: $(1-p)(1-q) = pq \implies p+q = 1$. On the other hand, Bob receives utility -1 if his policy $\boldsymbol{\pi}^B$ results in any outcome with $B_1 = \neg b_1$ and $B_2 = b_2$, or $B_1 = b_1$ and $B_2 = \neg b_2$, whatever the choice of $\boldsymbol{\pi}^A$. Therefore, we must have that $\pi^B(\neg b_1, b_2) + \pi^B(b_1, \neg b_2) < \pi^B(b_1, b_2) + \pi^2(\neg b_1, \neg b_2)$ and thus, by substituting in the result that $p + q = 1$: $(1-p)q + p(1-q) < pq + (1-p)(1-q) \implies (2p-1)^2 < 0.$. This contradiction implies that the MAID for Example 2 has no NE in behavioural policies.

To further understand this example, let us again write Bob's policy as a tuple (p,q), and suppose $\pi_A(a) = 0.5$. Then, either pure policy $(1,1)$ and $(0,0)$ is a best response for Bob with $EU^B = 0$. But, consider the convex combination of these best responses $0.5 \cdot (1,1) + 0.5 \cdot (0,0) = (0.5, 0.5)$. Under this policy, each of the eight outcomes in the payoff matrix is equally likely and so Bob's expected payoff drops to $(-1-1-1+1+1-1-1-1)/8 = -0.5$. Since a convex combination of best responses is no longer a best response, Bob's best response function is not convex-valued, and so nor is the grand best response function. The conditions of Kakutani's fixed point theorem are not satisfied, which explains why a Nash equilibrium need not exist.

Proof for Example 3 (absent-mindedness): First, observe from the normal-form game in Figure 3f that there is no NE in pure policies in this game. Next, suppose there exists a NE in behavioural policies and let Π_B be parameterised by $p \in [0,1]$, where $\pi_B(b) = p$ for $p \in [0,1]$. Alice's payoff only depends on her policy π^A when Bob plays bb or $\bar{b}\bar{b}$, for which Alice has pure best responses. This implies that, at an NE, $p^2 = (1-p)^2 \implies p = 0.5$. Therefore, Alice's policy is irrelevant and $EU^B = -1$ ($EU^B = 0$) if he does (doesn't) forfeit, which happens with probability 0.5. Therefore, Bob's policy is dominated by his pure policies, with worst-case payoff $EU^B = -1$. This contradicts the assumption of an NE in behavioural policies.

Explanation: If $\pi_A(a) = 0.5$, then $p = 0$ and $p' = 1$ are both best responses for Bob with $EU^B = 0$. However, the convex combination $0.5p + 0.5p'$ gives expected payoff to Bob $EU^B = 0.25 \cdot 1 + 0.25 \cdot (-1) + 0.5 \cdot (-10) = -5$ and is therefore not a best response. Again this is due to the fact that under behavioural policies, in situations of imperfect recall, a convex combination of pure policies can introduce outcomes that could not occur under either pure policy. Under a mixed combination of pure policies, Alice will always follow one or the other, and so no new outcomes are introduced. However, under a behavioural combination, two independent absent-minded draws from the same distribution over actions can come out differently, introducing new potential outcomes—in this case forfeit. □

Proposition 2. *Given a MAID $\mathscr{M}$ with any partial profile $\boldsymbol{\pi}^{-i}$ for agents $-i$, then if agent i is not absent-minded, for any behavioural policy $\boldsymbol{\pi}^i$ there exists a pure policy $\dot{\boldsymbol{\pi}}^i$ which yields a payoff at least as high against $\boldsymbol{\pi}^{-i}$. On the other hand, if agent i is absent-minded in $\mathscr{M}$ across a pair of decisions with descendants in $\boldsymbol{U}^i$, then there exists a parameterisation of $\mathscr{M}$ and a behavioural policy $\boldsymbol{\pi}^i$ which yields a payoff strictly higher than any payoff achievable by a pure policy.*

Proof. Let $\boldsymbol{\pi}^i$ be a behavioural policy and begin with any decision node $D \in \boldsymbol{D}^i$ with decision rule $\pi_D \in \boldsymbol{\pi}^i$. Now $\pi^i_D(d \mid \mathbf{pa}_D)$ is the probability of choosing $d \in dom(D)$ at D when $\mathbf{Pa}_D = \mathbf{pa}_D$ according to $\boldsymbol{\pi}^i$. Since agent i is not absent-minded, the expected payoff for agent i can be written $EU^i(\boldsymbol{\pi}^i, \boldsymbol{\pi}^{-i}) = \sum_{d \in dom(D)} \pi^i(d \mid \mathbf{pa}_D)\lambda_d + \nu$, where each coefficent λ_d and ν are independent of $\pi^i_D(d \mid \mathbf{pa}_D)$. Consider the action $\hat{d} \in dom(D)$ which achieves the highest λ_d (i.e., contributes most the expected utility) Setting $\pi^i_D(\hat{d} \mid \mathbf{pa}_D) = 1$ therefore yields a payoff at least as high. The first claim therefore follows by repeating this argument for every $D \in \boldsymbol{D}^i$.

For the converse claim, agent i is absent-minded, which means that at least two of agent i's decision nodes must draw from an identical distribution. Without loss of generality, call these D_l and D_m. Recall that for this to be the case, $dom(D_l) = dom(D_m)$ and $dom(\boldsymbol{Pa}_{D_l}) = dom(\boldsymbol{Pa}_{D_m})$. Now consider an outcome of the game $\hat{\boldsymbol{v}} \in dom(\boldsymbol{V})$ where $\mathbf{pa}_{D_l} = \mathbf{pa}_{D_m}$, but $d_l \neq d_m$. Since D_l and D_m have descendants in $\boldsymbol{U}^i$, Parameterise the MAID $\mathscr{M}$ such that $EU^i = 1$ if and only if $\boldsymbol{V} = \hat{\boldsymbol{v}}$. For all other game outcomes $\boldsymbol{v} \neq \hat{\boldsymbol{v}}$, let $EU^i = 0$. The claim follows since the outcome $\hat{\boldsymbol{v}}$ cannot be instantiated by any pure policy for agent i, but can be instantiated by any behavioural policy for agent i that has a (shared) decision rule for D_l and D_m that assigns a positive probability to both actions d_l and d_m. □

Proposition 3. *A MAID with sufficient information always has an NE in pure policies, a MAID with sufficient recall always has an NE in behavioural policies, and every MAID has an NE in mixed policies.*

Proof. The mixed policies case follows from Nash's theorem since all the finite number of random variables in a MAID have finite domains [37]. Hammond et al. proved the case with sufficient recall [18].

We now consider the sufficient information case where we show that a NE in pure policies must exist. Begin with an arbitrary policy profile across all decision nodes in the original MAID, $\mathscr{M}$. Decision rules associated with each $D \in \boldsymbol{D}$ can be optimised by iterating backwards through a subdiagram ordering $\mathscr{G}_1 \prec \cdots \prec \mathscr{G}_m$ of $\mathscr{M}$'s subdiagrams such that $\mathscr{G}_j \prec \mathscr{G}_k$ implies that $\mathscr{G}_j$ is *not* a subdiagram of $\mathscr{G}_k$. When $\mathscr{M}$ is a sufficient information game, this means that $\mathscr{G}_m$ contains just one decision node for some agent $i \in N$, and, for each subdiagram $\mathscr{G}_j$ where $1 \leq j < m$, $\mathscr{G}_{j-1}$ contains *at most* one additional decision variable. Several subdiagrams can have the same set of decisions, $\boldsymbol{D}_k$, so we choose a single subdiagram $\mathscr{G}_k$ (one with the fewest nodes $\boldsymbol{V}'$) for each $\boldsymbol{D}_k$ and discard the others. Each subdiagram in this ordering has an associated subgame for each setting of the nodes which have a child in $\boldsymbol{V}'$.

When considering each subgame $\mathscr{M}_{m-j}$ for $\mathscr{G}_{m-j}$, the decision rules for all decision nodes in proper subgames of $\mathscr{M}_{m-j}$ will have already been optimised and fixed in previous iterations, so these are now chance nodes in $\mathscr{M}_{m-j}$. In addition, the decision node D_{m-j} in $\mathscr{M}_{m-j}$ does not strategically rely on any of the decision nodes outside of $\mathscr{M}_{m-j}$. Therefore, this step is localised to computing only the optimal decision rule for D_{m-j}. Since this is a single-agent single-decision optimisation, we know that there must exist a pure decision rule best response. In the case of a tie, pick one arbitrarily. After repeating this optimisation process for all subgames in the MAID, we know that every decision node must have a pure decision rule, so we have found a NE in pure policies, as required. □

Proposition 4. *A MAID-CE in bounded treewidth MAIDs with sufficient recall can be found in poly-time.*

Proof sketch. We follow Huang and von Stengel's method for this result [20]. Our result comes from the observation that if there is sufficient recall in a MAID, then: (i) the set of decision contexts of every decision node in the MAID is in bijection with the set of all information sets in a corresponding EFG; and (ii) sufficient recall is sufficient for the ordering of decision contexts analogous to Huang and von Stengel's ordering of information sets. □

Lemma 1. *If* IS-BEST-RESPONSE *can be solved in poly-time, then agent i's expected utility under a best response to a partial policy profile* $\boldsymbol{\pi}^{-i}$ *in a MAID can be found in poly-time.*

Proof. This follows immediately from using binary search over agent i's policies and uses the fact that we are restricting parameters in the MAID to be rational numbers. □

Proposition 7. *If the in-degrees of* $\boldsymbol{D}^i$ *are bounded and* IS-BEST-RESPONSE *can be solved in poly-time, then a best response policy for agent i to a partial policy profile* $\boldsymbol{\pi}^{-i}$ *can be found in poly-time.*

Proof. Begin by constructing the MAID $\mathscr{M}(\boldsymbol{\pi}^{-i})$ by replacing decision nodes $\boldsymbol{D} \setminus \boldsymbol{D}^{-i}$ as chance nodes with CPDs given by $\boldsymbol{\pi}^{-i}$. Next, use Lemma 1 to compute agent i's expected utility under a best response policy in $\mathscr{M}(\boldsymbol{\pi}^{-i})$ and use this value as q. Take each of agent i's decision variables $D \in \boldsymbol{D}^i$ and build a new MAID $\mathscr{M}(\boldsymbol{\pi}^{-i}, \pi_D)$ for every possible decision rule of D (i.e., replace D as a chance node with CPD π_D). The fact that the in-degrees of agent i's decision nodes are bounded, bounds the number of these MAIDs. For each induced MAID, we can then use a poly-time algorithm for IS-BEST-RESPONSE to determine any decision rule π_D that makes up the best response policy for agent i. □

Proposition 10. *In a MAID with sufficient information, if the in-degrees of* $\boldsymbol{D}$ *are bounded and* IS-BEST-RESPONSE *can be solved in poly-time, then a pure NE can be found in poly-time.*

Proof. First, note that we can check whether a MAID is a sufficient information game in poly-time using s-reachability, a graphical criterion based on d-separation [43]. We can then follow the constructive procedure given for the proof of Proposition 3. Given Proposition 7, each optimisation step must take poly-time and since the in-degrees of all decision nodes are bounded by a constant, the number of subgames is also bounded by a constant. Therefore, the entire procedure takes poly-time. □

Joint Behavior and Common Belief

Meir Friedenberg
Department of Computer Science
Cornell University
meir@cs.cornell.edu

Joseph Y. Halpern
Department of Computer Science
Cornell University
halpern@cs.cornell.edu

For over 25 years, common belief has been widely viewed as necessary for joint behavior. But this is not quite correct. We show by example that what can naturally be thought of as joint behavior can occur without common belief. We then present two variants of common belief that can lead to joint behavior, even without standard common belief ever being achieved, and show that one of them, *action-stamped* common belief, is in a sense necessary and sufficient for joint behavior. These observations are significant because, as is well known, common belief is quite difficult to achieve in practice, whereas these variants are more easily achievable.

1 Introduction

The past few years have seen an uptick of interest in studying cooperative AI, that is, AI systems that are designed to be effective at cooperating. Indeed, a number of influential researchers recently argued that "[w]e need to build a science of cooperative AI ... progress towards socially valuable AI will be stunted unless we put the problem of cooperation at the centre of our research" [6].

One type of cooperative behavior is *joint behavior*, that is, collaboration scenarios where the success of the joint action is dependent on all agents doing their parts; one agent deviating can cause the efforts of others to be ineffective. The notion of joint behavior has been studied (in much detail) under various names such as "acting together", "teamwork", "collaborative plans", and "shared plans", and highly influential models of it were developed (see, e.g., [2, 4, 10, 11, 15, 24]). Efforts were also made to engineer some of these theories into real-world joint planning systems [23, 20]. Examples of the types of scenarios these works considered include drivers in a caravan, where if any agent deviates it might lead the entire caravan to get derailed, and a company of military helicopters, where deviation on the part of some agents can lead to the remaining agents being stranded or put in unnecessarily high-risk scenarios.

All the earlier work agrees on the importance of beliefs for this type of cooperation. In particular, because each agent would do her part only if she believed that all of the other agents would do their part as well, there is a widespread claim that *common belief* (often called *mutual belief*) of how the agents would behave was necessary. That is, not only did everyone have to believe all of the agents would act as desired, but everyone had to believe everyone believed it, and everyone had to believe that everyone believed everyone believed it, etc. This, they argued, followed from the fact that everyone acts only if they believe everyone else will. (See, e.g., [2, 4, 10, 11, 15, 24] for examples of this claim.)

As we show in this paper, this conclusion is not quite right. We do not need common belief for joint behavior; weaker variants suffice. Indeed, we provide a variant of common belief that we call *action-stamped* common belief that we show is, in a sense, necessary and sufficient for joint behavior. The key insight is that agents do not have to act simultaneously for there to be joint behavior. If agent 2 acts after agent 1, agent 1 does not have to believe, when he acts, that agent 2 currently believes that all agents will carry out their part of the joint behavior. Indeed, at the point that agent 1 acts, agent 2 might not even be aware of the joint action. It suffices that agent 2 believes *at the point that she carries out her part*

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 221–232, doi:10.4204/EPTCS.379.18

of the joint behavior that all the other agents will believe at the points where they are carrying out their parts of the joint behavior ... that everyone will act as desired at the appropriate time. If actions must occur simultaneously, then common belief is necessary [9]; the fact that we do not require simultaneous actions is what allows us to consider weaker variants of common belief.

Why does this matter? Common belief may be hard to obtain (see [9]); it may be possible to obtain action-stamped common belief in circumstances where common belief cannot be obtained. Thus, if we assume that we need common belief for joint behavior, we may end up mistakenly giving up on cooperative behavior when it is in fact quite feasible.

The rest of the paper is organized as follows. In the next section, we provide the background for the formal (Kripke-structure based) framework that we use throughout the paper. In Section 3, we give our first example showing that agents can have joint behavior without common belief, and define a variant of common belief that we call *time-stamped common belief* which enables it to happen. In Section 4, we give a modified version of the example where time-stamped common belief does not suffice for joint behavior, but *action-stamped common belief*, which is yet more general, does. In general, the group of agents involved in a joint behavior need not be static; it may change over time. For example, we would like to view the firefighters at the scene of a fire as acting jointly, but this group might change over time as additional firefighters arrive and some firefighters leave. In Section 5, we show how action-stamped (and time-stamped) common belief can be extended to deal with the group of agents changing over time. In Section 6, we go into more detail regarding the significance of these results. In Section 7, we show that there is a sense in which action-stamped common belief is necessary and sufficient for joint behavior. Finally, in Section 8, we conclude.

2 Background

To make our claims precise, we need to be able to talk formally about beliefs and time. To do so, we draw on standard ideas from modal logics and the runs-and-systems framework of Fagin et al. [9].

Our models have the form $M = (R, \Phi, \pi, \mathscr{B}_1, \dots, \mathscr{B}_n)$. R is a *system*, which, by definition, is a set of *runs*, each of which describes a way the system might develop over time. Given a run $r \in R$ and a time $n \in \mathbb{N}_{\geq 0}$ (for simplicity, we assume that time ranges over the natural numbers), we call (r,n) a *point* in the model; that is, it describes a point in time in one way the system might develop. Φ is the set of variables. In general, we will denote variables in Φ with uppercase letters (e.g., P) and values of those variables with lowercase ones (e.g., p). π is an *interpretation* that maps each point in the model and each variable $P \in \Phi$ to a value, denoting the value of P at that point. (Thus, the analogue of a primitive proposition for us is a formula of the form $P = p$: variable P takes on value p.) Finally, for each agent i, there is a *binary relation* $\mathscr{B}_i$ over the points in the model. Two points (r_1,n_1) and (r_2,n_2) are related by $\mathscr{B}_i$ (i.e., $(r_1,n_1),(r_2,n_2)) \in \mathscr{B}_i$) if the two points are indistinguishable to agent i; that is, if, at the point (r_1,n_1), agent i cannot tell if the true point is (r_1,n_1) or (r_2,n_2). We assume throughout that the $\mathscr{B}_i$ relations satisfy the standard properties of a belief relation: specifically, they are *serial* (for all points (r,n), there exists a point (r',n') such that $((r,n),(r',n')) \in \mathscr{B}_i$), *Euclidean* (if $((r_1,n_1),(r_2,n_2))$ and $((r_1,n_1),(r_3,n_3))$ are in $\mathscr{B}_i$, then so is $((r_2,n_2),(r_3,n_3)))$, and transitive. These assumptions ensure that the standard axioms for belief hold; see [9] for further discussion of these issues.

To talk about these models, we use the language generated by the following context-free grammar:

$$\varphi := P = p \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid B_i\psi \mid E_G\psi \mid C_G\psi,$$

where P is a variable in Φ, p is a possible value of P, and G is a non-empty subset of the agents. The intended reading of $B_i\psi$ is that agent i believes ψ; for $E_G\psi$ it is that ψ is believed by everyone in the group G; and for $C_G\psi$ it is that ψ is common belief among the group G.

We can inductively give semantics to formulas in this language relative to points in the above models. The propositional operators $\neg$ and $\wedge$ have the standard propositional semantics. The other operators are given semantics as follows:

- $(M,r,n) \models P = p$ if $\pi((r,n),P) = p$,
- $(M,r,n) \models B_i\psi$ if $(M,r',n') \models \psi$ for all points (r',n') such that $((r,n),(r',n')) \in \mathscr{B}_i$,
- $(M,r,n) \models E_G\psi$ if $(M,r,n) \models B_i\psi$ for all $i \in G$
- $(M,r,n) \models C_G\psi$ if $(M,r,n) \models E_G^k\psi$ for all $k \geq 1$, where $E_G^1\psi := E_G\psi$ and $E_G^{k+1}\psi := E_G(E_G^k\psi)$.

There are a number of axioms that are valid in these models. Since they are not relevant for the points we want to make here, we refer the reader to [9] for a discussion of them.

3 Time-Stamped Common Belief

We now give our first example showing that joint behavior does not require common belief. We do not define joint behavior here; indeed, as we said, there are a number of competing definitions in the literature [15, 4, 10, 11]. But we hope the reader will agree that, however we define it, the example gives an instance of it.

> General Y and her forces are standing on the top of a hill. Below them in the valley, the enemy is encamped. General Y knows that her forces are not strong enough to defeat the enemy on their own. She also knows that General Z and his troops, though knowing nothing of the encamped enemy, will arrive on the hill the next day at noon on the way back from a training exercise. Unfortunately though, General Y and her troops must move on before then. Thankfully, all generals are trained for how to deal with this situation. Just as her training recommends, General Y sets up traps that will delay the enemy's retreat, and leaves one soldier behind to inform General Z of the traps upon his arrival. At 11:30 the next morning, General Y receives a (false) message informing her that General Z and his troops have been captured, and thus (incorrectly) surmises that the enemy will live to fight another day. What in fact happens is that General Z's troops arrive at noon and attack the enemy, the enemy attempts to retreat and is stopped by General Y's traps, and the enemy is successfully defeated.

Clearly, Generals Y and Z jointly defeated the enemy. Yet they never achieved common belief of what they were doing. Before noon, General Z didn't even think that the enemy was there, and from 11:30 on, General Y thought that General Z would never arrive. It follows that there was no point at which they could have had common belief. So what is going on here?

What this example suggests is that there are times when a type of *time-stamped common belief* (cf., [9, 12]) suffices to enable joint behavior. Intuitively, on the first day, General Y believed that at noon on the second day General Z would act, attacking the enemy. Similarly, at noon on the second day, General Z believed that General Y had acted the day before, setting up the necessary traps. They also hold higher-order beliefs; for example, at the time she set the traps, general Y believed that at noon the next day general Z would believe that she had set the traps, otherwise she wouldn't have wasted the resources to set them, and so on. Much as in the usual case of common belief, these nested beliefs extend to arbitrary depths. What sets this example apart from those considered by earlier work is that, whereas in the earlier work agents needed to believe others would act as desired *at the same point*, here the agents need to believe only that others will act as desired *at the points where they're supposed to act for the joint behavior*. This suggests that time-stamped common belief can suffice for joint behavior.

We can capture this type of time-stamped common belief formally with the following additions to the logic and semantic models above. Syntactically, we add two more operators to the language, $E^t_G\psi$ and $C^t_G\psi$, where G is a set of agents. We then add to the semantic model a function t that maps each agent and run to a non-negative integer. The intended reading of these is "each agent $i \in G$ believes at the time $t(i,r)$ that ψ" and "it is time-stamped-by-t common belief among the agents in G that ψ", respectively. We give semantics to these operators as follows:

- $(M,r,n) \models E^t_G\psi$ if $(M,r,t(i,r)) \models B_i\psi$ for all $i \in G$
- $(M,r,n) \models C^t_G\psi$ if $(M,r,n) \models E^{t,k}_G\psi$ for all $k \geq 1$, where $E^{t,1}_G\psi := E^t_G\psi$ and $E^{t,k+1}_G\psi := E^t_G(E^{t,k}_G\psi)$.

These definitions are clearly very similar to the (standard) definitions given above for $E_G\psi$ and $C_G\psi$, except that the beliefs of each agent $i \in G$ in run r is considered at the time $t(i,r)$. It follows from the semantic definitions that $E^t_G\psi$ and $C^t_G\psi$ hold at either all points in a run or none of them.

In the example above, this notion of time-stamped common belief *is* achieved if we take $t(Y,r)$ to be the time in run r that Y laid the traps (which may be different times in different runs) and take $t(Z,r)$ to be the time that Z arrived in run r (which was noon in the actual run, but again, may be different times in different runs), provided that it is (time-stamped) common belief that both Y and Z will follow their training. That is, when Y lays the traps, Y must believe that Z will believe when he arrives that Y laid the traps, Z will believe when he arrives that Y believed when she laid the traps that he would believe when he arrived that Y laid the traps, and so on. The key point here is that time-stamped common belief can sometimes suffice for achieving cooperative behavior, even without standard common knowledge.

Our notion of time-stamped common belief is a generalization of (and was inspired by) Halpern and Moses' notion of (time-T) *time-stamped common knowledge*. Roughly speaking, for them, time-T time-stamped common knowledge of ϕ holds among the agents in a group G if every agent i in G knows ϕ at time T on her clock, all agents in G know at time T on their clock that all agents in G know ϕ at time T on their clock, and so on (where T is a fixed, specific time). If it is common knowledge that clocks are synchronized, then time-stamped common knowledge reduces to common knowledge. If we take $t(i,r)$ to be the time in run r that i's clock reads time T (and assume that it is commonly believed that each agent's clock reads time T at some point in every run), then their notion of time-stamped common knowledge becomes a special case of our time-stamped common belief. But note that with time-stamped common belief, we have the flexibility of referring to different times for different agents, and the time does not have to be a clock reading; it can be, for example, the time that an event like laying traps occurs.

4 Action-Stamped Common Belief

There is an even more general variant of common belief that can suffice for joint behavior. What really mattered in the previous example is that everyone had the requisite beliefs at the times that they were acting. But there need not necessarily be only one such point per agent per run; an agent might act multiple times as part of the plan, as the following modified version of the story illustrates:

> General Y and her forces arrive to the south of the town where the enemy forces are encamped. General Y knows that her forces are not strong enough to defeat the enemy on their own. She also knows that General Z and his troops are expected to arrive to the north of the city some time in the near future, though she and her troops must move on before then.

> The swiftly-coursing river prevents the enemy from escaping to the east. But unfortunately, they can still escape inland to the west. Thankfully, all generals are trained for how to deal with this situation as well. Just as her training recommends, General Y sets up traps that will delay the enemy's southward retreat and then, as she heads inland, also sets up traps to the west, finally leaving one soldier behind to go north and inform General Z of the traps upon his arrival. The next morning, General Y receives a (false) message informing her that General Z and his troops have been captured, and thus (incorrectly) surmises that the enemy will live to fight another day. What in fact happens is that General Z's troops arrive later that day and are informed by the remaining soldier that, not too long ago, General Y's troops set traps to the south and west. They attack the enemy, the enemy attempts to retreat and is stopped by General Y's traps, and the enemy is successfully defeated.

Again, Generals Y and Z jointly and collaboratively defeated the enemy, but time-stamped common belief doesn't suffice for this version of the story, because we cannot specify a single time for General Y's actions. Instead, what really matters is that when they are acting as part of a joint plan they hold the requisite (common) beliefs. The joint plan need not be known upfront; General Z does not know what he will need to do to achieve the common goal until he arrives at the scene. To capture this new requirement, we define a notion of *action-stamped common belief.*

We begin by adding a special Boolean variable $ACTING_{i,G}$ for any group G and agent $i \in G$. This variable is true (i.e., takes value 1, as opposed to 0) at a point (r,n) if the agent i is acting towards the group plan of G at (r,n) and false otherwise. So for the generals, $ACTING_{Y,G} = 1$ would be true when she lays the traps, $ACTING_{Z,G} = 1$ would be true at the point when he attacks, and they'd both be false otherwise (where $G = \{Y,Z\}$). We often write $ACTING_{i,G}$ and $\neg ACTING_{i,G}$ instead of $ACTING_{i,G} = 1$ and $ACTING_{i,G} = 0$, and similarly for other Boolean variables. By using $ACTING_{i,G}$, we can abstract away from what actions are performed; we just care that some action is performed by agent i towards the group plan, without worrying about what that action is.

As in the case of time-stamped common belief, we add two modal operators to the language (in addition to the variables $ACTING_{i,G}$). Let G be a set of agents. $E^{\mathbf{a}}_G\psi$ then expresses that, for each agent $i \in G$, whenever $ACTING_{i,G}$ holds (it may hold several times in a run, or never), i believes ψ. $C^{\mathbf{a}}_G\psi$ then defines the corresponding notion of common belief for the points at which agents act at part of the group.

We give semantics to these modal operators as follows:

- $(M,r,n) \models E^{\mathbf{a}}_G\psi$ if for all n' and all $i \in G$ such that $(M,r,n') \models ACTING_{i,G} = 1$, it is also the case that $(M,r,n') \models B_i\psi$.
- $(M,r,n) \models C^{\mathbf{a}}_G\psi$ if $(M,r,n) \models E^{\mathbf{a},k}_G\psi$ for all $k \geq 1$, where $E^{\mathbf{a},1}_G\psi := E^{\mathbf{a}}_G\psi$ and $E^{\mathbf{a},k+1}_G\psi := E^{\mathbf{a}}_G(E^{\mathbf{a},k}_G\psi)$.

Returning to the example, although the agents do not have time-stamped common belief at all the points when they act, they do have action-stamped common belief. General Z acted believing that General Y had acted as expected, and also believing that General Y acted believing that he would act as expected, and so on.

It is easy to see that time-stamped common belief can be viewed as a special case of action-stamped common belief: Given a time-stamping function t, we simply take $ACTING_{i,G}$ to be true at those points (r,n) such that $t(i,r)$ holds.

It is worth noting that, in both this and the previous section, the agents having a protocol in advance for how to deal with the situation is not really necessary for them to succeed. In the examples, consider a scenario where generals are in fact not trained for how to handle the situation, but instead General Y has the brilliant idea to lay traps and send a messenger to meet General Z upon arrival. As long as message delivery is reliable, action-stamped common belief can be achieved and they can successfully defeat the enemy.

5 Joint Behaviors Among Changing Groups

In practice, the members of groups change over time. For example, a group of firefighters may work together to safely clear a burning building, but (thankfully!) they don't need to wait until all the firefighters are on the scene, or even until it is known which firefighters are coming, in order for the first firefighters to begin. Instead, structures and guidelines allow the set of firefighters who are on the scene to act cooperatively, even without each firefighter knowing who else will show up.

The formalisms of the two previous sections assumed a fixed group G, so cannot capture this kind of scenario. But the changes necessary to do so are not complicated. Rather than considering (some variant of) common belief with respect to a fixed set G of agents, we consider it with respect to an *indexical* set S, one whose interpretation depends on the point. More precisely, an indexical set S is a function from points to sets of agents; intuitively, $S(r,n)$ denotes the members of the indexical group S at the point (r,n). We assume that a model is extended so as to provide the interpretation of S as a function.

Our semantics for action-stamped common belief with indexical sets are now a straightforward generalization of the semantics for rigid (non-indexical) sets:

- $(M,r,n) \models E_S^{\mathbf{a}}\psi$ if for all n' and all $i \in S(r,n')$ such that $(M,r,n') \models ACTING_{i,S}$, it is also the case that $(M,r,n') \models B_i\psi$.
- $(M,r,n) \models C_S^{\mathbf{a}}\psi$ if $(M,r,n) \models E_S^{\mathbf{a},k}\psi$ for all $k \geq 1$, where $E_S^{\mathbf{a},1}\psi := E_S^{\mathbf{a}}\psi$ and $E_S^{\mathbf{a},k+1}\psi := E_G^{\mathbf{a}}(E_S^{\mathbf{a},k}\psi)$.

The only change here is that in the semantics of $E_S^{\mathbf{a}}$, we need to check the agents in $S(r,n)$ for each point.

Of course, we can also allow indexical sets in time-stamped common belief in essentially the same way. Whereas in the semantics of $E_G^t\psi$, we required that $(M,r,n) \models E_G^t\psi$ if, for all $i \in G$, $(M,r,t(i,r)) \models B_i\psi$, now we require that $(M,r,n) \models E_S^t\psi$ if, for all agents i, if $i \in S(r,t(i,r))$, then $(M,r,t(i,r)) \models B_i\psi$. We care about what agent i believes at $(M,r,t(i,r))$ only if i is actually in group S at the point $(r,t(i,r))$.

6 Significance

In Sections 3-5 we showed that action-stamped common belief can suffice to enable joint behavior, whereas the prior work on the topic had assumed common belief was necessary. Why does this matter? We argue that it is important for two reasons: 1) misunderstanding the type of belief necessary can lead to mis-evaluation of cooperative capabilities, and 2) requiring common belief can unnecessarily make cooperation impossible in scenarios where it is in fact possible and could be quite beneficial.

As part of the recent push for more research on cooperative AI, some have argued that we should "construct more unified theory and vocabulary related to problems of cooperation" [7]. One important step in this program is (in our opinion) formalizing the requirements for various types of cooperation, including joint behavior. This, in turn, requires understanding the level and type of (common) belief needed for joint behavior. As our examples have shown, full-blown common belief is not necessary; weaker variants that are often easier to achieve can suffice. Relatedly, there has been a push to develop methods for *evaluating* the cooperative capabilities of agents, as a way of developing targets and guideposts for the community [5]. Again, this will require understanding (among other things) what type of beliefs are necessary for cooperation. Incorrect assumptions about the types of beliefs necessary can lead to incorrect conclusions about the feasibility of cooperation. For example, if an evaluation system takes as given the assumption that it is impossible for agents that cannot achieve common belief to behave cooperatively, it may in fact lead to effective cooperative agents being scored badly, leading to misdirected research.

A second reason that it is important to clarify the types of beliefs necessary for joint behavior is that misunderstanding them can lead to systems unnecessarily aborting important cooperative tasks. As is well known, achieving true common knowledge can be remarkably difficult in real-world systems, often requiring either a communication system that guarantees truly synchronous delivery or guaranteed bounded delivery time together with truly synchronized clocks [9]. Action-stamped common belief can sometimes be achieved when common belief cannot. To demonstrate the importance of this, we consider an example from the domain of urban search and rescue, a domain where 1) the use of multi-agent systems consisting of humans and AI agents has long been considered and advocated for, 2) the types of teamwork necessary can be complex, and 3) there is some evidence of potential adoption, having been used, for example, at a small scale in the aftermath of September 11th [3, 14, 13, 16, 19, 22]. Though the example we give is a simple, stylized case, the domain is sufficiently complex that we would expect these types of issues to arise in practice if systems were deployed at scale.

Example 1. *An earthquake occurs, causing a large building to collapse. The nearest search and rescue team arrives on scene, and the incident commander has to decide how to proceed. The team has determined that the structure is stable and will not collapse, and so is safe to enter. However, attempting to exit the building may disrupt the structure and cause harm. The incident commander determines that there are two reasonable options:*

1. *Wait a week for a heavy piece of machinery that will certainly be able to safely lift the roof of the collapsed building on its own and allow rescuers safe access to the building.*
2. *A team can enter the building and restabilize parts of the roof. The restabilization would not be enough to make it safe to exit—in fact, it would require adjusting the structure in ways that would make an attempt to exit even more risky—but it would allow a more easily accessible robotic system to safely remove the roof piece by piece, allowing the rescuers and anyone trapped inside to safely escape.*

The incident commander decides it is best not to wait, and so takes the second, joint-behavior-based approach. He sends the team of rescuers in to begin the necessary process, and tells them the full plan and that he expects it will be 2-3 hours before the robot arrives on scene. The group enters the wreckage and secures it in the necessary ways, as planned. But it turns out that the earthquake affected many buildings, so the robot is in high demand. It ends up taking close to 8 hours for the robot to arrive on scene. When the robot arrives, the incident commander enters the relevant information in the robot's system—namely, the full plan and that the restabilization has been carried out—so the robot can carry out its part of the specified cooperative plan.

If the robot's model of joint behavior requires common belief, a problem will arise. At no point is there ever common belief of the joint behavior. Before the robot arrives, the robot certainly has no belief about the joint behavior. And when the robot arrives, it must consider the possibility that, because of the delay, the rescuers have given up hope of the robot arriving and concluded that they may have to wait a full week until the larger piece of machinery is available. Even if this isn't actually the case, because the robot considers it possible, common belief will not be achieved. So if the robot assumes that joint behavior requires common belief, it will determine that the joint behavior cannot be carried out. Thus, everyone will have to wait a week for the heavier machinery, risking the lives of anyone trapped inside.

If, on the other hand, the robot's theory of joint behavior is based on action-stamped common belief, the task will be properly and safely carried out as soon as the robot arrives: When the rescuers perform *their part, they believe that the robot will arrive soon and perform its part of the task. Similarly, when the robot arrives and the incident commander enters the relevant information, the robot believes that the rescuers held those beliefs when acting (and therefore performed the required adjustments). The rescuers believed that the robot would hold these beliefs when it arrived, the robot believed they would, and so on. The fact that the robot arrived later than expected and the rescuers may have started to have uncertainty about the plan doesn't affect the requisite beliefs because all that matters are the beliefs of the agents at the points where they act.*

This example highlights the value of getting the types of beliefs necessary right; getting the theory right, and basing it on action-stamped common belief instead of standard common-belief, can enable cooperation in a range of important scenarios where standard common belief is impossible or difficult to achieve, whereas action-stamped common belief may be easily attainable.

7 On the Necessity and Sufficiency of Action-Stamped Common Belief for Joint Behavior

We've argued in this paper that the prior work was incorrect in asserting that common belief was necessary for joint behavior, and shown by example that action-stamped common belief can suffice. We now argue that an even stronger statement is true: there is a sense in which action-stamped common belief is necessary and sufficient for joint behavior. We say "in a sense" here, because much depends on the conception of joint behavior being considered. So what we do in this section is give a property that we would argue is one that we would want to hold of joint behavior, and then show that action-stamped common belief is necessary and sufficient for this property to hold.

What does it take to go from a collection of individual behaviors to a joint behavior? The following example may help illuminate some of the relevant issues.

> Jasper and Horace are both crooks, though neither is an evil genius by any stretch of the imagination. They both independently decide to rob the Great Bank of London on exactly the same day. As it turns out, neither of them did a good job preparing, and they each knew about only half of the bank's security systems, and so made plans to bypass only that half. By sheer dumb luck, between them they know about all the bank's security systems. So when each bypasses the part that they know about (at roughly the same time), the bank's security systems go down. They each make it in, steal a small fortune, and escape, none the wiser as to the other's behavior or that their plan was doomed to fail on its own.

Is Jasper and Horace robbing the bank an instance of joint behavior? We think not. One critical component that distinguishes this from a joint behavior is the beliefs of the agents. Joint behaviors are collective actions where people do their part because they believe that everyone else will do their part as well. Here, Jasper and Horace have no inkling that the other will help disable the system.

We now want to capture these intuitions more formally. We start by adding another special Boolean variable $SHOULD\_ACT_{i,S}$ for each agent i and indexical group S, specifying the points in each run where agent i is supposed to act towards the plan of group S. We then add a special formula χ_S to the language:[1]

- $(M,r,n) \models \chi_S$ if for all n' and all agents $i \in S(r,n')$, $(M,r,n') \models SHOULD\_ACT_{i,S} \rightarrow ACTING_{i,S}$

[1] As long as the set of agents is finite (which we implicitly assume it is), we can express χ_S in a language that includes a standard modal operator $\boxdot$, where $\boxdot\varphi$ is true at a point (r,n) iff φ is true at all points (r,n') in the run. For ease of exposition, we do not introduce the richer modal logic here.

The formula χ_S is thus true at a point (r,n) if, at all points in run r, each agent i in the indexical group S plays its part in the group plan whenever it is supposed to.

If we think of $ACTING_{i,S}$ as "i is taking part in the joint behavior of the group S", then the property JB$_S$ (for "Joint Behavior, group S") that we now specify essentially says that to have truly joint behavior, each agent in S must believe when she acts that all of the members of the (indexical) group S will do what they're supposed to; if they don't all have that belief, then it's not really joint behavior. Formally, JB$_S$ is a property of an indexical group S in a model M:

[JB$_S$:] For all points (r,n) and agents $i \in S(r,n)$, $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\chi_S$.

Requiring JB$_S$ for joint behavior makes action-stamped common belief of χ_S necessary for joint behavior.

Theorem 7.1. *If JB$_S$ holds in a model M, then $(M,r,n) \models C^a_S\chi_S$ for all points (r,n).*

Proof. We begin by defining a notion of *a-reachability*: A point (r',n') is S-a-reachable from (r,n) in k steps if there exists a sequence $(r_0,n_0),\dots,(r_k,n_k)$ of points such that $(r_0,n_0) = (r,n)$, $(r_k,n_k) = (r',n')$, and for all $0 \le l < k$, there exists a point (r_l,n'_l) and an agent $i \in S(r_l,n'_l)$ such that $(M,r_l,n'_l) \models ACTING_{i,S}$ and $((r_l,n'_l),(r_{l+1},n_{l+1})) \in \mathscr{B}_i$.

By the semantics of C^a_S, $C^a_S\chi_S$ holds at (r,n) iff χ_S holds at every point (r',n') that is S-a-reachable from (r,n) in 1 or more steps. Consider any such point (r',n'). Then, by the definition of reachability, there exists some point (r'',n'') and some agent $i \in S(r'',n'')$ such that $(M,r'',n'') \models ACTING_{i,S}$ and $((r'',n''),(r',n')) \in \mathscr{B}_i$. Because $(M,r'',n'') \models ACTING_{i,S}$, we get by JB$_S$ that $(M,r'',n'') \models B_i\chi_S$. Then by the semantics of B_i and the fact that $((r'',n''),(r',n')) \in \mathscr{B}_i$ we get that $(M,r',n') \models \chi_S$. But (r',n') was an arbitrary point S-a-reachable from (r,n) in 1 or more steps, so χ_S holds at all such points, and we have that $(M,r,n) \models C^a_S\chi_S$. But (r,n) was also arbitrary, so $C^a_S\chi_S$ holds at all points. □

The converse to Theorem 7.1 also holds; that is, action-stamped common belief of χ_S suffices for JB$_S$ to hold. Put another way, action-stamped common belief is exactly the ingredient that we need to meet the belief requirements of the property that we used to characterize joint behavior.

Theorem 7.2. *If $(M,r,n) \models C^a_S\chi_S$ for all points (r,n), then JB$_S$ holds in M.*

Proof. Consider an arbitrary point (r,n) and agent $i \in S(r,n)$ such that $(M,r,n) \models ACTING_{i,S}$. By assumption, $(M,r,n) \models C^a_S\chi_S$. So, by the semantics of C^a_S, it follows that $(M,r,n) \models E^a_S\chi_S$. In turn, it follows from the semantics of E^a_S that $(M,r,n) \models B_i\chi_S$ (because $(M,r,n) \models ACTING_{i,S}$). But r, n, and i were arbitrary, so we have that $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\chi_S$ for all such points and agents. Thus, JB$_S$ holds in M. □

The astute reader will have noticed that the proofs of Theorem 7.1 and 7.2 did not depend in any way on χ_S. The formula χ_S in these theorems can be replaced by an arbitrary formula φ. In other words, if all the agents in S believe φ at the point when they act, then φ is action-stamped common belief, and if φ is action-stamped common belief, then all agents in S must believe φ at the point when they act. Formally, the proofs of Theorem 7.1 and 7.2 also show the following:

Theorem 7.3. *If $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\varphi$ for all points (r,n) and agents $i \in S(r,n)$, then $(M,r,n) \models C^a_S\varphi$ for all points (r,n).*

Theorem 7.4. *If $(M,r,n) \models C^a_S\varphi$ for all points (r,n), then $(M,r,n) \models ACTING_{i,S} \rightarrow B_i\varphi$ for all points (r,n) and agents $i \in S(r,n)$.*

8 Conclusion and Future Work

We have argued here that, contrary to what was suggested in earlier work, common belief is not necessary for joint behavior. We have presented a new notion, *action-stamped* common belief, and shown that it is, in a sense, necessary and sufficient for joint behavior, and can be achieved in scenarios where standard common belief cannot. This is important because modelling the conditions needed for joint behavior correctly can enable cooperation in important scenarios, such as search and rescue, where it might not otherwise be possible. We chose to use the term *joint behavior* in this paper because it sounded to our ears like it most accurately captured the notion we were considering; no doubt to some readers other terms will sound like a better fit. As we showed in Section 7, action-stamped common belief characterizes scenarios where individuals do their part only if they believe others will do the same, whatever terminology we use.

We suspect that, for some readers, the idea that action-stamped common belief is sufficient for joint behavior will seem obvious. In a certain sense, we agree; in retrospect, it *does* feel like the obviously correct notion for joint behaviors. That said, while action-stamped common belief seems quite natural, it does not seem to have been studied in any prior literature.

With that in mind, it is worth briefly discussing the connection between the ideas in this paper and some of the prior work that has been done. First, note that action-stamped common belief is in some ways the natural variant of common belief for extensive-form games. Because an agent i's information sets are usually specified only at nodes at which agent i moves, it is possible to reason about agent i's beliefs only at points where agent i acts. This makes it all the more surprising that action-stamped common belief has not been formalized and studied in its own right; in some sense, it captures what epistemic game theorists have been implicitly considering in the case of extensive-form games.

In this paper, we have considered the types of *beliefs* necessary for joint behavior, but that may not be the only factor involved (nor do we claim it is; we are just focused in this paper on the belief component). For example, in much of the literature, *intent* is taken to play an important role in various cooperative behaviors. Dunin-Keplicz and Verbrugge [8] proposed a three-part notion of "collective commitment", with the levels of belief (e.g., no one believes, everyone believes, it is common belief) held at each of the three parts leading to various types of collective commitment. Their work is in some ways orthogonal to ours; it can be thought of as considering various types of cooperative behaviors that can occur, while ours just focuses on one particularly strict form, joint behavior. One way of interpreting our work in the context of theirs is as saying that the top level of belief to consider for cooperation should in fact be action-stamped common belief.

Blomberg [1] gives an insightful argument that common belief (and variants thereof) of intentions is not necessary for a joint intentional act. Roughly speaking, an agent may (incorrectly) believe that other agents do not share his intent, as long as what he believes they intend would still lead them to act in the manner conducive to his goals. We find his counterexample and arguments compelling. But this is perfectly consistent with our results. Theorems 7.1 and 7.2 show that action-stamped common belief (or in the case of simultaneous acts, standard common belief) that agents will do the necessary acts is required for joint behaviors. We place no requirements on what agents have to believe about other agents' intentions. Put a different way, Blomberg makes a compelling case that, when characterizing joint behavior, it is a mistake to instantiate the φ in Theorems 7.3 and 7.4 with formulas about shared intents. That is to say, it is not a necessary property of cooperative behavior that agents act only if they are sure others are acting with the same purpose.

Ludwig [17, 18] also presents an argument that common belief is not necessary for joint (intentional) action. Putting aside the question of whether his argument is correct, it is not relevant to our considerations here as it relies on a much broader notion of cooperative behavior than the joint behaviors we consider in this paper (though he calls it "joint action"). That is, we certainly agree with his conclusion that for *some* types of cooperative behavior agents don't need to be sure that others will do their part and so don't need common belief, but for the types of cooperative behaviors we consider in this paper we have shown in Theorem 7.3 that (action-stamped) common belief is in fact necessary.

Lastly, Roy and Schwenkenbecher [21] consider a novel notion of belief that they call "pooled knowledge", which is related to distributed knowledge, and argue that it is both weaker than common knowledge and sufficient for shared intentions. The basic idea behind the argument is that if agents are rational, then pooled knowledge would induce some agent to act as a coordinator to guide the behavior of the group. It's certainly an interesting proposal, and one that deserves further study. From the perspective of this paper, it would be interesting to try to formally analyze under what conditions pooled knowledge/belief would lead to action-stamped common knowledge/belief.

The present work suggests two areas that are ripe for future work. The first is to more fully explore the logical aspects of action-stamped common belief. Can a sound and complete axiomatization be provided? What is the complexity of the model-checking and validity problems for a language involving action-stamped common knowledge? How can we practically engineer systems that rely on action-stamped common belief? The second area we think worth exploring is that of understanding better what levels of group knowledge are required for other aspects of joint behavior and other types of cooperation. We focused on one aspect, revealing a nuanced but important error in earlier thinking. We think that there may well be other aspects of cooperation that are worth digging into in this fine-grained way. Given the importance of cooperative AI, we hope that others will join us in exploring these questions.

Acknowledgments

Meir Friedenberg and Joe Halpern were supported in part by ARO grant W911NF-22-1-0061 and MURI grant W911NF-19-1-0217. Halpern was also supported in part by a AFOSR grant FA9550-12-1-0040 and a grant from the Algorand Centres of Excellence program managed by Algorand Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

[1] O. Blomberg (2016): *Common knowledge and reductionism about shared agency*. *Austalasian Journal of Philosophy* 94(2), pp. 315–326, doi:10.1080/00048402.2015.1055581.

[2] M. E. Bratman (1992): *Shared cooperative activity*. *The Philosophical Review* 101(2), pp. 327–341, doi:10.2307/2185537.

[3] J. Casper & R. R. Murphy (2003): *Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center*. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 33(3), pp. 367–385, doi:10.1109/TSMCB.2003.811794.

[4] P. R. Cohen & H. J. Levesque (1991): *Teamwork*. *Nous* 25(4), pp. 487–512, doi:10.2307/2216075.

[5] Cooperative AI Foundation (CAIF) (2022): *Evaluation for Cooperative AI: Call for Proposals*. https://www.cooperativeai.com/calls-for-proposals/evaluation-for-cooperative-ai; accessed October. 25, 2022.

[6] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson & T. Graepel (2021): *Cooperative AI: machines must learn to find common ground*.

[7] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson & T. Graepel (2020): *Open problems in cooperative AI.* Available at https://arxiv.org/pdf/2012.08630.pdf.

[8] B. Dunin-Kęplicz & R. Verbrugge (2004): *A tuning machine for cooperative problem solving*. Fundamenta Informaticae 63(2–3), pp. 283–307.

[9] R. Fagin, J. Y. Halpern, Y. Moses & M. Y. Vardi (1995): *Reasoning About Knowledge*. MIT Press, Cambridge, MA. A slightly revised paperback version was published in 2003.

[10] B. J. Grosz & S. Kraus (1996): *Collaborative plans for complex group action*. Artificial Intelligence 86(2), pp. 269–357, doi:10.1016/0004-3702(95)00103-4.

[11] B. J. Grosz & C. L. Sidner (1990): *Plans for discourse*. In P. R. Cohen, J. L. Morgan & M. E. Pollack, editors: Intentions in Communication, chapter 20, MIT Press.

[12] J. Y. Halpern & Y. Moses (1990): *Knowledge and common knowledge in a distributed environment*. Journal of the ACM 37(3), pp. 549–587, doi:10.1145/79147.79161.

[13] H. Kitano & S. Tadokoro (2001): *Robocup rescue: A grand challenge for multiagent and intelligent systems*. AI magazine 22(1), pp. 39–39, doi:10.1609/aimag.v37i1.2642.

[14] H. Kitano, S. Tadokoro, I. Noda, H. Matsubara, T. Takahashi, A. Shinjou & S. Shimada (1999): *Robocup rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research*. In: *1999* IEEE International Conference on Systems, Man, and Cybernetics, 6, pp. 739–743.

[15] H. J. Levesque, P. R. Cohen & J. H. T. Nunes (1990): *On acting together*. In: Proc. Eighth National Conference on Artificial Intelligence (AAAI '90), pp. 94–99.

[16] Y. Liu & G. Nejat (2013): *Robotic urban search and rescue: A survey from the control perspective*. Journal of Intelligent & Robotic Systems 72(2), pp. 147–165, doi:10.1007/s10846-013-9822-x.

[17] K. Ludwig (2007): *Collective intentional behavior from the standpoint of semantics*. Noûs 41(3), pp. 355–393, doi:10.1111/j.1468-0068.2007.00652.x.

[18] K. Ludwig (2016): *From Individual to Plural Agency: Collective Action*. Oxford University Press, doi:10.1093/acprof:oso/9780198755623.001.0001.

[19] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju & T. Westerlund (2020): *Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision*. IEEE Access 8, pp. 191617–191643, doi:10.1109/ACCESS.2020.3030190.

[20] C. Rich & C. L. Sidner (1997): *COLLAGEN: When agents collaborate with people*. In: Proceedings of the First International Conference on Autonomous Agents, pp. 284–291, doi:10.1145/267658.267730.

[21] O. Roy & A. Schwenkenbecher (2021): *Shared intentions, loose groups, and pooled knowledge*. Synthese 198(5), pp. 4523–4541, doi:10.1007/s11229-019-02355-x.

[22] J. Scholtz, J. Young, J. L. Drury & H. A. Yanco (2004): *Evaluation of human-robot interaction awareness in search and rescue*. In: Proc. IEEE International Conference on Robotics and Automation, 2004 (ICRA'04), 3, pp. 2327–2332, doi:10.1109/ROBOT.2004.1307409.

[23] M. Tambe (1997): *Towards flexible teamwork*. Journal of A.I. Research 7, pp. 83–124.

[24] R. Tuomela (2005): *We-intentions revisited*. Philosophical Studies 125(3), pp. 327–369, doi:10.1007/s11098-005-7781-1.

Presumptive Reasoning in a Paraconsistent Setting

Sabine Frittella Daniil Kozhemiachenko
INSA Centre Val de Loire, Univ. Orléans, LIFO EA 4022, France*
{sabine.frittella,daniil.kozhemiachenko}@insa-cvl.fr

Bart Verheij
Bernoulli Institute, Rijksuniversiteit Groningen
Groningen, the Netherlands†
bart.verheij@rug.nl

We explore presumptive reasoning in the paraconsistent case. Specifically, we provide semantics for non-trivial reasoning with presumptive arguments with contradictory assumptions or conclusions. We adapt the case models proposed by Verheij [25, 26] and define the paraconsistent analogues of the three types of validity defined therein: coherent, presumptively valid, and conclusive ones. To formalise the reasoning, we define case models that use $\mathsf{BD}\triangle$, an expansion of the Belnap–Dunn logic with the Baaz Delta operator. We also show how to recover presumptive reasoning in the original, classical context from our paraconsistent version of case models. Finally, we construct a two-layered logic over $\mathsf{BD}\triangle$ and biG (an expansion of Gödel logic with a coimplication $\prec$ or $\triangle$) and obtain a faithful translation of presumptive arguments into formulas.

1 Introduction

When arguing for a given statement, it can happen that a person uses contradictory assumptions. From the classical standpoint, every statement trivially follows from a contradiction. This, however, is counterintuitive as an agent may not be willing to accept a completely arbitrary statement just because their premises contain a contradiction.

In general, an argument from ϕ to χ (written formally as $\langle\phi,\chi\rangle$) can be either *deductive* (when ϕ *entails* χ) or *presumptive* (otherwise). I.e., to verify the correctness of a deductive argument, it suffices to utilise purely logical means while establishing the correctness (acceptability) of a presumptive one requires an extra-logical framework. Thus, from *the classical standpoint*, every argument from a contradictory premise is *deductive*. Hence, if one wants to formalise *non-trivial* presumptive reasoning from contradictory premises, one has to use a *paraconsistent* logic, i.e., a logic where the explosion principle $p,\neg p\models q$ is not valid.

Dung's argumentative semantics vs case models An influential approach to the formalisation of argumentation focuses on argument attack [11]. The main idea is to represent the argumentative framework as a directed graph where $\mathscr{A}\rightarrow\mathscr{B}$ is interpreted as 'argument $\mathscr{A}$ attacks $\mathscr{B}$' (here, arguments are treated as unified statements, and premises and conclusions are not singled out). Then, $\mathscr{A}$ is *acceptable* if it responds to every attack (or, formally, if $\mathscr{A}\rightarrow\mathscr{B}$ for every $\mathscr{B}$ s.t. $\mathscr{B}\rightarrow\mathscr{A}$). An argument's correctness depends on the argumentation semantics choice.

However, the connection of Dung's approach to standard logical semantics may not be straightforward. In addition, the support of arguments is abstracted. Both issues have been addressed in several ways (cf., e.g. [9, 5, 14, 21]). One such alternative to Dung's approach was proposed in [25] and further

[*] The research of Sabine Frittella and Daniil Kozhemiachenko was funded by the grant ANR JCJC 2019, project PRELAP (ANR-19-CE48-0006).

[†] Bart Verheij participates in the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 233–244, doi:10.4204/EPTCS.379.19

developed in [26]. In these works, the interpretation of presumptive arguments was given by means of *case models*: sets of classically incompatible satisfiable propositional formulas called 'cases' (whence the name) with a preference relation defined thereon. An argument in this framework has the following form: $\mathscr{A} = \langle\phi,\chi\rangle$. Here, ϕ is the premise, χ is the conclusion, and, in addition, $\mathscr{A}$ presents a *case* — $\phi\wedge\chi$. Three kinds of acceptable arguments were studied: coherent (both the premise and the conclusion are supported by at least one case), presumptively valid (both the premise and conclusion are supported by the most preferred case), and conclusive (the conclusion is supported by all cases that support the premise). Furthermore, a representation of case models in terms of sample spaces with probability measures was devised and a correspondence between different arguments and probabilities of the corresponding events was provided. An important distinction between case models and Dung's semantics is that the validities in the former are defined via the entailment in the classical logic. Thus, one can produce a non-classical counterpart to case models by changing the underlying entailment relation.

Non-trivial contradictory arguments In both approaches discussed above, it is assumed that the acceptable arguments are not *self-contradictory*. Namely, if ϕ is an argument in Dung's framework, it should be classically satisfiable, and if $\langle\chi,\psi\rangle$ is an argument over a case model, then both χ and ψ must be classically satisfiable. This restriction is easy to explain in Dung's approach: indeed, we can claim that a contradictory argument attacks itself. In the case model setting, however, it makes sense to consider contradictory arguments and cases under the following interpretation.

Every 'case' in the model can be thought of as a source that gives some information regarding a given set of statements. Accordingly, the preference relation on cases shows which sources are trusted more or less. In this interpretation, it is clear that even if a source is trusted, it can provide a contradictory response to a question (e.g., a police officer testifying in court can first claim that they were unarmed while on patrol and then say 'when I saw the suspect, I immediately drew my pistol out of the holster') or fail to provide any information at all.

Let us now introduce the running example to illustrate the contexts that we aim to formalise.

Running example, part 1 (Witnessing a robbery). An investigator reads a report by a police officer who questioned several witnesses on a bank robbery. The relevant information is whether the perpetrator had a limp (l), whether they had a big bag for the robbed valuables (b), and whether they used the lift or the staircase (s) to leave the office. The report contains the following testimonies.

- c_1 tells that the perpetrator indeed had a limp but cannot say anything about how they left the office; moreover, c_1 tells that the robber put all the loot in the pockets.
- c_2 tells nothing about whether the perpetrator was limping and mentions that the perpetrator had a big shoulder bag; unfortunately, c_2 is confused: they claim that they saw the robber using the lift but are also saying that 'the lift has been out of order for half a year'.
- c_3 testifies that the robber had a limp but walked down the stairs; c_3's account is also contradictory: they describe the bag as 'huge' but then say that the robber put it into the pocket.

All witnesses gave non-classical (incomplete or contradictory) responses, whence we cannot straightforwardly represent them in the case models, nor in Dung's framework. An investigator, however, needs to draw conclusions from the accounts at hand. E.g., they might want to know how the perpetrator in fact left the building and for that, they need to know whether the perpetrator had a limp.

Plan of the paper In this paper, we adapt the case models presented in [25, 26] to the presumptive reasoning with possibly contradictory statements. To this end, we will use $\mathsf{BD}\triangle$ — the expansion of the Belnap–Dunn logic [1, 12, 4] with a Baaz $\triangle$ operator (cf. [2] for the original presentation of $\triangle$ in the context of fuzzy logics) originating from [24]. We define the analogues of coherent, presumptively valid, and conclusive arguments and show their relations to one another. Finally, we are going to provide a logical representation of all these arguments.

The remainder of the paper is structured as follows. In Section 2, we present the syntax and semantics of $\mathsf{BD}\triangle$. Then, in Section 3, we develop the $\mathsf{BD}\triangle$ case models. In Section 4, we present a logic that formalises reasoning in $\mathsf{BD}\triangle$ models and construct a faithful translation of arguments to formulas. Finally, in Section 5, we summarise the results and provide a plan for future research.

2 Logical preliminaries

The language of $\mathscr{L}_{\mathsf{BD}\triangle}$ and its $\triangle$-free fragment $\mathscr{L}_{\mathsf{BD}}$ are defined via the following grammar ($\mathtt{Prop}$ is a fixed countable set of propositional variables).

$$\mathscr{L}_{\mathsf{BD}\triangle} \ni \phi \coloneqq p \in \mathtt{Prop} \mid \neg\phi \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid \triangle\phi$$

There are several semantics for BD and its expansions (cf. [20] for the examples). One of the simplest is a truth-table semantics from [4]. There, a formula can have one of the following four values corresponding to the available information regarding a statement ϕ. **T** stands for ‘there is only information in support of ϕ’; **F** for ‘there is only information denying ϕ’; **N** for ‘there is information neither in support nor in denial of ϕ’; **B** for ‘there is information both in support and denial of ϕ’. We also use frame semantics (cf., e.g., [8, 16, 7]) as it is more convenient for the logical representation of case models.

Definition 1 (Truth-table semantics of $\mathsf{BD}\triangle$)**.** A **4**-valuation is a map $v_{\mathbf{4}} : \mathtt{Prop} \to \{\mathbf{T}, \mathbf{B}, \mathbf{N}, \mathbf{F}\}$ that is extended to complex formulas using the following definitions.

$\neg$	
T	**F**
B	**B**
N	**N**
F	**T**

$\triangle$	
T	**T**
B	**T**
N	**F**
F	**F**

$\wedge$	**T**	**B**	**N**	**F**
T	**T**	**B**	**N**	**F**
B	**B**	**B**	**F**	**F**
N	**N**	**F**	**N**	**F**
F	**F**	**F**	**F**	**F**

$\vee$	**T**	**B**	**N**	**F**
T	**T**	**T**	**T**	**T**
B	**T**	**B**	**T**	**B**
N	**T**	**T**	**N**	**N**
F	**T**	**B**	**N**	**F**

Definition 2 (Frame semantics for $\mathsf{BD}\triangle$)**.** Let $\phi, \phi' \in \mathscr{L}_{\mathsf{BD}\triangle}$. For a model $\mathfrak{M} = \langle W, v^+, v^- \rangle$ with $v^+, v^- : \mathtt{Prop} \to 2^W$, we define notions of $w \vDash^+ \phi$ and $w \vDash^- \phi$ for $w \in W$ as follows.

$$\begin{array}{rll@{\qquad}rll}
w \vDash^+ p & \text{iff} & w \in v^+(p) & w \vDash^- p & \text{iff} & w \in v^-(p) \\
w \vDash^+ \neg\phi & \text{iff} & w \vDash^- \phi & w \vDash^- \neg\phi & \text{iff} & w \vDash^+ \phi \\
w \vDash^+ \phi \wedge \phi' & \text{iff} & w \vDash^+ \phi \text{ and } w \vDash^+ \phi' & w \vDash^- \phi \wedge \phi' & \text{iff} & \mathrm{w}\vDash^- \phi \text{ or } w \vDash^- \phi' \\
w \vDash^+ \phi \vee \phi' & \text{iff} & w \vDash^+ \phi \text{ or } w \vDash^+ \phi' & w \vDash^- \phi \vee \phi' & \text{iff} & w \vDash^- \phi \text{ and } w \vDash^- \phi' \\
w \vDash^+ \triangle\phi & \text{iff} & w \vDash^+ \phi & w \vDash^- \triangle\phi & \text{iff} & w \nvDash^+ \phi
\end{array}$$

We define the *positive* and *negative interpretations of* ϕ as follows: $|\phi|^+ = \{w \in W \mid w \vDash^+ \phi\}$; $|\phi|^- = \{w \in W \mid w \vDash^- \phi\}$.

We say that a sequent $\phi \vdash \chi$ is *satisfied on* $\mathfrak{M}$ (denoted, $\mathfrak{M} \models [\phi \vdash \chi]$) iff $|\phi|^+ \subseteq |\chi|^+$ and $|\chi|^- \subseteq |\phi|^-$. $\phi \vdash \chi$ is *valid* iff it is satisfied on every model. In this case, we say that ϕ *entails* χ and write $\phi \models_{\mathsf{BD}\triangle} \chi$.

Let us make several quick observations regarding $\mathsf{BD}\triangle$. First, the semantical conditions of $\neg$, $\wedge$, and $\vee$ coincide with those from the classical logic. On the other hand, it is more intuitive to interpret $w \vdash^+ \phi$ as ‘w gives evidence for (confirms) ϕ’ and $w \vdash^- \phi$ as ‘w gives evidence against (denies) ϕ’. Thus, w confirms $\phi \wedge \phi'$ when both conjuncts are confirmed by w and w denies $\phi \wedge \phi'$ when at least one conjunct is denied.

The difference is that in $\mathsf{BD}\triangle$ the truth and falsity of a formula *are independent*. Thus, in contrast to the classical logic, neither $p\wedge\neg p\vdash q$ nor $p\vdash q\vee\neg q$ is valid. Second, the addition of $\triangle$ (read 'it is true that') to BD makes it weakly functionally complete (cf. [19] for further details). This allows us to represent every testimony a source can give regarding ϕ (i.e., confirm ϕ, contradict itself regarding ϕ, say nothing about ϕ or deny ϕ) as follows:

$$\mathbf{t}(\phi) := \triangle\phi\wedge\neg\triangle\neg\phi \qquad \mathbf{b}(\phi) := \triangle\phi\wedge\triangle\neg\phi \qquad \mathbf{n}(\phi) := \neg\triangle\phi\wedge\neg\triangle\neg\phi \qquad \mathbf{f}(\phi) := \neg\triangle\phi\wedge\triangle\neg\phi$$

Note that $v_\mathbf{4}(\mathbf{x}(\phi))=\mathbf{T}$ if $v_\mathbf{4}(\phi)=\mathbf{X}$; and $v_\mathbf{4}(\mathbf{x}(\phi))=\mathbf{F}$ otherwise (with $\mathbf{x}\in\{\mathbf{t},\mathbf{b},\mathbf{n},\mathbf{f}\}$ and $\mathbf{X}\in\{\mathbf{T},\mathbf{B},\mathbf{N},\mathbf{F}\}$). Furthermore, it is possible to define $\bot$ and $\top$ s.t. $|\top|^+=W$, $|\top|^-=\varnothing$, $|\bot|^+=\varnothing$, $|\bot|^-=W$ as follows: $\top := \triangle p\vee\neg\triangle p$; $\bot := \neg\top$. $\triangle$ also allows for the internalisation of entailment: for $\mathbf{x},\mathbf{x}'\in\{\mathbf{t},\mathbf{b},\mathbf{n},\mathbf{f}\}$, the formula below is valid iff $\phi\models_{\mathsf{BD}\triangle}\chi$.

$$\phi\Rrightarrow\chi := \bigvee_{\mathbf{x}\leq_\mathbf{4}\mathbf{x}'}(\mathbf{x}(\phi)\wedge\mathbf{x}'(\chi)) \qquad\qquad (\mathbf{f}\leq_\mathbf{4}\mathbf{b},\mathbf{n}\leq_\mathbf{4}\mathbf{t};\ \mathbf{b}\not\leq_\mathbf{4}\mathbf{n};\ \mathbf{n}\not\leq_\mathbf{4}\mathbf{b})$$

The following property will be useful in showing how classical case models can be simulated in $\mathsf{BD}\triangle$.

Proposition 1. *Let $\phi\in\mathscr{L}_{\mathsf{BD}\triangle}$ be s.t. every occurrence of every variable p in ϕ is in the scope of $\triangle$. Then for every $\mathsf{BD}\triangle$ model $\mathfrak{M}$ and $w\in\mathfrak{M}$, exactly one of the following holds: $w\vDash^+\phi$ and $w\nvDash^-\phi$, or $w\nvDash^+\phi$ and $w\vDash^-\phi$.*

Proof. Observe that ϕ is constructed from the formulas of the form $\triangle\chi$ using $\neg$, $\wedge$, and $\vee$. We can now proceed by induction on ϕ. The basis case is simple. From Definition 2, we see that $|\triangle\chi|^+=W\setminus|\triangle\chi|^-$, whence, indeed, either $w\vDash^+\triangle\chi$ and $w\nvDash^-\triangle\chi$ or $w\nvDash^+\triangle\chi$ and $w\vDash^-\triangle\chi$. The cases of $\phi=\psi\vee\psi'$, $\phi=\psi\wedge\psi'$, and $\phi=\neg\psi$ can be shown by straightforward application of the induction hypothesis. □

3 $\mathsf{BD}\triangle$ case models

In this section, we introduce the $\mathsf{BD}\triangle$ case models. To make the presentation clearer, let us first recall the case models from [25, 26] and types of arguments over them that we will henceforth call *classical case models* and *classical arguments* since they use the classical logic as background.

Definition 3 (Classical case models)**.** A *classical case model* is a tuple $\mathfrak{C}_{\mathsf{CPL}}=\langle\mathsf{C},\preceq\rangle$ s.t. C is a finite set of pairwise incompatible classically satisfiable formulas and $\preceq$ is a total preorder on C.

The strict preorder associated with $\preceq$ is interpreted as a preference relation on the set of cases. I.e., $\phi\prec\phi'$ means that the agent prefers ϕ' to ϕ (or trusts in ϕ' more than in ϕ).

Definition 4 (Classical arguments and their types)**.** An *argument* is a tuple $\langle\phi,\phi'\rangle$ of classical propositional formulas. The *case* is the statement $\phi\wedge\phi'$, while a *premise* (conclusion) is any formula χ s.t. $\phi\models_{\mathsf{CPL}}\chi$ ($\phi'\models_{\mathsf{CPL}}\chi$). We say that the argument is *presumptive* iff $\phi\not\models_{\mathsf{CPL}}\phi'$.

An argument $\langle\phi,\chi\rangle$ over a classical case model $\mathfrak{C}_{\mathsf{CPL}}=\langle\mathsf{C},\preceq\rangle$ is

- *classically coherent* iff there is $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{CPL}}\phi\wedge\chi$;
- *classically conclusive* iff it is classically coherent and it holds $\psi\models_{\mathsf{CPL}}\phi\wedge\chi$ for every $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{CPL}}\phi$;
- *classically presumptively valid* iff it is classically coherent and there is $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{CPL}}\phi\wedge\psi$ and $\psi\succeq\psi'$ for every ψ' s.t. $\psi'\models_{\mathsf{CPL}}\phi$.

Let us now present $\mathsf{BD}\triangle$ case models and the counterparts to the coherent, conclusive, and presumptively valid arguments.

Definition 5 ($\mathsf{BD}\triangle$ case models)**.** A $\mathsf{BD}\triangle$ *case model* is a tuple $\mathfrak{C}_{\mathsf{BD}\triangle} = \langle \mathsf{C}, \preceq \rangle$ with C being a finite set of $\mathscr{L}_{\mathsf{BD}\triangle}$ formulas s.t. for any $\phi, \phi' \in \mathsf{C}$, it holds that $\phi \not\models_{\mathsf{BD}\triangle} \bot$ and $\phi \wedge \phi' \models_{\mathsf{BD}\triangle} \bot$, and $\preceq$ a total preorder on C.

Definition 6 (Arguments)**.** An *argument* is a tuple $\langle \phi, \phi' \rangle$ with $\phi, \phi' \in \mathscr{L}_{\mathsf{BD}\triangle}$. The *case* is the statement $\phi \wedge \phi'$, while a *premise* (conclusion) is any formula χ s.t. $\phi \models_{\mathsf{BD}\triangle} \chi$ ($\phi' \models_{\mathsf{BD}\triangle} \chi$). We say that the argument is *presumptive* iff $\phi \not\models_{\mathsf{BD}\triangle} \phi'$.

We can interpret $\psi \in \mathsf{C}$ as witnesses' testimonies. A testimony might be contradictory or omit information relevant to the case. Thus, given an argument $\langle \phi, \chi \rangle$, it makes sense to differentiate between three kinds of conclusions.

1. Given ϕ, χ is claimed to be *true* but nothing is said whether it is also non-false.
2. Given ϕ, χ is claimed to be *non-false* but nothing is said about whether it is true as well.
3. Given ϕ, χ is claimed to be *true and non-false*.

Let us now recall part 1 of the running example and build a case model.

Running example, part 2 (Witnessing a robbery, formalisation). The investigator in part 1 can build the following case model $\mathfrak{C}$ (we omit the ordering for now):

$$\mathsf{C} = \{\underbrace{\mathbf{t}(l) \wedge \mathbf{n}(s) \wedge \mathbf{f}(b)}_{\mathsf{c}_1}, \underbrace{\mathbf{n}(l) \wedge \mathbf{b}(s) \wedge \mathbf{t}(b)}_{\mathsf{c}_2}, \underbrace{\mathbf{t}(l) \wedge \mathbf{t}(s) \wedge \mathbf{b}(b)}_{\mathsf{c}_3}\}$$

Using part 2 of the running example, we define the counterparts to coherent and conclusive arguments from [25, 26].

Definition 7 (Coherent arguments)**.** Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$. $\langle \phi, \chi \rangle$ is

- *negatively coherent* (denoted $\mathfrak{C} \models \phi \mapsto^{-} \chi$) over $\mathfrak{C}$ iff there is $\psi \in \mathsf{C}$ s.t. $\chi \models_{\mathsf{BD}\triangle} \phi \wedge \neg\triangle\neg\chi$;
- *positively coherent* (denoted $\mathfrak{C} \models \phi \mapsto^{+} \chi$) over $\mathfrak{C}$ iff there is $\psi \in \mathsf{C}$ s.t. $\chi \models_{\mathsf{BD}\triangle} \phi \wedge \triangle\chi$;
- *strongly coherent* (denoted $\mathfrak{C} \models \phi \mapsto^{\pm} \chi$) over $\mathfrak{C}$ iff there is $\psi \in \mathsf{C}$ s.t. $\chi \models_{\mathsf{BD}\triangle} \phi \wedge \mathbf{t}(\chi)$.

Definition 8 (Conclusive arguments)**.** Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$. $\langle \phi, \chi \rangle$ is

- *negatively conclusive* over $\mathfrak{C}$ (denoted $\mathfrak{C} \models \phi \Rightarrow^{-} \chi$) iff it is negatively coherent and it holds that if $\chi \models_{\mathsf{BD}\triangle} \phi$, then $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \neg\triangle\neg\chi$ for any $\psi \in \mathsf{C}$;
- *positively conclusive* over $\mathfrak{C}$ (denoted $\mathfrak{C} \models \phi \Rightarrow^{+} \chi$) iff it is positively coherent and it holds that if $\psi \models_{\mathsf{BD}\triangle} \phi$, then $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \triangle\chi$ for any $\psi \in \mathsf{C}$;
- *strongly conclusive* over $\mathfrak{C}$ (denoted $\mathfrak{C} \models \phi \Rightarrow^{\pm} \chi$) iff it is strongly coherent, and it holds that if $\psi \models_{\mathsf{BD}\triangle} \phi$, then $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \mathbf{t}(\chi)$ for any $\psi \in \mathsf{C}$.

Remark 1. Let us provide an intuitive explanation of coherent and conclusive arguments. We begin with coherent arguments:

- for an argument to be *negatively coherent*, there has to be a case that supports the premise and *does not contradict the conclusion*;
- for an argument to be *positively coherent*, there has to be a case that supports the premise and also *supports the conclusion*;

- for an argument to be *strongly coherent*, there has to be a case that supports the premise, *does not contradict the conclusion*, and *supports it*.

Conclusive arguments can be construed as follows:

- for an argument to be *negatively conclusive*, no case satisfying the premise should *contradict* the conclusion of a argument;
- for an argument to be *positively conclusive*, all cases satisfying the premises of an argument should support its conclusion.

Observe that the arguments that are both positively and negatively conclusive are strongly conclusive as well. On the other hand, $\langle\phi,\chi\rangle$ can be both positively and negatively coherent but not strongly coherent if there is no case c s.t. $\mathsf{c}\models_{\mathsf{BD}\triangle}\mathbf{t}(\chi)$.

Remark 2 ($\mathsf{BD}\triangle$ and classical arguments). Note that while there is no classical case model over which both $\mathscr{A}=\langle\phi,\chi\rangle$ and $\mathscr{B}=\langle\phi,\neg\chi\rangle$ are classically conclusive (albeit, they can be presumptively valid), it is possible that they are both *positively conclusive* (*negatively conclusive*) if $\mathsf{c}\models_{\mathsf{BD}\triangle}\mathbf{b}(\chi)$ (resp., $\mathsf{c}\models_{\mathsf{BD}\triangle}\mathbf{n}(\chi)$) for every $\mathsf{c}\in\mathsf{C}$. Still, there is no $\mathsf{BD}\triangle$ case model over which $\mathscr{A}$ and $\mathscr{B}$ are *strongly conclusive*.

In addition, it is clear that no argument of the form $\langle\phi,\neg\phi\rangle$ is *classically coherent* since $\phi\wedge\neg\phi$ is classically unsatisfiable. On the other hand, $\langle s,\neg s\rangle$ is *positively coherent* (by c_2) in the model from the part 2 of the running example.

Finally, it is easy to see that every coherent deductive classical argument $\langle\phi,\chi\rangle$ (i.e., the one where $\phi\models_{\mathsf{CPL}}\chi$) is also classically conclusive. In the case of $\mathsf{BD}\triangle$ arguments, only the weaker statement holds: *if* $\phi\models_{\mathsf{BD}\triangle}\chi$ *and* $\langle\phi,\chi\rangle$ *is positively coherent, then it is positively conclusive as well.* E.g., $p\wedge\neg p\wedge q\models_{\mathsf{BD}\triangle}p\wedge\neg p$ but $\mathbf{t}(p\wedge\neg p)$ always has value $\mathbf{F}$, whence $\langle p\wedge\neg p\wedge q,p\wedge\neg p\rangle$ can never be negatively or strongly coherent (and thus, negatively or strongly conclusive).

Let us now define the $\mathsf{BD}\triangle$ counterparts of presumptively valid arguments.

Definition 9 (Presumptively valid arguments)**.** An argument $\mathscr{A}=\langle\phi,\chi\rangle$ is:

- *positively presumptively valid* (denoted $\mathfrak{C}\models\phi\rightsquigarrow^{+}\chi$) iff there is $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{BD}\triangle}\phi\wedge\triangle\chi$ and $\psi\succeq\psi'$ for any ψ' s.t. $\psi'\models\phi$;
- *negatively presumptively valid* (denoted $\mathfrak{C}\models\phi\rightsquigarrow^{-}\chi$) iff there is $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{BD}\triangle}\phi\wedge\neg\triangle\neg\chi$ and $\psi\succeq\psi'$ for any ψ' s.t. $\psi'\models\phi$;
- *strongly presumptively valid* (denoted $\mathfrak{C}\models\phi\rightsquigarrow^{\pm}\chi$) iff there is $\psi\in\mathsf{C}$ s.t. $\psi\models_{\mathsf{BD}\triangle}\phi\wedge\mathbf{t}(\chi)$ and $\psi\succeq\psi'$ for any ψ' s.t. $\psi'\models\phi$.

Convention 1. We will further call ψ the *witnessing case for* $\mathscr{A}$.

Remark 3. We can now explain presumptively valid arguments similarly to how we interpreted coherent and conclusive ones.

- An argument is *negatively coherent* when there is the most preferred case that supports its premise and *does not contradict the conclusion*.
- An argument is *positively coherent* when there is the most preferred case that supports *both its premise and conclusion*.

Running example, part 3 (Witnessing a robbery, preferences). We return to the model in part 2. The investigator now wants to find out how the robber escaped from the office. It is clear that neither $\langle\top,s\rangle$ nor $\langle\top,\neg s\rangle$ is strongly conclusive. On the other hand, nobody *explicitly denied* that the robber was

$$\begin{tikzcd}
\phi \mapsto^{+} \chi \arrow[rr, phantom, "\supseteq"] \arrow[dr, phantom, sloped, "\supseteq"] & & \phi \rightsquigarrow^{+} \chi \arrow[rr, phantom, "\supseteq"] \arrow[dr, phantom, sloped, "\supseteq"] & & \phi \Rightarrow^{+} \chi \arrow[dr, phantom, sloped, "\supseteq"] & \\
& \phi \mapsto^{\pm} \chi \arrow[rr, phantom, "\supseteq"] & & \phi \rightsquigarrow^{\pm} \chi \arrow[rr, phantom, "\supseteq"] & & \phi \Rightarrow^{\pm} \chi \\
\phi \mapsto^{-} \chi \arrow[rr, phantom, "\supseteq"] \arrow[ur, phantom, sloped, "\supseteq"] & & \phi \rightsquigarrow^{-} \chi \arrow[rr, phantom, "\supseteq"] \arrow[ur, phantom, sloped, "\supseteq"] & & \phi \Rightarrow^{-} \chi \arrow[ur, phantom, sloped, "\supseteq"] &
\end{tikzcd}$$

Figure 1: Conclusive ($\Rightarrow$), presumptively valid ($\rightsquigarrow$), and coherent ($\mapsto$) arguments with same statements.

limping, whence $\langle\top, l\rangle$ is *negatively conclusive*. The investigator thinks that it is reasonable to take l to be true. Unfortunately, even assuming l, neither $\langle l, s\rangle$ nor $\langle l, \neg s\rangle$ is conclusive.

The investigator rereads the accounts of c_1, c_2, and c_3 and notices that c_3 was the only one to follow the robber out of the office. On the other hand, c_1 hid under the table and was sitting there during the robbery. Thus, the preference is given as $\mathsf{c}_1 \prec \mathsf{c}_2 \prec \mathsf{c}_3$. Hence, $\langle l, s\rangle$ is *strongly presumptively valid.*

Remark 4. It is important to note that *both following statements are false*:

- $\langle\phi, \chi\rangle$ is negatively coherent (resp., presumptively valid, conclusive) iff $\langle\phi, \neg\chi\rangle$ is positively coherent (resp., presumptively valid, conclusive);
- $\langle\phi, \chi\rangle$ is positively coherent (resp., presumptively valid, conclusive) iff $\langle\phi, \neg\chi\rangle$ is negatively coherent (resp., presumptively valid, conclusive).

Indeed, recall part 2 of the running example. $\langle\top, l\rangle$ is negatively coherent while $\langle\top, \neg l\rangle$ is not positively coherent. $\langle b, \neg s\rangle$ is negatively presumptively valid but $\langle s, \neg b\rangle$ is positively presumptively valid but $\langle s, b\rangle$ is not negatively presumptively valid.

The following statement establishes the expected relations between coherent, presumptively valid, and conclusive arguments and follows immediately from Definitions 7–9.

Proposition 2. *The diagram in Fig. 1 depicts the inclusions between different types of arguments.*

It is instructive to see that $\mathsf{BD}\triangle$ models allow classical presumptive reasoning if the values of formulas in the cases are classical. We define a class of $\mathsf{BD}\triangle$ case models 'simulating' the classical ones.

Definition 10 (Quasi-classical case models)**.** A $\mathsf{BD}\triangle$ case model $\mathfrak{C} = \langle\mathsf{C}, \preceq\rangle$ is called *quasi-classical* iff every $\chi \in \mathsf{C}$ is constructed from $\mathbf{t}(p)$'s via applications of $\neg$, $\wedge$, and $\vee$.

Proposition 3. *Let* $\mathfrak{C}$ *be a quasi-classical* $\mathsf{BD}\triangle$ *case model and* $\rhd \in \{\mapsto, \rightsquigarrow, \Rightarrow\}$. *Then* $\mathfrak{C} \models \phi \rhd^{+} \chi$ *iff* $\mathfrak{C} \models \phi \rhd^{-} \chi$ *iff* $\mathfrak{C} \models \phi \rhd^{\pm} \chi$.

Proof. We only consider the case of coherent arguments as conclusive and presumptively valid ones can be tackled similarly. It suffices to prove that positively coherent arguments and negatively coherent arguments are strongly coherent. Let $\mathfrak{C}$ be quasi-classical and $\mathfrak{C} \models \phi \mapsto^{+} \chi$. Then, there is $\psi \in \mathfrak{C}$ s.t. $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \triangle\chi$. But then, from Definition 2 and Proposition 1, it is clear that $\psi \models_{\mathsf{BD}\triangle} \mathbf{t}(\chi)$. Likewise, let $\mathfrak{C} \models \phi \mapsto^{-} \chi$, and, accordingly, $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \neg\triangle\neg\chi$. Again, using Definition 2 and Proposition 1, we have $\psi \models_{\mathsf{BD}\triangle} \mathbf{t}(\chi)$. The result now follows. □

We finish the section by showing how to build a $\mathsf{BD}\triangle$ counterpart of a classical case model that preserves all arguments.

Definition 11. Let $\mathfrak{C} = \langle\mathsf{C}, \preceq\rangle$ be a classical case model s.t. all formulas in C are over $\{\neg, \wedge, \vee\}$. In addition, for $\phi \in \mathscr{L}_{\mathsf{BD}}$, denote $\phi^{\mathbf{t}}$ the result of substitution of every variable p occurring in ϕ with $\mathbf{t}(p)$.

The $\mathsf{BD}\triangle$ *counterpart* of $\mathfrak{C}$ is $\mathfrak{C}_{\mathsf{BD}\triangle} = \langle\mathsf{C}_{\mathsf{BD}\triangle}, \preceq_{\mathsf{BD}\triangle}\rangle$ with $\mathsf{C}_{\mathsf{BD}\triangle} = \{\chi^{\mathbf{t}} : \chi \in \mathsf{C}\}$ and $\chi^{\mathbf{t}} \preceq_{\mathsf{BD}\triangle} \chi'^{\mathbf{t}}$ iff $\chi \preceq_{\mathsf{BD}} \chi'$.

Theorem 1. *Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a classical case model s.t. all formulas in C are over $\{\neg, \wedge, \vee\}$.*

1. *The $\mathsf{BD}\triangle$ counterpart $\mathfrak{C}_{\mathsf{BD}\triangle}$ of $\mathfrak{C}$ is quasi-classical.*
2. *$\langle \phi, \chi \rangle$ is coherent (resp., presumptively valid, conclusive) in $\mathfrak{C}$ iff $\langle \phi^{\mathbf{t}}, \chi^{\mathbf{t}} \rangle$ is strongly coherent (resp., strongly presumptively valid, strongly conclusive) in $\mathfrak{C}_{\mathsf{BD}\triangle}$.*

Proof. 1. holds by Definitions 3 and 11. Let us now consider 2. We will only tackle the case of presumptively valid arguments since coherent and conclusive ones can be dealt with similarly.

Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a classical case model and $\langle \phi, \chi \rangle$ presumptively valid on $\mathfrak{C}$. Then, there is $\psi \in \mathsf{C}$ s.t. $\psi \models_{\mathsf{CPL}} \phi \wedge \chi$ and $\psi \succeq \psi'$ for every $\psi' \in \mathsf{C}$ s.t. $\psi' \models_{\mathsf{CPL}} \phi$. Now observe from Definition 1 that $\neg$, $\wedge$, and $\vee$ behave classically on $\mathbf{T}$ and $\mathbf{F}$. Thus, it is clear that $\tau \models_{\mathsf{CPL}} \tau'$ iff $\tau^{\mathbf{t}} \models_{\mathsf{BD}\triangle} \tau'^{\mathbf{t}}$ for every $\tau, \tau' \in \mathscr{L}_{\mathsf{BD}}$. It now follows that $\psi^{\mathbf{t}} \models_{\mathsf{BD}\triangle} \phi^{\mathbf{t}} \wedge \mathbf{t}(\chi^{\mathbf{t}})$ and $\psi^{\mathbf{t}} \succeq_{\mathsf{BD}\triangle} \psi'^{\mathbf{t}}$ for every $\psi'^{\mathbf{t}} \models_{\mathsf{BD}\triangle} \phi^{\mathbf{t}}$, as required. Conversely, let $\langle \phi, \chi \rangle$ be not presumptively valid on $\mathfrak{C}$. Then, there is no $\psi \in \mathsf{C}$ s.t. $\psi \models_{\mathsf{CPL}} \phi \wedge \chi$ and $\psi \succeq \psi'$ for every $\psi' \in \mathsf{C}$ s.t. $\psi' \models_{\mathsf{CPL}} \phi$. Again, from Definition 1, it follows that there is no $\psi^{\mathbf{t}}$ s.t. $\psi^{\mathbf{t}} \models_{\mathsf{BD}\triangle} \phi^{\mathbf{t}} \wedge \mathbf{t}(\chi^{\mathbf{t}})$ and $\psi^{\mathbf{t}} \succeq_{\mathsf{BD}\triangle} \psi'^{\mathbf{t}}$ for every $\psi'^{\mathbf{t}} \models_{\mathsf{BD}\triangle} \phi^{\mathbf{t}}$. □

4 A two-layered logic for case models

Conclusive and presumptively valid arguments on *classical* case models can be represented in terms of conditional probabilities [25, 26]. In this section, we provide a representation of $\mathsf{BD}\triangle$ models and arguments on them in terms of a paraconsistent two-layered logic.

Two-layered logics form a class of formalisms designed to reason about uncertainty: their languages consist of *inner-layer* formulas that describe events and *outer-layer* formulas composed of *modal atoms* of the form $\mathsf{M}\phi$ (ϕ being an inner-layer formula and M the modality interpreted as a measure on the set of events). There are two-layered logics formalising reasoning with classical probabilities [3] and their paraconsistent counterparts [7] presented in [16].[1] Furthermore, there are two-layered logics formalising paraconsistent reasoning with belief and plausibility functions [7].

These papers usually study *quantitative* representations of uncertainty. On the other hand, case models provide a *qualitative one* via their preference relations. This shows a degree of affinity between case models and representations of different uncertainty measures by means of total preorders as studied in [17] (for the case of probabilities) and [28, 27] (belief functions). In [6], two-layered logics formalising reasoning with the qualitative counterparts of belief functions and probabilities were presented.

In this section, we present a two-layered logic $\mathsf{QG}_{\mathsf{BD}\triangle}$ which is a modification of QG from [6]. The inner layer of $\mathsf{MCB}_{\triangle}$ is $\mathsf{BD}\triangle$, the outer one is biG — an expansion of Gödel logic (cf., e.g., [15]) with a coimplication $\prec$ or the Baaz Delta operator $\triangle$. To connect the layers, we use $\mathtt{B}$ (with $\mathtt{B}\phi$ read as 'the agent believes in ϕ'). Since we do not impose any restrictions on $\preceq$ in case models, we are interpreting $\mathtt{B}$ as a *capacity* on the set of events W, i.e., via a map $\mu : 2^W \to [0,1]$ which is monotone w.r.t. $\subseteq$ with $\mu(W) = 1$ and $\mu(\varnothing) = 0$. The main goal of the paper is to establish a correspondence between case models and $\mathsf{QG}_{\mathsf{BD}\triangle}$ models as well as to show how given an argument $\langle \phi, \phi' \rangle$ to construct a $\mathsf{QG}_{\mathsf{BD}\triangle}$ formula that is true in the corresponding model iff the argument is (positively, negatively, or strongly) coherent, conclusive, or presumptively valid.

Let us now recall biG.

[1] Note that [16] is not the only paraconsistent interpretation of probabilities: alternative approaches can be found. e.g., in [18, 13, 10, 23].

Definition 12. The bi-Gödel algebra $[0,1]_{\mathsf{G}} = \langle [0,1], 0, 1, \wedge_{\mathsf{G}}, \vee_{\mathsf{G}}, \rightarrow_{\mathsf{G}}, \prec, \sim_{\mathsf{G}}, \triangle_{\mathsf{G}} \rangle$ is defined as follows: for all $a, b \in [0,1]$, $\wedge_{\mathsf{G}}$ and $\vee_{\mathsf{G}}$ are given by $a \wedge_{\mathsf{G}} b := \min(a,b)$, $a \vee_{\mathsf{G}} b := \max(a,b)$. The remaining operations are defined below:

$$a \rightarrow_{\mathsf{G}} b = \begin{cases} 1, \text{ if } a \leq b \\ b \text{ else} \end{cases} \qquad a \prec_{\mathsf{G}} b = \begin{cases} 0, \text{ if } a \leq b \\ a \text{ else} \end{cases} \qquad \sim_{\mathsf{G}} a = \begin{cases} 0, \text{ if } a > 0 \\ 1 \text{ else} \end{cases} \qquad \triangle_{\mathsf{G}} a = \begin{cases} 0, \text{ if } a < 1 \\ 1 \text{ else} \end{cases}$$

Remark 5. Note that constants $\top$ and $\bot$ are definable as, respectively, $p \rightarrow p$ and $p \prec p$, and that $\triangle$ and $\prec$ are interdefinable as follows: $\triangle\phi := \top \prec (\top \prec \phi)$, $\phi \prec \phi' := \phi \wedge {\sim}\triangle(\phi \rightarrow \phi')$.

Definition 13 (Language and semantics of biG)**.** We fix a countable set $\texttt{Prop}$ of propositional variables and consider the following language.

$$\mathscr{L}_{\mathsf{biG}} \ni \phi := p \in \texttt{Prop} \mid {\sim}\phi \mid (\phi \wedge \phi) \mid (\phi \vee \phi) \mid (\phi \rightarrow \phi) \mid (\phi \prec \phi) \mid \triangle\phi$$

Let $e : \mathsf{Prop} \rightarrow [0,1]$. For the complex formulas, we define $e(\phi \circ \phi') = e(\phi) \circ_{\mathsf{G}} e(\phi')$.
Finally, let $\Gamma \cup \{\phi\} \subseteq \mathscr{L}_{\mathsf{biG}}$. We define: $\Gamma \models_{\mathsf{biG}} \phi$ iff $\forall e : \inf\{e(\psi) : \psi \in \Gamma\} \leq e(\phi)$.

Using biG, we can define $\mathsf{QG}_{\mathsf{BD}\triangle}$ as follows.

Definition 14. The language of $\mathsf{QG}_{\mathsf{BD}\triangle}$ is defined via the following grammar: $\mathscr{L}_{\mathsf{QG}_{\mathsf{BD}\triangle}} \ni \alpha := \mathtt{B}\phi \mid \alpha \circ \alpha$ $(\circ \in \{\sim, \wedge, \vee, \rightarrow, \prec, \triangle\}, \phi \in \mathscr{L}_{\mathsf{BD}\triangle})$. A $\mathsf{QG}_{\mathsf{BD}\triangle}$ model is a tuple $\mathscr{M} = \langle W, v^+, v^-, \mu, e \rangle$ with $\langle W, v^+, v^- \rangle$ being a $\mathsf{BD}\triangle$ model (cf. Definition 2), $\mu : 2^W \rightarrow [0,1]$ being a capacity. Semantic conditions of $\mathscr{L}_{\mathsf{QG}_{\mathsf{BD}\triangle}}$ formulas are as follows: $e(\mathtt{B}\phi) = \mu(|\phi|^+)$ for modal atoms; the values of complex formulas are computed according to Definition 13.

For a given model $\mathscr{M}$, we write $\mathscr{M} \models \alpha$ to designate $e(\alpha) = 1$. For a frame $\mathbb{F} = \langle W, \pi \rangle$ on a $\mathsf{QG}_{\mathsf{BD}\triangle}$ model $\mathscr{M}$, we say that $\alpha \in \mathscr{L}_{\mathsf{MCB}_\triangle}$ is valid on $\mathbb{F}$ ($\mathbb{F} \models \alpha$) iff $e(\alpha) = 1$ for every e on $\mathbb{F}$. Finally, for $\Psi \cup \{\alpha\} \subseteq \mathscr{L}_{\mathsf{QG}_{\mathsf{BD}\triangle}}$, we define the same entailment relation as in Definition 13.

Let us now establish the correspondence results for coherent, conclusive, and presumptively valid arguments. To do this, we define a class of *μ-counterparts* for every $\mathsf{BD}\triangle$ model.

Definition 15 (μ-counterparts)**.** Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a $\mathsf{BD}\triangle$ case model and $\mathsf{C} = \{\mathsf{c}_1, \ldots, \mathsf{c}_n\}$. Its μ-counterpart is a $\mathsf{QG}_{\mathsf{BD}\triangle}$-model $\mathscr{M}_{\mathfrak{C}} = \langle \{w_1, \ldots, w_n\}, v^+, v^-, \mu_{\preceq}, e \rangle$ for which the following holds.

1. For every $\mathsf{c}_i \in \mathsf{C}$ and every ϕ, if $\mathsf{c}_i \models_{\mathsf{BD}\triangle} \phi$ ($\mathsf{c}_i \models_{\mathsf{BD}\triangle} \neg\phi$), then $w_i \vDash^+ \phi$ ($w_i \vDash^- \phi$).
2. For every $\mathsf{c}_i, \mathsf{c}_j \in \mathsf{C}$, $\mathsf{c}_i \preceq \mathsf{c}_j$ iff $\mu_{\preceq}(\{w_i\}) \leq \mu_{\preceq}(\{w_j\})$.
3. For every $\mathsf{c}_i \in \mathsf{C}$, $\mu_{\preceq}(\{\mathsf{c}_i\}) > 0$.

Theorem 2. *Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a $\mathsf{BD}\triangle$ case model and $\mathscr{M}_{\mathfrak{C}}$ its μ-counterpart. Then the following holds.*

1. $\mathfrak{C} \models \phi \mapsto^+ \phi'$ *iff* $\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathtt{B}(\phi \wedge \triangle\phi')$.
2. $\mathfrak{C} \models \phi \mapsto^- \phi'$ *iff* $\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathtt{B}(\phi \wedge \neg\triangle\neg\phi')$.
3. $\mathfrak{C} \models \phi \mapsto^\pm \phi'$ *iff* $\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathtt{B}(\phi \wedge \mathbf{t}(\phi'))$.

Proof. We consider 2. Other cases can be proved in the same way. Let $\mathfrak{C} \models \phi \mapsto^- \phi'$. Then, there is $\mathsf{c}_i \in \mathfrak{C}$ s.t. $\mathsf{c}_i \models_{\mathsf{BD}\triangle} \phi \wedge \neg\triangle\neg\phi'$, whence $w_i \vDash^+ \phi \wedge \neg\triangle\neg\phi'$ and $\mu(|\phi \wedge \neg\triangle\neg\phi'|^+) > 0$. Thus, $e(\mathtt{B}(\phi \wedge \neg\triangle\neg\phi')) > 0$ and $\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathtt{B}(\phi \wedge \neg\triangle\neg\phi')$, as required. Conversely, let $\mathfrak{C} \not\models \phi \mapsto^- \phi'$. Then, for every $\mathsf{c}_i \in \mathfrak{C}$, $\mathsf{c}_i \not\models_{\mathsf{BD}\triangle} \phi \wedge \neg\triangle\neg\phi'$, whence there is no w_i s.t. $w_i \vDash^+ \phi \wedge \neg\triangle\neg\phi'$. Hence, $|\phi \wedge \neg\triangle\neg\phi'|^+ = \varnothing$ and $\mu(|\phi \wedge \neg\triangle\neg\phi'|^+) = 0$. Thus, ${\sim\sim}e(\mathtt{B}(|\phi \wedge \neg\triangle\neg\phi'|^+)) = 0$, as required. □

Observe from Definitions 8 and 9 that the classes of strongly conclusive (presumptively valid) arguments on the one hand and both positively and negatively conclusive (presumptively valid) arguments on the other hand coincide. Thus, it suffices to provide representation for positively and negatively conclusive (presumptively valid) arguments only.

Theorem 3. *Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a $\mathsf{BD}\triangle$ case model and $\mathscr{M}_{\mathfrak{C}}$ its μ-counterpart. Then the following holds.*

1. $\mathfrak{C} \models \phi \Rightarrow^{+} \phi'$ *iff* $\mathscr{M}_{\mathfrak{C}} \models {\sim}\mathsf{B}(\phi \wedge \neg\triangle\phi') \wedge {\sim\sim}\mathsf{B}(\phi \wedge \triangle\phi')$.
2. $\mathfrak{C} \models \phi \Rightarrow^{-} \phi'$ *iff* $\mathscr{M}_{\mathfrak{C}} \models {\sim}\mathsf{B}(\phi \wedge \triangle\neg\phi') \wedge {\sim\sim}\mathsf{B}(\phi \wedge \neg\triangle\neg\phi')$.

Proof. Again, for the sake of brevity, we consider only 1. We let $\mathfrak{C} \models \phi \Rightarrow^{+} \phi'$. Then, $\langle \phi, \phi' \rangle$ is positively coherent on $\mathfrak{C}$ and thus (by Theorem 2), $\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathsf{B}(\phi \wedge \triangle\phi')$. Furthermore, since $\psi \models_{\mathsf{BD}\triangle} \phi \wedge \triangle\phi'$ for every $\psi \in \mathfrak{C}$ s.t. $\psi \models_{\mathsf{BD}\triangle} \phi$, we have that $|\phi|^{+} \cap (\mathsf{C} \setminus |\triangle\phi'|^{+}) = \varnothing$, whence $\mathscr{M}_{\mathfrak{C}} \models {\sim}\mathsf{B}(\phi \wedge \neg\triangle\phi')$, as required. As the converse direction can be proved in the same manner, the result follows. □

Theorem 4. *Let $\mathfrak{C} = \langle \mathsf{C}, \preceq \rangle$ be a $\mathsf{BD}\triangle$ case model and $\mathscr{M}_{\mathfrak{C}}$ its μ-counterpart. Then the following holds.*

1. $\mathfrak{C} \models \phi \rightsquigarrow^{+} \phi'$ *and χ is $\langle \phi, \phi' \rangle$'s witnessing case iff*
$$\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathsf{B}(\phi \wedge \triangle\phi') \wedge \triangle\mathsf{B}(\chi \Rrightarrow (\phi \wedge \triangle\phi')) \wedge \textstyle\bigwedge_{\chi' \in \mathsf{C}} (\triangle\mathsf{B}(\chi \Rrightarrow \phi) \rightarrow \triangle(\mathsf{B}\chi' \rightarrow \mathsf{B}\chi))$$
2. $\mathfrak{C} \models \phi \rightsquigarrow^{-} \phi'$ *and χ is $\langle \phi, \phi' \rangle$'s witnessing case iff*
$$\mathscr{M}_{\mathfrak{C}} \models {\sim\sim}\mathsf{B}(\phi \wedge \neg\triangle\neg\phi') \wedge \triangle\mathsf{B}(\chi' \Rrightarrow (\phi \wedge \neg\triangle\neg\phi')) \wedge \textstyle\bigwedge_{\chi' \in \mathsf{C}} (\triangle\mathsf{B}(\chi' \Rrightarrow \phi) \rightarrow \triangle(\mathsf{B}\chi' \rightarrow \mathsf{B}\chi))$$

Proof. We prove 1. as 2. can be proven in the same manner. Assume that $\langle \phi, \phi' \rangle$ is positively presumptively valid over $\mathfrak{C}$ and that χ is its witnessing case. Then, $\langle \phi, \phi' \rangle$ is positively coherent (whence, $\mathscr{M} \models {\sim\sim}\mathsf{B}(\phi \wedge \neg\triangle\neg\phi')$) and $\chi \models_{\mathsf{BD}\triangle} \phi \wedge \triangle\phi'$. Thus, $|\chi|^{+} \subseteq |\phi \wedge \triangle\phi'|^{+}$ and $|\chi|^{-} \supseteq |\phi \wedge \triangle\phi'|^{-}$ *for every model $\mathscr{M}$*, whence $\mathscr{M} \models \triangle\mathsf{B}(\chi \Rrightarrow (\phi \wedge \neg\triangle\neg\phi'))$. Finally, we also have that $\chi' \preceq \chi$ for every $\chi' \in \mathsf{C}$ s.t. $\chi \models_{\mathsf{BD}\triangle} \phi$. But this means that for every such χ', $\mu(|\chi'|^{+}) \leq \mu(|\chi|^{+})$[2] and thus, $\mathscr{M} \models \triangle(\mathsf{B}\chi' \rightarrow \mathsf{B}\chi)$. Hence, $\mathscr{M} \models \bigwedge_{\chi' \in \mathsf{C}} (\triangle\mathsf{B}(\chi' \Rrightarrow \phi) \rightarrow \triangle(\mathsf{B}\chi' \rightarrow \mathsf{B}\chi))$, as required.

For the converse, let $\langle \phi, \phi' \rangle$ be *not* positively presumptively valid argument with χ as the witnessing case. Then at least one of the following holds: (1) $\langle \phi, \phi' \rangle$ is not positively coherent; (2) $\chi \not\models_{\mathsf{BD}\triangle} \phi \wedge \triangle\phi'$; (3) there is some $\chi' \in \mathsf{C}$ s.t. $\chi' \models_{\mathsf{BD}\triangle} \phi$ but $\chi' \succ \chi$. Now, for (1), $\mathscr{M} \not\models {\sim\sim}\mathsf{B}(\phi \wedge \neg\triangle\neg\phi')$; for (2), $\mathscr{M} \not\models \triangle\mathsf{B}(\chi \Rrightarrow (\phi \wedge \triangle\phi'))$; and finally, for (3), $\mathscr{M} \not\models \bigwedge_{\chi' \in \mathsf{C}} (\triangle\mathsf{B}(\chi' \Rrightarrow \phi) \rightarrow \triangle(\mathsf{B}\chi' \rightarrow \mathsf{B}\chi))$. □

Recall that conclusive and coherent arguments over *classical* case models were represented by means of *conditional probabilities* in [25, 26]. Here, we did not need conditionalisations on capacities as we used a purely logical representation and could employ $\rightarrow$ in order to 'simulate' conditionalised measures.

5 Conclusion

In this paper, we provided paraconsistent counterparts to the case models discussed in [25, 26] that use $\mathsf{BD}\triangle$ as their underlying logic. We showed how to recover classical presumptive reasoning from $\mathsf{BD}\triangle$ case models (Theorem 1). Moreover, we constructed a two-layered logic $\mathsf{QG}_{\mathsf{BD}\triangle}$ over $\mathsf{BD}\triangle$ and biG and used it to establish a representation of arguments with $\mathsf{QG}_{\mathsf{BD}\triangle}$ formulas (Theorems 2–4).

The natural next steps would be as follows. First, it is instructive to provide a complete axiomatisation of $\mathsf{QG}_{\mathsf{BD}\triangle}$. Second, while in this paper we used a linear preference relation (as it is traditionally done,

[2] Recall that since $\mathsf{c}_i \wedge \mathsf{c}_j \models_{\mathsf{BD}\triangle} \bot$ for every $\mathsf{c}_i, \mathsf{c}_j \in \mathsf{C}$, we have that $\mu(|\mathsf{c}_k|^{+}) = \mu(\{w_k\})$ for all $\mathsf{c}_k \in \mathsf{C}$.

cf., e.g., [22]), one could argue that if an agent cannot choose between two cases c and c', it does not mean that they prefer them to the same degree. It is, hence, reasonable to explore case models whose preference relation is a *partial* preorder. Finally, we managed to represent preference relations on case models as capacities on their $\mathsf{BD}\triangle$ counterparts. An expected question to ask is which properties we have to require from the case model so that its preference relation be represented as a stronger measure: e.g., a belief function, a plausibility function, or a probability measure.

References

[1] A.R. Anderson & N.D. Belnap (1962): *Tautological entailments*. Philosophical Studies 13(1), pp. 9–24, doi:10.1007/BF00818100.

[2] M. Baaz (1996): *Infinite-valued Gödel logics with* 0-1*-projections and relativizations*. In: Gödel'96: Logical foundations of mathematics, computer science and physics—Kurt Gödel's legacy, Brno, Czech Republic, August 1996, proceedings, Association for Symbolic Logic, pp. 23–33.

[3] P. Baldi, P. Cintula & C. Noguera (2020): *Classical and Fuzzy Two-Layered Modal Logics for Uncertainty: Translations and Proof-Theory*. International Journal of Computational Intelligence Systems 13, pp. 988–1001, doi:10.2991/ijcis.d.200703.001.

[4] N.D. Belnap (2019): *How a Computer Should Think*. In H. Omori & H. Wansing, editors: New Essays on Belnap-Dunn Logic, Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science) 418, Springer, Cham, pp. 35–53, doi:10.1007/978-3-030-31136-0_4.

[5] P. Besnard & A. Hunter (2001): *A logic-based theory of deductive arguments*. Artificial Intelligence 128(1–2), pp. 203–235, doi:10.1016/S0004-3702(01)00071-6.

[6] M. Bílková, S. Frittella, D. Kozhemiachenko & O. Majer (2023): *Qualitative reasoning in a two-layered framework*. International Journal Approximate Reasoning 154, pp. 84–108, doi:10.1016/j.ijar.2022.12.011.

[7] M. Bílková, S. Frittella, D. Kozhemiachenko, O. Majer & S. Nazari (2022): *Reasoning with belief functions over Belnap–Dunn logic*. `https://arxiv.org/abs/2203.01060`.

[8] M. Bílková, S. Frittella, O. Majer & S. Nazari (2020): *Belief Based on Inconsistent Information*. In M.A. Martins & I. Sedlár, editors: Dynamic Logic. New Trends and Applications, Springer International Publishing, Cham, pp. 68–86, doi:10.1007/978-3-030-65840-3_5.

[9] A. Bondarenko, P.M. Dung, R.A. Kowalski & F. Toni (1997): *An abstract, argumentation-theoretic approach to default reasoning*. Artificial Intelligence 93(1-2), pp. 63–101, doi:10.1016/S0004-3702(97)00015-5.

[10] J. Bueno-Soler & W. Carnielli (2016): *Paraconsistent probabilities: Consistency, contradictions and Bayes' theorem*. Entropy (Basel) 18(9), p. 325, doi:10.3390/e18090325.

[11] P.M. Dung (1995): *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. Artificial Intelligence 77(2), pp. 321–357, doi:10.1016/0004-3702(94)00041-X.

[12] J.M. Dunn (1976): *Intuitive semantics for first-degree entailments and 'coupled trees'*. Philosophical Studies 29(3), pp. 149–168, doi:10.1007/BF00373152.

[13] J.M. Dunn (2010): *Contradictory information: Too much of a good thing*. Journal of Philosophical Logic 39, pp. 425–452, doi:10.1007/s10992-010-9134-6.

[14] A.J. García & G.R. Simari (2004): *Defeasible logic programming: an argumentative approach*. Theory and Practice of Logic Programming 4(1+2), pp. 95–138, doi:10.1017/S1471068403001674.

[15] P. Hájek (1998): *Metamathematics of Fuzzy Logic*. Trends in Logic 4, Springer, Dordrecht, doi:10.1007/978-94-011-5300-3_5.

[16] D. Klein, O. Majer & S. Rafiee Rad (2021): *Probabilities with gaps and gluts*. Journal of Philosophical Logic 50(5), pp. 1107–1141, doi:10.1007/s10992-021-09592-x.

[17] C.H. Kraft, J.W. Pratt & A. Seidenberg (1959): *Intuitive probability on finite sets*. The Annals of Mathematical Statistics 30(2), pp. 408–419, doi:10.1214/aoms/1177706260.

[18] E.D. Mares (1997): *Paraconsistent Probability Theory and Paraconsistent Bayesianism*. Logique et Analyse 40(160), pp. 375–384.

[19] H. Omori & K. Sano (2015): *Generalizing Functional Completeness in Belnap-Dunn Logic*. Studia Logica 103(5), pp. 883–917, doi:10.1007/s11225-014-9597-5.

[20] H. Omori & H. Wansing (2017): *40 years of FDE: An Introductory Overview*. Studia Logica 105(6), pp. 1021–1049, doi:10.1007/s11225-017-9748-6.

[21] H. Prakken (2010): *An abstract framework for argumentation with structured arguments*. Argument & Computation 1(2), pp. 93–124, doi:10.1080/19462160903564592.

[22] F.S. Roberts (1985): *Measurement Theory*. Cambridge University Press.

[23] A. Rodrigues, J. Bueno-Soler & W. Carnielli (2021): *Measuring evidence: a probabilistic approach to an extension of Belnap–Dunn logic*. Synthese 198(S22), pp. 5451–5480, doi:10.1007/s11229-020-02571-w.

[24] K. Sano & H. Omori (2014): *An expansion of first-order Belnap–Dunn logic*. Logic Journal of IGPL 22(3), pp. 458–481, doi:10.1093/jigpal/jzt044.

[25] B. Verheij (2016): *Correct grounded reasoning with presumptive arguments*. In: Logics in Artificial Intelligence, Lecture notes in computer science, Springer International Publishing, Cham, pp. 481–496, doi:10.1007/978-3-031-10769-6_26.

[26] B. Verheij (2017): *Proof with and without probabilities*. Artificial Intelligence and Law 25(1), pp. 127–154, doi:10.1007/s10506-017-9199-4.

[27] S.K.M. Wong, Y. Yao & P. Bollmann (1992): *Characterization of comparative belief structures*. International journal of man-machine studies 37(1), pp. 123–133, doi:10.1016/0020-7373(92)90094-2.

[28] S.K.M. Wong, Y.Y. Yao, P. Bollmann & H.C. Bürger (1991): *Axiomatization of qualitative belief structure*. IEEE Transactions on Systems, Man, and Cybernetics 21(4), pp. 726–734, doi:10.1109/21.108290.

Optimal Mechanism Design for Agents with DSL Strategies: The Case of Sybil Attacks in Combinatorial Auctions

Yotam Gafni
Technion
Haifa, Israel
yotam.gafni@campus.technion.ac.il

Moshe Tennenholtz
Technion
Haifa, Israel
moshet@ie.technion.ac.il

In robust decision making under uncertainty, a natural choice is to go with safety (aka security) level strategies. However, in many important cases, most notably auctions, there is a large multitude of safety level strategies, thus making the choice unclear. We consider two refined notions:

- a term we call DSL (distinguishable safety level), and is based on the notion of "discrimin" [7], which uses a pairwise comparison of actions while removing trivial equivalencies. This captures the fact that when comparing two actions an agent should not care about payoffs in situations where they lead to identical payoffs.
- The well-known Leximin notion from social choice theory, which we apply for robust decision-making. In particular, the leximin is always DSL but not vice-versa [7].

We study the relations of these notions to other robust notions, and illustrate the results of their use in auctions and other settings. Economic design aims to maximize social welfare when facing self-motivated participants. In online environments, such as the Web, participants' incentives take a novel form originating from the lack of clear agent identity—the ability to create Sybil attacks, i.e., the ability of each participant to act using multiple identities. It is well-known that Sybil attacks are a major obstacle for welfare-maximization. Our main result proves that when DSL attackers face uncertainty over the auction's bids, the celebrated VCG mechanism is welfare-maximizing even under Sybil attacks. Altogether, our work shows a successful fundamental synergy between robustness under uncertainty, economic design, and agents' strategic manipulations in online multi-agent systems.

1 Introduction

Consider an agent who needs to decide on her action in an environment consisting of other agents. In certain cases there is a uniquely defined optimal action for the agent, but in most cases this "agent perspective" is an open challenge. Given the above, both AI and economics care about an adequate modeling of an agent, and its ramifications in a variety of multi agent contexts, for example, on social welfare.

We consider a notion for agent modeling we term DSL (Distinguishable Safety-Level). The notion was previously suggested in the context of constraint-satisfaction problems and fuzzy logic, and was termed "discrimin" [7]. In game theoretic settings, the notion was previously applied [6] as a solution concept for bargaining in Boolean games [11]. To the best of our knowledge, it was not previously considered in the context of auctions, voting, and more generally mechanism design, i.e., when considering the robustness of economic mechanisms' performance when facing strategic agents.

There are two ways to think of the DSL solution concept, when applied to agent modeling. One is as a solution concept adapted to capture the behavioral phenomenon of the loss aversion cognitive bias in agents, particularly when probabilities over nature states are unknown. The other is as a form of robust strategy choice under uncertainty, that may be required in volatile and unpredictable environments that

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 245–259, doi:10.4204/EPTCS.379.20

do not admit a stable Bayesian description. We show its usefulness in auctions. In the full version of this paper, we also study its behavior in other prominent strategic settings, such as voting. In our main result we consider the celebrated welfare maximizing VCG mechanism in combinatorial auctions setting, where it is known to fail under false name (aka Sybil) attacks. We show that DSL agents lead to optimal social welfare.

1.1 Reasoning under Uncertainty

A classic distinction [15] separates reasoning under risk, where the actors are rational and there is a commonly known distribution about their environment (also known as the stochastic or Bayesian setting), and reasoning under uncertainty, where the general structure of strategies and outcomes is known, but there is no probabilistic information about the environment. Moreover, even assumptions regarding actors' rationality or behavior characteristics may not be guaranteed . For such cases, a robust or worst-case approach seems appropriate, and various notions exist to capture it. Ideally, a dominant strategy solution exists, but this is usually not the case (and indeed it is not the case in all the cases we analyze in this paper). A minimal robust notion is that of a safety level strategy, which uses a max-min approach over all possible outcomes given a strategy choice. However, though it yields interesting results in some cases [2, 20], in many other cases it does not tell us much about what strategy to choose, in particular in auctions settings, where we derive our most interesting results. As we see, this is because in auctions the natural safety level is 0 (which happens when the bidder loses the auction), and any strategy that does not overbid guarantees it. It is thus hard to choose among these strategies without considering a more refined notion. Existing refined notions are the lexicographic max-min (originally defined in [19]) and min-max regret [18]. We overview their comparison to the notion of DSL in Section 3 and Appendix A, respectively.

1.2 VCG, Sybil Attacks, and Welfare

VCG is a well known mechanism which can be applied for combinatorial auctions. VCG has good qualities such as being dominant strategy incentive compatible and achieving optimal social welfare. However, under the possibility of false-name attacks [22], it is no longer truthful. Coming up with other mechanisms does not solve the basic conundrum: In the full information settings, any false-name proof mechanism performs poorly in terms of welfare [12].

A possible avenue to solving the issue is by limiting the discussed valuation classes. However, an example in [13] shows that even when all bidders have sub-modular valuations, VCG is no longer dominant strategy incentive-compatible under false-name attacks. Notably though, even with this example, VCG still arrives at the socially optimal allocation, and in fact as [1] show, this observation is true in general up to a constant with sub-modular (and near sub-modular) bidders. However, in the full version of our paper, we show an example where for the XOS valuation class, which extends the sub-modular class, there is such an attack so that VCG arrives at an arbitrarily sub-optimal allocation. The attack we describe is enabled by the full information settings. Without full information, the attack is risky for the attacker, since it could lead to negative utility, as the attacker overbids her true valuation.

A useful approach, that can lead to better welfare guarantees than dominant strategy mechanism design, is Bayesian mechanism design. Assuming that the bidder distributions are common knowledge, recent work has shown that selling each item separately leads to good constant approximation welfare guarantees for XOS [4] and sub-additive [8] valuations. Though the works do not explicitly consider

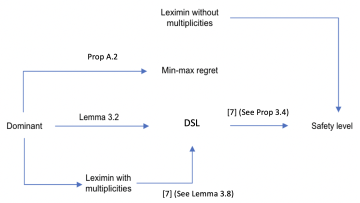


Figure 1: Hierarchy for robust decision under uncertainty

false-name attacks, their constructions use the false-name-proof first and second price auctions to auction items separately, and so their results naturally extend to Bayesian false-name mechanism design.

It is important to note, that many of the above positive results for welfare guarantees under false-name attack assume some form of risk-aversion; most importantly, that bidders do not overbid, i.e., they choose only strategies that are individually rational (under any possible nature state). This condition is equivalent to limiting the strategy space only to safety level strategies (as in this case of combinatorial auctions, the safety level is 0). In [10] the authors do not make this assumption, but their positive welfare optimality results are limited as they only consider the homogeneous single-minded case with two items. We thus believe that it is natural to ask: Under our definition of DSL, which is a strong risk-aversion notion (compared, e.g., to the safety level strategy), what welfare guarantees can be obtained? Surprisingly, the answer is optimal, as we show in our main result in Theorem 4.2.

1.3 Our Results

In Section 2 we formally define our solution concept, and apply it to the first-price and discrete first-price auctions. In Section 3 (with the additional discussion of min-max regret in Appendix A) we describe a hierarchy of solution concepts and their relations to the solution concept we introduce (DSL), as summarized in Figure 1.

In Section 4, we present our main result. We discuss VCG as a combinatorial auction under false-name attacks, when bidders may create shill identities to send bids. It is known that VCG is not dominant strategy truthful in these settings, and previous results were limited in establishing good welfare guarantees for combinatorial auctions generally under false-name attacks. We show that when bidders use DSL strategies, VCG achieves optimal welfare even under the threat of false-name attacks.

2 DSL: Definition

When defining DSL strategies, we take the perspective of a single agent i facing uncertainty. The agent has a utility function u_i that determines her utility given the state of the world, which is comprised of her own action a_i, others' actions a_{-i}, and agent i's type θ_i. Formally, $u_i(a_i, a_{-i}|\theta_i)$. We denote by A_i the set of all agent i's pure actions, and by $\Delta(A_i)$ the set of all agent i's mixed actions. An action a_i may be from either of these action sets depending on the context. For mixed strategies, $u_i(a_i, a_{-i}|\theta_i) = \mathbb{E}_{a \sim a_i}[u_i(a, a_{-i}|\theta_i)]$. We denote by Θ_i the set of all agent i's types.

Definition 2.1. *We say that an action a_i of agent i is **DSL** (given a type θ_i) if for any other action a_i', over the set of outcomes where agent i's utility differs between the actions, the minimal utility attained using a_i is at least as good as that attained by a_i'. Formally, let*

$$D_{\theta_i}(a_i, a_i') = \{a_{-i} \quad s.t. \quad u_i(a_i, a_{-i}|\theta_i) \neq u_i(a_i', a_{-i}|\theta_i)\}.$$

Then, an action a_i is pure/mixed DSL if $\forall a'_i \in A_i$,[1]

$$\min_{a_{-i}\in D_{\theta_i}(a_i,a'_i)} u_i(a_i,a_{-i}|\theta_i) \geq \min_{a_{-i}\in D_{\theta_i}(a_i,a'_i)} u_i(a'_i,a_{-i}|\theta_i).^2$$

We say that a strategy $s_i : \Theta_i \to A_i$ is pure DSL if it maps any type θ_i to a corresponding DSL pure action a_i. We say that $s_i : \Theta_i \to \Delta(A_i)$ is mixed DSL if it maps any type θ_i to a corresponding DSL mixed action a_i.

Notice that in our definition we compare pure strategies only with other pure strategies, i.e., they are DSL with respect to this strategy set. Mixed strategies are DSL w.r.t. all strategies (mixed and pure). We use the term "nature state" to mean the actions a_{-i}, which may result from either uncertainty over others' types or over their strategic choice: What matters to the agent in the end is what are all of their possible actions. There is seemingly some loss of generality in that we assume that all possible a_{-i} are fixed vectors of actions, and not more generally random variables over actions. But, as we show in the full version of our paper, allowing for the latter loses the usefulness of the DSL notion.

3 Relations to Prominent Game-theoretic Solution Concepts

Note: Missing proofs in this section appear in the full version of our paper. For completeness, we state the connection of DSL to safety level and what we call Multi-Leximin strategies, although these claims are already established in the literature characterizing the notion of discrimin (see, e.g., [7]).

3.1 Dominant Strategy

Definition 3.1. *A weakly dominant action a_i satisfies that for any other action a'_i:*
(1) For any nature state a_{-i},

$$u_i(a_i,a_{-i}|\theta_i) \geq u_i(a'_i,a_{-i}|\theta_i),$$

and (2) there is such nature state a_{-i} so that the above inequality is strict.
A weakly dominant strategy is such that maps types to weakly dominant actions.

The following result is natural:

Lemma 3.2. *Every weakly dominant strategy is DSL.*

3.2 Safety Level Strategy and Individual Rationality

Safety level strategies in non-cooperative games are such strategies that yield a best possible guarantee of utility for a player, without the need to reason about the types or strategies chosen by other players. The example of [2] makes a compelling argument for choosing such strategies: There are games where the Nash Equilibrium does not guarantee more than the safety level. In such cases, choosing the equilibrium strategy runs the unnecessary risk of a lower outcome. [21] extends this insight and shows a class of games where the safety level strategy guarantees a large constant fraction of the Nash equilibrium outcome, without its involved risks.

[1] Or, in the mixed case: $\forall a'_i \in \Delta(A_i)$

[2] We use the term minimum loosely: When taken over infinite sets that do not have a minimum the definition uses the infimum.

Individual rationality is a common requirement in game theory analysis (see, e.g., [17]), that requires either that an agent does not participate in a game where it gains negative utility, or that it does not choose a strategy that may yield negative outcomes. We define:

Definition 3.3. *A safety level strategy s_i [2] is a strategy (mixed or pure) of player i such that for any type θ_i it chooses an action a_i so that for any nature state a_{-i} of the other agents, $u_i(a_i, a_{-i}|\theta_i) \geq \max_a \min_{a_{-i}} u_i(a, a_{-i}|\theta_i)$. I.e., the strategy guarantees the safety level $L \overset{def}{=} \max_{a'} \min_{a'_{-i}} u_i(a', a'_{-i}|\theta_i)$.*

Individual Rationality of a strategy s_i of player i satisfies that for any type θ_i it chooses an action a_i so that for any nature state a_{-i} of the other agents, $u_i(a_i, a_{-i}|\theta_i) \geq 0$. I.e., the strategy guarantees a non-negative utility for the player.

The two notions are quite similar, as individual rationality can be seen as a minimal safety level requirement; in auctions they are in fact equivalent to a third notion of no over-bidding, under some reasonable conditions (the auction does not charge payments from non-winners, and never charges a winner more than her declared value). We claim:

Proposition 3.4. *A DSL strategy is a safety level strategy, but not necessarily vice-versa.*

Corollary 3.5. *When there is a finite amount of safety level strategies, and a finite amount of nature states, a DSL strategy is guaranteed to exist.*

The corollary is a result of Lemma 3.8 and Lemma 3.9. We prove both during our discussion of the lexicographic max-min in the next subsection.

3.3 Lexicographic Max-min

A very interesting comparison is with another robust solution notion, the lexicographic max-min (also commonly known as leximin). The leximin is especially prevalent in the fair allocation literature, see, e.g., [16]. We consider two possible ways to define it:

Definition 3.6. *Leximin - Let U_{a_i} be the **set** of all possible utility outcomes of the action a_i by agent i, ordered from small to large, and let $U_{a_i}[j]$ be the j element of U_{a_i} in this ordering. An action a_i lexicographically weakly dominates (LD) another action a'_i if $\min U_{a_i} > \min U_{a'_i}$, or $\min U_{a_i} = \min U_{a'_i}$ and $U_{a_i} \setminus \min U_{a_i}$ LDs $U_{a'_i} \setminus \min U_{a'_i}$ (a recursive definition). We call an action that LDs all other actions a leximin. A strategy is leximin if it maps all types to leximin actions.*

*Multi-Leximin - Let U_{a_i} be the **multiset** of all possible utility outcomes of the action a_i by agent i, ordered from small to large. The rest of the definition follows similarly, where importantly in the recursive definition we remove only one copy of the minimum element at each step.*

Note that the (Multi-)leximin notions are only clearly defined when there is a finite amount of nature states a_{-i}, otherwise the recursive definition of LD may not terminate.

We first note that both definitions give stronger notions than safety level strategies.

Lemma 3.7. *(Multi-)leximin is a safety level strategy, but not necessarily vice-versa.*

Despite some similarity in the definition with DSL, the notion of leximin does not have a special relationship with it: neither implies the other. We demonstrate it using the discrete first-price auction in Example B.7 Appendix B.

The notion of multi-leximin is much more closely related to the DSL notion. In fact, it is a stronger notion:

Lemma 3.8. *Multi-leximin is a DSL strategy, but not necessarily vice-versa.*

Lemma 3.9. *When there is a finite amount of safety level strategies, multi-leximin is guaranteed to exist.*

An important advantage of the DSL definition is that it naturally extends to settings with continuous outcomes. It is not clear how to extend the leximin definition to such cases. Thus, one possible way of thinking about the DSL notion is that it is a somewhat weaker notion of multi-leximin, that can be used in continuous settings, as well as discrete ones.

4 Main Result: Application to VCG under False-name Attacks

We now move on to present our main result and through it the usefulness of the DSL notion. False-name attacks by an agent i in a combinatorial auction are where instead of sending one combinatorial bid, the agent sends multiple combinatorial bids (a vector $\mathbf{b_i}$ rather than a single bid b_i). The agent then gets all the items allocated to the "agents" (which we call Sybil agents or Sybil bids) $1 \leq j \leq |\mathbf{b_i}|$, and pays the sum of all their payments. Before formally introducing the VCG notations, we note three complexities that are present in our notations: (1) We consider both the notion of DSL strategies (which has the single agent perspective vs. nature states) and social welfare (which accounts for n different agents). (2) We consider welfare for the real n underlying agents of the auction, but since each may use Sybil identities, the VCG allocations are in terms of the Sybil identities. We allow for both by using sub-indexing. (3) Similar to the case of the first-price auction, discretization of the bid space is essential to the result (a counter-example for continuous VCG appears in the full version of our paper. To further simplify the proof, we also assume that the valuation space is discrete, though this assumption can be removed. We allow more granularity to the bid space: valuations are on an ε grid, while bids are on an $\frac{\varepsilon}{2|M|!}$ grid.

Definition 4.1. *$Grid(\varepsilon) = \{\varepsilon k\}_{k \in \mathcal{N}} = \{0, \varepsilon, 2\varepsilon, \ldots\}$. A combinatorial bid $b \in B$ over an item set M is a function $b : P(M) \to Grid(\frac{\varepsilon}{2|M|!})$ from the power set of all subsets of M to a non-negative bid value. A combinatorial valuation v is similarly $v : P(M) \to Grid(\varepsilon)$. With the possibility of Sybil attacks, an agent i with valuation (type) v_i sends a vector of bids (action) $\mathbf{b_i} \in B^*$ (i.e., any amount of combinatorial bids), and faces a nature state $\mathbf{b_{-i}} \in B^*$.*

Let $\eta_i = |\mathbf{b_i}|, \eta_{-i} = |\mathbf{b_{-i}}|$ be the number of (Sybil) agents in each vector. An allocation $\alpha^S(\mathbf{b_i}, \mathbf{b_{-i}})$ maps the bid vectors into a partition of S into subsets. We allow indexing $\alpha_1, \ldots, \alpha_n$ to mean the union of items allocated to the Sybil identities of each real agent, as well as sub-indexing $\alpha_{i_1}, \ldots, \alpha_{i_{\eta_i}}$ to mean the items allocated to a specific Sybil identity of agent i. We denote $SW_\alpha^{Obs} = \sum_{i=1}^n \sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}), SW_\alpha^{Real} = \sum_{i=1}^n v_i(\alpha_i)$ for the observed social welfare of an allocation as specified in the (possibly Sybil) bids, and the real social welfare of the agents, respectively. We denote $truth_i = v_i$ for the truthful bid.

The VCG combinatorial auction is the pair of allocation rule

$$\alpha^M(\mathbf{b_i}, \mathbf{b_{-i}}) = \underset{\mathbf{\tilde{a}^M}(\mathbf{b_i}, \mathbf{b_{-i}})}{\arg\max} \left(SW^{Obs}_{\mathbf{\tilde{a}^M}(\mathbf{b_i}, \mathbf{b_{-i}})}\right),$$

and the payment rule

$$p^M_{i_j}(\mathbf{b_i}, \mathbf{b_{-i}}) = SW^{Obs}_{\alpha^M}(\mathbf{b_i}, \mathbf{b_{-i}}) - SW^{Obs}_{\alpha^{M \setminus a^M_{i_j}}}(\mathbf{b_i}, \mathbf{b_{-i}}).$$

Finally, the utility of agent i is $u_i(\mathbf{b_i}, \mathbf{b_{-i}} | v_i) = v_i \left(\bigcup_{1 \leq j \leq \eta_i} a^M_{i_j}(\mathbf{b_i}, \mathbf{b_{-i}}) \right) - \sum_{j=1}^{\eta_i} p^M_{i_j}(\mathbf{b_i}, \mathbf{b_{-i}}).$

Theorem 4.2. *When all bidders play DSL strategies, discrete VCG achieves optimal welfare, even under the possibility of false-name attacks and with general valuations.*

Proof. Our proof follows the following structure: First, we define overbidding Sybil attacks and show that they are not DSL. We then define underbidding attacks and show that they are not DSL. For any of the remaining attacks, which we call exact-bidding (bidding truthfully is also exact-bidding, but not exclusively so), we show that even though they are not necessarily truthful, they yield maximal welfare. However, this still does not guarantee that one of the remaining strategies is in fact DSL. For this purpose, we show that there exists a DSL strategy: being truthful[3].

First, we show that if the Sybil bids are overbidding v (in a sense that will be immediately defined), then, similarly to our proof for the first-price auction (see Appendix B), it is not safety level and thus not DSL. This requires slightly more care since the bids are combinatorial and there are several Sybil bids. We say that $\mathbf{b_i} = (b_{i_1}, \dots, b_{i_{\eta_i}})$ is overbidding if there is a set S and an allocation $\alpha^S(\mathbf{b_i})$ so that $\sum_{j=1}^{\eta_i} b_{i_j}(\alpha^S_{i_j}(\mathbf{b_i})) > v_i(S)$.

Claim 4.3. *Overbidding $\implies$ not DSL.*

Proof. Let us choose a maximizing allocation $\alpha^S(\mathbf{b_i})$ for S.

We denote $\bar{b} = \max_{1 \leq j \leq \eta_i} \max_{S' \subseteq M} b_{i_j}(S') + v_i(S')$ for a number high enough that if some other agent bids it for any subset of M, both the truthful bid v_i or the Sybil attack $\mathbf{b_i}$ will lose that subset. We will use it in our construction of nature states. By the overbidding condition, we can take the average $\tilde{b} = \frac{v_i(S)}{2} + \frac{1}{2}\sum_{j=1}^{\eta_i} b_{i_j}(a^S_{i_j})$, so that $\sum_{j=1}^{\eta_i} b_{i_j}(\alpha^S_{i_j}) > \tilde{b} > v_i(S)$. Consider a nature state where the false-name attacker faces exactly one additive bidder b' that has for any good $g \in M \setminus S$, $b'(g) = \bar{b}$, and for any good $g \in S$, $b'(g) = \frac{\tilde{b}}{|S|}$. The optimal observed welfare allocation is to allocate all goods in $M \setminus S$ to b', and allocate the set S as in $\alpha_{\mathbf{i}}(S)$. The payment of bidder b_i must be at least $b'(S) = \tilde{b} > v_i(S)$. Therefore, the attacker has negative utility in this case, while truthfulness is individually rational: i.e., it is not a safety level strategy and so also not DSL. □

We say that $b_{i_1}, b_{i_{\eta_i}}$ are underbidding if there is a set S so that for any allocation $\alpha^S \overset{def}{=} \alpha^S(\mathbf{b_i})$ so that $\sum_{j=1}^{\eta_i} b_{i_j}(\alpha^S_{i_j}) < v_i(S)$.

Claim 4.4. *Underbidding $\implies$ not DSL.*

Proof. Let $\tilde{b} = \frac{1}{2}\sum_{j=1}^{\eta_i} b_{i_j}(\alpha^S_{i_j}) + \frac{v_i(S)}{2}$, then

$$\sum_{j=1}^{\eta_i} b_{i_j}(\alpha^S_{i_j}(\mathbf{b_i})) < \tilde{b} < v_i(S).$$

Let b' be constructed as in the overbidding case. The allocation $\alpha^M(\mathbf{b_i}, b')$ allocates no items to the Sybil bidders of agent i. However, the allocation given agent i bids truthfully $\alpha^M(truth_i, b')$, allocates the set S to her with payment $\tilde{b}$, which yields agent i a positive utility $v_i(S) - \tilde{b}$. This yields

$$\min_{\mathbf{b_{-i}} \in D_{v_i}(\mathbf{b_i}, truth_i)} u_i(\mathbf{b_i}, \mathbf{b_{-i}} | v_i) = 0.$$

On the other hand, we claim that since we know DSL strategies are not overbidding, there are no nature states for which an underbidding Sybil attack gets positive utility while bidding truthfully gets 0 utility. Assume towards contradiction $truth_i$ gets 0 utility. It then either does not win any item, or wins some set

[3] Another, albeit non-constructive method to show there exists a DSL strategy is by showing the finiteness of undominated exact-bidding Sybil attacks, and then use Corollary 3.5

S and pays $v_i(S)$ for it. Let S be the set that the Sybil bidders win to gain positive utility. As there is no overbidding, this set can be won by $truth_i$ as well (in the respective maximizing allocation)[4]. Then,

$$\begin{aligned}
& SW^{Obs}_{\alpha^{M\setminus S}}(\mathbf{b_i},\mathbf{b_{-i}}) && \text{(No overbidding)}\\
& = SW^{Obs}_{\alpha^{M\setminus S}}(truth_i,\mathbf{b_{-i}}) && \text{(i wins only S)}\\
& = SW^{Obs}_{\alpha^{M\setminus S}}(\mathbf{b_{-i}}) && \text{(i's truthful payment)}\\
& = SW^{Obs}_{\alpha^{M}}(\mathbf{b_{-i}}) - v_i(S) && \text{(More bids)}\\
& \leq SW^{Obs}_{\alpha^{M}}(\mathbf{b_i},\mathbf{b_{-i}}) - v_i(S)
\end{aligned}$$

So

$$v_i(S) \leq SW^{Obs}_{\alpha^M(\mathbf{b_i},\mathbf{b_{-i}})} - SW^{Obs}_{\alpha^{M\setminus S}(\mathbf{b_i},\mathbf{b_{-i}})} \tag{1}$$

Since our choice of S assumes the Sybil bids win exactly it, we have

$$SW^{Obs}_{\alpha^{M\setminus S}(\mathbf{b_i},\mathbf{b_{-i}})} + SW^{Obs}_{\alpha^{S}(\mathbf{b_i})} = SW^{Obs}_{\alpha^M(\mathbf{b_i},\mathbf{b_{-i}})},$$

and so, together with Eq. 1,

$$v_i(S) \leq SW^{Obs}_{\alpha^M(\mathbf{b_i},\mathbf{b_{-i}})} - SW^{Obs}_{\alpha^{M\setminus S}(\mathbf{b_i},\mathbf{b_{-i}})} = SW^{Obs}_{\alpha^{S}(\mathbf{b_i})}.$$

Since there is no overbidding, $SW^{Obs}_{\alpha^{S}(\mathbf{b_i})} = v_i(S)$.

We now show that any Sybil bidder j pays $v_i(S) - \sum_{1\leq t\neq j\leq \eta_i} b_{i_j}(\alpha_{i_j})$. Since S is allocated to the Sybil bidders and $M\setminus S$ to others,

$$\begin{aligned}
& SW^{Obs}_{\alpha^{M\setminus a^M_{i_j}}}(\mathbf{b_i},\mathbf{b_{-i}}) = \\
& SW^{Obs}_{\alpha^{M\setminus S}}(\mathbf{b_i},\mathbf{b_{-i}}) + SW^{Obs}_{\alpha^{S\setminus a^M_{i_j}}}(\mathbf{b_i},\mathbf{b_{-i}})
\end{aligned} \tag{2}$$

Then,

$$\begin{aligned}
p^M_{i_j} &= SW^{Obs}_{\alpha^M}(\mathbf{b_i},\mathbf{b_{-i}}) - SW^{Obs}_{\alpha^{M\setminus a^M_{i_j}}}(\mathbf{b_i},\mathbf{b_{-i}})\\
&= SW^{Obs}_{\alpha^M}(\mathbf{b_i},\mathbf{b_{-i}}) - SW^{Obs}_{\alpha^{M\setminus S}}(\mathbf{b_i},\mathbf{b_{-i}})\\
&\qquad\qquad - SW^{Obs}_{\alpha^{S\setminus a^M_{i_j}}}(\mathbf{b_i},\mathbf{b_{-i}})\\
&= v_i(S) - SW^{Obs}_{\alpha^{S\setminus a^M_{i_j}}}(\mathbf{b_i},\mathbf{b_{-i}})\\
&= v_i(S) - \sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j})
\end{aligned}$$

The total payment of agent i is then

$$\begin{aligned}
& \sum_{j=1}^{\eta_i} p^M_{i_j} = \sum_{j=1}^{\eta_i} v_i(S) - \sum_{1\leq t\neq j\leq \eta_i} b_{i_j}(\alpha_{i_j}) = \\
& \eta_i \cdot v_i(S) - (\eta_i - 1)\sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}) = v_i(S).
\end{aligned}$$

[4] In full generality, $truth_i$ may win a set s that has partial intersection with S. The analysis of this case is essentially the same, and stems from the fact that the alternative value for the items forces zero utility on the truthful agent.

This concludes that whenever the utility of $truth_i$ is 0, then the utility for the Sybil attack is 0 as well.

In any other case, the utility of $truth_i$ must be strictly positive, and since the bids are discrete the minimum over all these cases satisfies

$$\min_{\mathbf{b_{-i}} \in D_{v_i}(\mathbf{b_i}, truth_i)} u_i(truth_i, \mathbf{b_{-i}} | v_i) \geq \frac{1}{2|M|!}.$$

Therefore, underbidding is not DSL. □

We consider *exact-bidding* such Sybil bids that have for any set of items S, $\max_{\alpha^S(\mathbf{b_i})} \sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}(S)) = v_i(S)$. These are exactly all the Sybil attacks that are neither overbidding nor underbidding. $truth_i$ is also exact-bidding.

Claim 4.5. *Exact-bidding* $\implies$ *optimal welfare.*

Proof. Consider an allocation $\alpha_F \overset{def}{=} \alpha^M(\mathbf{b_i}, \mathbf{b_{-i}})$ attained when all players choose an exact-bidding attack, vs $\alpha_T \overset{def}{=} \alpha^M(truth_i, \mathbf{truth_{-i}})$. We have

$$\begin{aligned} SW_{\alpha_T}^{Real} \leq & \quad \text{(Truthful)} \\ SW_{\alpha_T}^{Obs} \leq & \quad \text{(No underbidding)} \\ SW_{\alpha_F}^{Obs} \leq & \quad \text{(No overbidding)} \\ SW_{\alpha_F}^{Real} & \end{aligned}$$

In words, since there is no underbidding in the Sybil attack, if we take the set allocated to each agent i under the allocation that maximizes welfare under truthfulness, there are Sybil bidders $i_{j_1}, \dots, i_{j_k}$ with the same aggregate valuation for it. So, $SW_{\alpha_F}^{Obs}$ is lower bounded by the optimal truthful welfare. Since there is also no overbidding, whatever allocation is chosen as α_F^{Obs} is at least as good to each agent i as is declared. □

Claim 4.6. *$truth_i$ is DSL.*

Proof. Consider some exact-bidding Sybil attack $\mathbf{b_i}$.

Case 1: There is a set S so that $\forall 1 \leq j \leq \eta_i, b_{i_j}(S) < v_i(S)$. Then, by the exact-bidding condition there must be some allocation $\alpha^S(\mathbf{b_i})$ (with at least two non-empty allocations $\alpha_{i_j}^S$) so that $\max_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}^S) < \sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}^S) = v_i(S)$. Consider the nature state where there is one bid b' so that $b'(\alpha_{i_j}^S) = v_i(\alpha_{i_j}^S)$ for any $1 \leq j \leq \eta_i$, and the rest of the sets are defined upward-monotonely: They inherit the largest value of a subset. With this nature state, the Sybil attack has utility 0. On the other hand, $truth_i$ has positive utility of $v_i(S) - \max_{j=1}^{\eta_i} v_i(\alpha_{i_j}) > 0$. Since $truth_i$ is individually rational, it is thus DSL w.r.t. such Sybil attacks.

Case 2: For every set S, there is such j' with $b_{i_{j'}}(S) = v_i(S)$. It must hold by the exact-bidding condition that for any allocation $\alpha^S(\mathbf{b_i})$, $\sum_{j=1}^{\eta_i} b_{i_j}(\alpha_{i_j}^S) \leq v_i(S) = b_{i_{j'}}(S)$. We may assume that VCG prefers to assign larger bundles when tie-breaking between possible assignments. Then, it must be that any allocation to the Sybil bidders is given to one Sybil bidder as a whole bundle. It is then weakly better to send only $b_{i_{j'}}$ as a single bid instead of $\mathbf{b_i}$. Furthermore, it is then weakly better to send $truth_i$, since truthfulness is dominant for single bid VCG. Since this is true given any nature state, the Sybil attack is weakly dominated by $truth_i$, which implies $truth_i$ is DSL with respect to it.

This covers all the exact-bidding Sybil attacks. DSL strategies with respect to overbidding and underbidding attacks are implied by the relevant discussion. Overall this covers all Sybil attacks. □

□

5 Discussion and Future Directions

In the example of the discrete first-price auction in Section 2, as well as in our main result in Section 4, the DSL solution concept leads to optimal results: truthfulness (or near truthfulness), and optimal revenue or welfare. In Appendix B we study the first-price auction, and show similar results for its discrete variant. However, for the classic setting of voting, we show in the full version of this paper that this is not the case, and that solutions may have various surprising forms.

A robust notion missing from our discussion in Section 3 is min-max regret. We show in Appendix A it does not imply or is implied by our notion of DSL, and give further characteristics of it. It is also compared with our notion as part of our discussion of the discrete first-price auction in Appendix B.

In our definition of DSL, we consider only pure nature states. We justify this choice in Appendix C of the full version, by showing that if we consider mixed nature states as well, then the DSL and safety level notions become one. In Appendix D of the full version, we show a possible refinement of our notion of DSL, and demonstrate why it may be useful.

A few immediate open questions follow our work:

- We find that DSL is a stronger notion than safety level. In settings previously studied that proved performance guarantees for safety level strategies, do DSL strategies exist? Can they yield better performance guarantees?
- In the case of single-item auctions, our analysis of the discrete first price auction implies that with DSL bidders, it is possible to achieve optimal welfare and revenue. Does this extend to combinatorial auctions? If so, does it hold even when the discretization must be polynomially bound?
- In the presence of partial knowledge or the option to elicitate it (similar to the ideas in [14]), what would the DSL action be? This is relevant, for example, when agents arrive sequentially, and so the set of feasible nature states diminishes for later agents.

Acknowledgements

Yotam Gafni and Moshe Tennenholtz were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 740435).

References

[1] Colleen Alkalay-Houlihan & Adrian Vetta (2014): *False-Name Bidding and Economic Efficiency in Combinatorial Auctions*. Proceedings of the AAAI Conference on Artificial Intelligence 28(1), doi:10.1609/aaai.v28i1.8828.

[2] Robert J Aumann (1985): *On the non-transferable utility value: A comment on the Roth-Shafer examples*. Econometrica: Journal of the Econometric Society, pp. 667–677, doi:10.2307/1911661.

[3] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia (2016): *Handbook of Computational Social Choice*, 1st edition. Cambridge University Press, USA, doi:10.1017/CBO9781107446984.

[4] George Christodoulou, Annamária Kovács & Michael Schapira (2016): *Bayesian Combinatorial Auctions*. Journal of the ACM 63(2). Available at `https://doi.org/10.1145/2835172`.

[5] Michael Suk-Young Chwe (1989): *The discrete bid first auction*. Economics Letters 31(4), pp. 303–306, doi:10.1016/0165-1765(89)90019-0.

[6] Sofie De Clercq, Steven Schockaert, Ann Nowé & Martine De Cock (2018): *Modelling incomplete information in Boolean games using possibilistic logic*. International Journal of Approximate Reasoning 93, pp. 1–23, doi:10.1016/j.ijar.2017.10.017.

[7] Didier Dubois & Philippe Fortemps (1999): *Computing improved optimal solutions to max–min flexible constraint satisfaction problems*. European Journal of Operational Research 118(1), pp. 95–126, doi:10.1016/S0377-2217(98)00307-5.

[8] Michal Feldman, Hu Fu, Nick Gravin & Brendan Lucier (2013): *Simultaneous Auctions Are (Almost) Efficient*. In: Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC '13, Association for Computing Machinery, New York, NY, USA, p. 201–210. Available at `https://doi.org/10.1145/2488608.2488634`.

[9] John A. Ferejohn & Morris P. Fiorina (1974): *The Paradox of Not Voting: A Decision Theoretic Analysis*. American Political Science Review 68(2), p. 525–536, doi:10.2307/1959502.

[10] Yotam Gafni, Ron Lavi & Moshe Tennenholtz (2020): *VCG under Sybil (False-Name) Attacks - A Bayesian Analysis*. Proceedings of the AAAI Conference on Artificial Intelligence 34(02), pp. 1966–1973, doi:10.1609/aaai.v34i02.5567.

[11] Paul Harrenstein, Wiebe van der Hoek, John-Jules Meyer & Cees Witteveen (2001): *Boolean games*. In: Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge, pp. 287–298.

[12] Atsushi Iwasaki, Vincent Conitzer, Yoshifusa Omori, Yuko Sakurai, Taiki Todo, Mingyu Guo & Makoto Yokoo (2010): *Worst-Case Efficiency Ratio in False-Name-Proof Combinatorial Auction Mechanisms*. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1, AAMAS '10, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 633–640.

[13] Benny Lehmann, Daniel Lehmann & Noam Nisan (2006): *Combinatorial auctions with decreasing marginal utilities*. Games and Economic Behavior 55(2), pp. 270–296, doi:10.1016/j.geb.2005.02.006.

[14] Tyler Lu & Craig Boutilier (2011): *Robust Approximation and Incremental Elicitation in Voting Protocols*. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume One, IJCAI'11, AAAI Press, p. 287–293, doi:10.5591/978-1-57735-516-8/IJCAI11-058.

[15] S. Merrill (1982): *Strategic Voting in Multicandidate Elections under Uncertainty and under Risk*. In Manfred J. Holler, editor: Power, Voting, and Voting Power, Physica-Verlag HD, Heidelberg, pp. 179–187.

[16] Hervé Moulin (2003): *Fair Division and Collective Welfare*. The MIT Press. Available at `https://doi.org/10.7551/mitpress/2954.001.0001`.

[17] Noam Nisan, Tim Roughgarden, Éva Tardos & Vijay V. Vazirani (2007): *Algorithmic Game Theory*. Cambridge University Press, doi:10.1017/CBO9780511800481.

[18] L. J. Savage (1951): *The Theory of Statistical Decision*. Journal of the American Statistical Association 46(253), pp. 55–67, doi:10.1080/01621459.1951.10500768.

[19] A. Sen (1970): *Collective Choice and Social Welfare*. Mathematical economics texts, Holden-Day. Available at `https://books.google.co.il/books?id=kzq6AAAAIAAJ`.

[20] Moshe Tennenholtz (2001): *Rational Competitive Analysis*. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 1067–1072.

[21] Moshe Tennenholtz (2002): *Competitive safety analysis: Robust decision-making in multi-agent systems*. Journal of Artificial Intelligence Research 17, pp. 363–378, doi:10.1613/jair.1065.

[22] Makoto Yokoo, Yuko Sakurai & Shigeo Matsubara (2004): *The effect of false-name bids in combinatorial auctions: New fraud in Internet auctions*. Games and Economic Behavior 46(1), pp. 174–188, doi:10.1016/S0899-8256(03)00045-9.

A Min-max Regret

Another robust solution notion is the min-max regret [18]. The notion has many uses in voting: [9] showed it can be used to explain why voters choose to participate in elections, and [15] used it to "resolve" the Gibbard-Satterthwaite impossibility theorem (see, e.g., [3]), by showing that plurality voting (for example) is truthful under this notion. [14] showed how when only partial preferences are known, voting rules can use this notion to decide a winner, and design good elicitation schemes.

Definition A.1. *Regret for an action a_i and nature state a_{-i} given a type θ_i is*

$$Reg(a_i, a_{-i}|\theta_i) = \max_{a'_i} u(a'_i, a_{-i}|\theta_i) - u(a_i, a_{-i}|\theta_i).$$

Max regret for an action a_i given a type θ_i is

$$Reg(a_i, a_{-i}|\theta_i).$$

A min-max regret action belongs to

$$\arg\min_{a_i} \max_{a_{-i}, a'_i} u(a'_i, a_{-i}|\theta_i) - u(a_i, a_{-i}|\theta_i).$$

In words, the regret of an action a_i under nature state a_{-i} is the maximal lost utility $u(a'_i, a_{-i}|\theta_i) - u(a_i, a_{-i}|\theta_i)$ of choosing a_i instead of a'_i, over all possible actions a'_i (this regret is non-negative, as there is always the option of choosing a_i itself). Max regret is the maximal such regret over all nature states, and the min-max regret action is the action a_i that has minimal max regret.

Proposition A.2. *Dominant strategy $\implies$ min-max regret*

Proof. Consider a dominant strategy s, fix a type θ_i, and let $a = s(\theta_i)$. For any a', a_{-i}, we have that $u(a', a_{-i}) - u(a, a_{-i}) \leq 0$, i.e., the max regret for a is 0, the minimum possible, and so a is a min-max regret action. Since this holds for all types, s is a min-max regret strategy. □

Example A.3. *Min-max regret $\not\Longrightarrow$ safety level*

Consider two players, with actions a,b, and A,B respectively. Consider $u_1(a,A) = u_1(a,B) = 0, u_1(b,A) = -1, u_1(b,B) = 100$. The max regret of a for player 1 is 100, and the max regret of b is 1, and so b is the min-max regret strategy, while a is the unique safety level strategy.

B DSL Strategies: Application to the First-price Auction

B.1 The First-price Auction

As an illustrative example, we demonstrate the usage of our solution concept using the first price and discrete first price single item auctions. Interestingly, we show that in the first-price auction, there are no DSL strategies. However, moving to a discrete setting, we show that in the discrete first-price auction [5], there is a unique DSL strategy, which achieves maximal welfare and near maximal revenue.

Definition B.1. *An agent i has a value (type) v_i for an item. The agent's bid b_i (action) and nature states b_{-i} are from the same bid space. The auctioneer allocates the item to the highest bidder (either the agent or nature, tie-breaking towards nature) and if the agent wins it receives $v_i - b_i$, and otherwise 0.*

First-price auction (FPA): bid space is $0 \leq b_i \leq v_i$.

Discrete first-price auction (DFPA): bid space is $b_i \in \{\varepsilon \cdot k | \varepsilon \cdot k \leq v_i\}_{k \in \mathcal{N}}$.

Note that we reformulate the auctions to suits our agent perspective formulation. Moreover, we omit strategies that are not individually rational (in the (discrete) first-price auction, overbidding has negative utility in some nature states), which is justified by our later discussion in Proposition 3.4. We also ignore multitude in nature states that does not change the auction outcome. I.e., we only consider the highest bids by others as the nature state, and not the entire bid vector. For the DFPA, we denote $\varepsilon_{net}(v_i) = \varepsilon \cdot \max_{\varepsilon n \leq v_i} n$, i.e., the closest possible bid below the agent's value of the item.

In the first-price auction, the notion of DSL strategies is not of much help:

Lemma B.2. *In the first-price auction, there are no DSL bid strategies.*

Proof. First, consider some bid $0 \leq b_i < v_i$. Compare it with another bid b'_i that satisfies $b_i < b'_i < v_i$.[5] Consider a nature state b_{-i} so that $b_i < b_{-i} < b'_i$. Then, $0 = u_i(b_i, b_{-i}|v_i) \neq u_i(b'_i, b_{-i}|v_i) = v_i - b'_i$. Thus,

$$\min_{b_{-i} \in D_{v_i}(b_i, b'_i)} u_i(b_i, b_{-i}|v_i) = 0.$$

On the other hand, for the bid b'_i and for some nature state b_{-i}, $u_i(b'_i, b_{-i}) = 0$ if and only if $b_{-i} \geq b'_i$. In all such cases, it also holds that $u_i(b_i, b_{-i}|v_i) = 0$. In all other cases, i.e., when $b_{-i} < b'_i$, the utility of the bidder satisfies $u_i(b'_i, b_{-i}|v_i) = v_i - b'_i$. We conclude that

$$\begin{aligned} &\min_{b_{-i} \in D_{v_i}(b_i, b'_i)} u_i(b'_i, b_{-i}|v_i) = v_i - b'_i \\ > &\min_{b_{-i} \in D_{v_i}(b_i, b'_i)} u_i(b_i, b_{-i}|v_i) = 0, \end{aligned}$$

and the bid strategy b_i is not DSL.

If $b_i = v_i$, then for any nature state b_{-i}, $u_i(b_i, b_{-i}|v_i) = 0$. For some $0 \leq b'_i < b_i$, for any nature state b_{-i} where its utility is non-zero, we have $u_i(b'_i, b_{-i}|v_i) = v_i - b'_i > 0$, and so similarly to before $b_i = v_i$ is not DSL. □

However, things get more interesting with the DFPA:

Lemma B.3. *In the discrete first-price auction:*

For types that have $\varepsilon_{net}(v_i) \neq v_i$, and types with $\varepsilon_{net}(v_i) = v_i = 0$, bidding $\varepsilon_{net}(v_i)$ is the unique DSL bid.

For types with $\varepsilon_{net}(v_i) = v_i \neq 0$, the unique DSL bid is $\varepsilon_{net}(v_i) - \varepsilon$.

We first give a proof for the pure DSL case.

Proof. The argument why any other bid strategy is not DSL follows a discretized version of the proof for Lemma B.2.

Case 1: $\varepsilon_{net}(v_i) \neq v_i$

Consider some bid with $0 \leq b'_i < \varepsilon_{net}(v_i)$. By the same argument as in the first part of the proof of Lemma B.2, bidding $\varepsilon_{net}(v_i)$ is DSL w.r.t. b'_i. Since there are no bids with $\varepsilon_{net}(v_i) < b'_i \leq v_i$ by the definition of ε_{net}, we conclude that $\varepsilon_{net}(v_i)$ is DSL w.r.t. all other bids, i.e., DSL.

Case 2: $\varepsilon_{net}(v_i) = v_i = 0$

The unique safety level bid is to bid 0, and so by Proposition 3.4 it is also the unique DSL strategy.

Case 3: $\varepsilon_{net}(v_i) = v_i \neq 0$

Similar to the first case, with the difference that bidding v_i always leads to utility 0, and so the DSL bid bracket is $v_i - \varepsilon$. □

[5] Note that the proof is written for the pure DSL case. However, it immediately generalizes to the mixed case, by adapting "$b_i < v_i$" to "has a positive probability to satisfy $b_i < v_i$", etc.

The following lemma completes the mixed DSL case:

Lemma B.4. *In the discrete first-price auction with mixed strategies, following the unique DSL pure strategy is the unique DSL strategy.*

Proof. We show the proof for case 1 where $\varepsilon_{net}(v_i) \neq v_i$. The other cases are done similarly.

Let s_i be the stated strategy, v_i the valuation (type) and the bid $b = s_i(\theta_i)$. Let b' be some other bid: since $b \neq b'$, the bracket $\varepsilon_{net}(v_i)$ has probability $p < 1$ of being the actualized bid. Consider the case b_{-i} where another bidder bids $\varepsilon_{net}(v) - \varepsilon$, and ties are broken in favor of the other bidder. Then, $u_i(b', b_{-i}|v_i) = \mathbb{E}_{\tilde{b}' \sim b'}[u_i(\tilde{b}', b_{-i}|v_i)] = p \cdot (v_i - \varepsilon_{net}(v_i)) + (1-p)\mathbb{1}[\tilde{b}' > \varepsilon_{net}(v_i)] \cdot (v_i - \tilde{b}') = p \cdot (v_i - \varepsilon_{net}(v_i)) + (1-p)\mathbb{1}[\tilde{b}' > v_i](v_i - \tilde{b}') < p \cdot (v_i - \varepsilon_{net}(v_i)) < v_i - \varepsilon_{net}(v_i)$. In any nature state and actualized outcome over the mixed bid b', if b does not win the item, then b' does not win the item, or, alternatively, it wins it and receives negative utility. So, $\min_{b_{-i} \in D_{v_i}(b,b')} u_i(b, b_{-i}|v_i) \geq v_i - \varepsilon_{net}(v_i) > \min_{b_{-i} \in D_{v_i}(b,b')} u(b', b_{-i}|v_i)$, and so by the DSL condition b' is not DSL (and b is DSL w.r.t. b').

□

The simple intuition as to why the discrete first-price auction "works" (to guarantee a DSL strategy) and the first-price auction does not, is that in the first-price auction there is always a "safer" bid that would guarantee winning the item in more nature states. In the discrete first-price auction, due to bracketing, the highest bracket that can have positive utility is that DSL bid. Note that this is "almost" truthful: When $\varepsilon_{net}(v_i) \neq v_i$, it is the closest bracket to v_i, and it is less than ε away from it. When $\varepsilon_{net}(v_i) = v_i$ (which should be seen as a rare case, where the value precisely matches the epsilon net), it is not the truthful bracket, but it is ε close to it. It is also very close to optimal revenue for the auctioneer: If n individually rational agents participate, the most the auctioneer can get is $\max_{1 \leq i \leq n} v_i$. If they play DSL strategies, she will get at least $\max_{1 \leq i \leq n} v_i - \varepsilon$.

We note that for the discrete first-price auction, DSL identifies with multi-leximin.

Corollary B.5. *The unique DSL strategy of the discrete first-price auction is also the unique multi-leximin strategy.*

Proof. For an agent i with value v_i there is a finite amount of safety level strategies, namely all the strategies with $b_i \leq v_i$, the amount of which is at most $\lceil \frac{v_i}{\varepsilon} \rceil + 1$. By Lemma 3.9, there must exist a multi-leximin strategy. By Lemma 3.8 it is also DSL. Since there is a unique DSL strategy by Lemma B.3, it must also be the unique multi-leximin. □

On the other hand, we now see that min-max regret yields a different solution to the discrete first-price auction than DSL, i.e., the two notions do not imply each other. [20] previously applied min-max regret in auction settings, and in particular discussed the DFPA in their Claim 3.1, which we restate adapted to our notations:

Claim B.6. *In the discrete first-price auction, the min-max regret strategy is to bid $\varepsilon_{net}(\frac{v_i}{2})$.*

Proof. For any bid b_i, the maximum regret is either b_i itself (in the case when no other bidders show up and it was possible to bid and pay 0), or $v_i - (b_i + \varepsilon)$ (in the case when another bidder bids b_i and the item goes to her.[6] We are thus looking for $\arg\min_{b_i} \max\{b_i, v_i - b_i - \varepsilon\}$, among the ε_{net} feasible bids.

[6] This is true under worst-case arbitrary tie-breaking. If tie-breaking is uniformly random between bidders of the same bracket, this is still true as the limiting regret when there are $n \to \infty$ bidders in the same bracket

For $b_i' < \varepsilon_{net}(\frac{v_i}{2})$, the regret is thus at least

$$\begin{aligned}
&Reg(b_i') \geq v_i - b_i - \varepsilon \\
&\geq v_i - (\varepsilon_{net}(\frac{v_i}{2}) - \varepsilon) - \varepsilon = v_i - \varepsilon_{net}(\frac{v_i}{2}) \\
&\geq \max\{v_i - \varepsilon_{net}(\frac{v_i}{2}) - \varepsilon, \frac{v_i}{2}\} \\
&\geq \max\{v_i - \varepsilon_{net}(\frac{v_i}{2}) - \varepsilon, \varepsilon_{net}(\frac{v_i}{2})\} = Reg(b_i).
\end{aligned}$$

For $b_i' > \varepsilon_{net}(\frac{v_i}{2})$, the regret is at least

$$\begin{aligned}
&Reg(b_i') \geq b_i' \geq \varepsilon_{net}(\frac{v_i}{2}) + \varepsilon \\
&\geq \max\{\varepsilon_{net}(\frac{v_i}{2}), \frac{v_i}{2}\} \\
&\geq \max\{\varepsilon_{net}(\frac{v_i}{2}), v_i - \varepsilon_{net}(\frac{v_i}{2}) - \varepsilon\} = Reg(b_i).
\end{aligned}$$

We conclude that $\varepsilon_{net}(\frac{v_i}{2})$ is the min-max regret bid strategy. □

Finally, we use the discrete first-price auction to demonstrate the difference between leximin and DSL strategies.

Example B.7. *We demonstrate that leximin is different from DSL using the discrete first-price auction. Bidding* 0 *is the leximin action, as its set of outcomes is simply the set of two items* $U_0 = \{0, v_i\}$*: This is the leximin since any other bid* $b_i > 0$ *has* $U_{b_i} = \{0, v_i - b_i\}$.

Satisfiability of Arbitrary Public Announcement Logic with Common Knowledge is Σ^1_1-hard

Rustam Galimullin
University of Bergen
Bergen, Norway
rustam.galimullin@uib.no

Louwe B. Kuijer
University of Liverpool
Liverpool, UK
lbkuijer@liverpool.ac.uk

Arbitrary Public Announcement Logic with Common Knowledge (APALC) is an extension of Public Announcement Logic with common knowledge modality and quantifiers over announcements. We show that the satisfiability problem of APALC on $S5$-models, as well as that of two other related logics with quantification and common knowledge, is Σ^1_1-hard. This implies that neither the validities nor the satisfiable formulas of APALC are recursively enumerable. Which, in turn, implies that APALC is not finitely axiomatisable.

1 Introduction

Quantified Public Announcement Logics. *Epistemic logic* (EL) [21] is one of the better-known formalisms for reasoning about knowledge of agents in multi-agent systems. It extends the language of propositional logic with constructs $\Box_a \varphi$ meaning that 'agent a knows φ'. Formulas of EL are interpreted on epistemic models (or, equivalently, $S5$-models) that comprise a set of states, equivalence relations for each agent between states, and a valuation function that specifies in which states propositional variables are true. However, EL provides only a static description of distribution of knowledge in a system. Extensions of the logic that allow one to reason about how information of individual agents and groups thereof changes as a result of some epistemic event are generally collectively known as *dynamic epistemic logics* (DELs) [10].

The prime example of a DEL and arguably the most well-studied logic in the family is *public announcement logic* (PAL) [25]. A public announcement is an event of all agents publicly and simultaneously receiving the same piece of information. The language of PAL extends that of EL with formulas $[\psi]\varphi$ that are read as 'after public announcement of ψ, φ is true'.

Quantification over various epistemic actions, and in particular over public announcements, has been explored in the last 15 or so years [9]. Adding quantification over public announcements allows one to shift the emphasis from the effects of a particular announcement to the question of (non-)existence of an announcement leading to a desired epistemic goal. In this paper, we focus on the three, perhaps most well-known, *quantified PALs* (QPALs). The first of the three is *arbitrary PAL* (APAL) [6] that extends the language of PAL with constructs $[!]\varphi$ meaning 'after *any* public announcement, φ is true'. A formula with the dual existential quantifier $\langle ! \rangle \varphi$ is read as '*there is* a public announcement, after which φ is true'.

Observe that quantifiers of APAL do not specify whether an announcement can be made by any of the agents, or groups thereof, modelled in a system. Hence, a more 'agent-centric' quantified PAL was proposed. *Group announcement logic* (GAL) [1] extends the language of PAL with formulas $[G]\varphi$ meaning 'after *any* announcement by agents from group G, φ is true'. A formula with the dual of the universal GAL quantifier is $\langle G \rangle \varphi$ that is read '*there is* an announcement by agents from group G that makes φ true'.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 260–271, doi:10.4204/EPTCS.379.21

Once we start reasoning about what groups of agents can achieve by making public announcements, it is only too natural to consider their abilities in a game-theoretic setting. In particular, we may let agents outside of the group make their own announcements in an attempt to preclude the group from reaching their epistemic goals. A QPAL with such a competitive flavour to it is called *coalition announcement logic* (CAL) [2, 15]. The logic extends PAL with modalities $[\langle G\rangle]\varphi$ that are read as '*whatever* agents from coalition G announce, *there is* a counter-announcement by the anti-coalition that makes φ true'. The diamond version $\langle[G]\rangle\varphi$ is then means that '*there is* an announcement by coalition G, such that *whatever* the anti-coalition announces at the same time, they cannot avoid φ'. Observe, that compared to APAL and GAL, modalities of CAL contain double quantification: $\forall\exists$ and $\exists\forall$ correspondingly. As the name of the logic suggests, modalities of CAL were inspired by coalition logic [24], and they capture game-theoretic notions of α- and β-effectivity [5].

Some Logical Properties of QPALs. One of the most pressing open problems in the area is the existence of finitary axiomatisations of QPALs. Both finitary and infinitary axiom systems for APAL were proposed in [6], but later the finitary version was shown to be unsound [19]. The infinitary axiomatisation is, however, sound and complete [7]. As the axiomatisation of GAL [1] is quite similar to that of APAL, its finitary version is also not sound [13, Footnote 4], and its infinitary version can be shown to be sound and complete by a modification of the proof from [7]. To the best of our knowledge, there are no known sound and complete proof systems, finitary or infinitary, for CAL[1].

The satisfiability problem for QPALs is known to be undecidable [3]. The result is achieved by a reduction from the classic tiling problem that consists in answering the question whether a given finite set of tiles can tile the $\mathbb{N}\times\mathbb{N}$ plane. Since this problem is co-RE-complete [8, 17], or, equivalently, Π^0_1-complete, the reduction amounts to the fact that the satisfiability problem for QPALs is co-RE-hard (or Π^0_1-hard). Note that this result does not rule out the existence of finitary axiomatisations of QPALs. A prime example of a logic with a co-RE-complete satisfiability problem and a finitary axiomatisation is first-order logic.

Overview of the paper and our result. In this paper we consider extensions of QPALs with *common knowledge* [12], which is a classic variant of group knowledge in multi-agent systems. Its intuitive meaning is that 'φ is common knowledge among agents in group G if everyone in G knows φ, everyone in G knows that everyone in G knows φ and so on ad infinitum'. Semantically, common knowledge among agents from G corresponds to the reflexive transitive closure of equivalence relations of all agents from group G. We call extensions of APAL, GAL, and CAL with common knowledge APALC [4], GALC, and CALC, correspondingly, or QPALCs if we refer to all of them at the same time.

The result we prove in this paper is that the satisfiability problems for QPALCs are Σ^1_1-hard. We do this by showing that the *recurring tiling problem*, which is known to be Σ^1_1-complete [18], can be reduced to satisfiability of QPALC formulas. Because the satisfiability problems are Σ^1_1-hard, it follows that, in particular, the set of valid QPALC formulas is not recursively enumerable. That, in turn, implies that QPALCs have no finitary axiomatisations. The non-existence of a finitary axiomatisation of a somewhat related arbitrary arrow update logic [11] with common knowledge was shown in [20] by the reduction from the non-halting problem. Moreover, the recurring tiling problem was used in [22] to demonstrate that the satisfiability problem of PAL with iterated announcements and common knowledge is Σ^1_1-complete.

The use of common knowledge is instrumental in our paper, since it allows us to have a 'tighter' grid than the ones from [3] and [14]. We deem our result important in at least two ways. First, the non-existence of finitary axiomatisations of QPALCs is interesting in its own right as it demonstrates

[1] A complete infinitary axiomatisation with CAL modalities and additional operators was given in [16]

that presence of common knowledge in QPALCs is a sufficient condition for Σ_1^1-hardness. Second, having both our construction (with common knowledge) and the constructions from [3] and [14] side by side, allows one to flesh out crucial differences between Σ_1^1-hardness and Σ_1^0-hardness arguments, and, hopefully, move closer to tackling the open problem of (non-)existence of finitary axiomatisations of QPALs.

Outline of the paper. The rest of the paper is organised as follows. In Section 2 we cover the background on QPALCs. After that, in Section 3, we prove the main claim of this paper, and, finally, we conclude in Section 4.

2 Quantified Public Announcement Logics with Common Knowledge

Let A be a finite set of agents, and P be a countable set of propositional variables.

Definition 2.1. The *languages of arbitrary public announcement logic with common knowledge* APALC, *group announcement logic with common knowledge* GALC, and *coalition announcement logic with common knowledge* CALC are inductively defined as

$$\begin{array}{lll}\mathsf{APALC} & \ni & \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a\varphi \mid [\varphi]\varphi \mid \blacksquare_G\varphi \mid [!]\varphi \\ \mathsf{GALC} & \ni & \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a\varphi \mid [\varphi]\varphi \mid \blacksquare_G\varphi \mid [G]\varphi \\ \mathsf{CALC} & \ni & \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a\varphi \mid [\varphi]\varphi \mid \blacksquare_G\varphi \mid [\langle G\rangle]\varphi\end{array}$$

where $p \in P$, $a \in A$, and $G \subseteq A$. Duals are defined as $\Diamond_a\varphi := \neg\Box_a\neg\varphi$, $\langle\psi\rangle\varphi := \neg[\psi]\neg\varphi$, $\blacklozenge_G\varphi := \neg\blacksquare_G\neg\varphi$, $\langle!\rangle\varphi := \neg[!]\neg\varphi$, $\langle G\rangle\varphi := \neg[G]\neg\varphi$ and $\langle[G]\rangle\varphi := \neg[\langle G\rangle]\neg\varphi$.

The fragment of APALC without $[!]\varphi$ is called *public announcement logic with common knowledge* PALC; the latter without $[\varphi]\varphi$ is *epistemic logic with common knowledge* ELC; PALC and ELC minus $\blacksquare_G\varphi$ are, correspondingly, *public announcement logic* PAL and *epistemic logic* EL. Finally, fragments of APALC, GALC and CALC without $\blacksquare_G\varphi$ are called *arbitrary public announcement logic* APAL, *group announcement logic* GAL and *coalition announcement logic* CAL respectively.

Definition 2.2. A *model* M is a tuple $(S, \sim, V)$, where S is a non-empty set of states, $\sim : A \to 2^{S\times S}$ gives an equivalence relation for each agent, and $V : P \to 2^S$ is the valuation function. By $\sim_G$ we mean reflexive transitive closure of $\bigcup_{a\in G} \sim_a$. We will denote model M with a distinguished state s as M_s.

We would like to stress that agent relations in our models are *equivalence relations* (and hence our models are $S5$ models). The results of this paper do not generalise to arbitrary agent relations in any obvious way.

It is assumed that for group announcements, agents know the formulas they announce. In the following, we write $\mathsf{PALC}^G = \{\bigwedge_{i\in G}\Box_i\psi_i \mid \text{for all } i \in G, \psi_i \in \mathsf{PALC}\}$ to denote the set of all possible announcements by agents from group G. We will use ψ_G to denote arbitrary elements of PALC^G.

Definition 2.3. Let $M_s = (S, R, V)$ be a model, $p \in P$, $G \subseteq A$, and $\varphi, \psi \in \mathsf{APALC} \cup \mathsf{GALC} \cup \mathsf{CALC}$.

$$\begin{array}{lll}
M_s \models p & \text{iff} & s \in V(p) \\
M_s \models \neg\varphi & \text{iff} & M_s \not\models \varphi \\
M_s \models \varphi \wedge \psi & \text{iff} & M_s \models \varphi \text{ and } M_s \models \psi \\
M_s \models \Box_a \varphi & \text{iff} & \forall t \in S : s \sim_a t \text{ implies } M_t \models \varphi \\
M_s \models \blacksquare_G \varphi & \text{iff} & \forall t \in S : s \sim_G t \text{ implies } M_t \models \varphi \\
M_s \models [\psi]\varphi & \text{iff} & M_s \models \psi \text{ implies } M_s^\psi \models \varphi \\
M_s \models [!]\varphi & \text{iff} & \forall \psi \in \mathsf{PALC} : M_s \models [\psi]\varphi \\
M_s \models [G]\varphi & \text{iff} & \forall \psi_G \in \mathsf{PALC}^G : M_s \models [\psi_G]\varphi \\
M_s \models [\langle G \rangle]\varphi & \text{iff} & \forall \psi_G \in \mathsf{PALC}^G, \exists \chi_{A \setminus G} \in \mathsf{PALC}^{A \setminus G} : M_s \models \psi_G \text{ implies } M_s \models \langle \psi_G \wedge \chi_{A \setminus G} \rangle \varphi
\end{array}$$

where $M_s^\psi = (S^\psi, R^\psi, V^\psi)$ with $S^\psi = \{s \in S \mid M_s \models \psi\}$, $R^\psi(a)$ is the restriction of $R(a)$ to S^ψ for all $a \in A$, and $V^\psi(p) = V(p) \cap S^\psi$ for all $p \in P$.

Observe, that it follows from the definition of the semantics that in the case of the grand coalition A, $M_s \models [A]\varphi$ if and only if $M_s \models [\langle A \rangle]\varphi$. For the case of the empty group $\varnothing$, we assume that the conjunction of an empty set of formulas is a tautology.

Remark 1. For APAL, GAL, and CAL, we assume that quantification ranges over a quantifier-free fragment of the language, i.e. over PAL, which is equally expressive as EL [25]. This is, however, not as straightforward once we consider ELC and PALC. The latter is strictly more expressive than ELC [10, Theorem 8.48], and ELC, in its turn, is strictly more expressive than EL, and thus it matters, expressivity-wise, which quantifer-free fragment of a QPALC the quantification ranges over. These matters are explored in [4], where also infinitary axiomatisations of APALC and GALC are given. For our current purposes, though, the difference in the range of quantification does not play a role.

3 The Satisfiability Problem of QPALCs is Σ_1^1-hard

We prove the Σ_1^1-hardness of the satisfiability problem of QPALCs via a reduction from the recurring tiling problem [17].

Definition 3.1. Let C be a finite set of *colours*. A *tile* is a function $\tau : \{\mathsf{north}, \mathsf{south}, \mathsf{east}, \mathsf{west}\} \to C$. A finite set of tiles T is called an *instance* of the tiling problem. A *solution* to an instance of the tiling problem is a function[2] $f : \mathbb{N} \times \mathbb{N} \to \mathrm{T}$ such that for all $(i, j) \in \mathbb{N} \times \mathbb{N}$,

$$f(i, j)(\mathsf{north}) = f(i, j+1)(\mathsf{south}) \text{ and } f(i, j)(\mathsf{east}) = f(i+1, j)(\mathsf{west}).$$

Definition 3.2. Let T be a finite set of tiles with a designated tile $\tau^* \in \mathrm{T}$. The *recurring tiling problem* is the problem to determine whether there is a solution to instance T of the tiling problem such that τ^* appears *infinitely* often in the first column.

We assume without loss of generality that the designated tile τ^* occurs only in the first column.

[2] Throughout the paper we assume that $0 \in \mathbb{N}$.

3.1 Encoding a Tiling

For our construction we will require five propositional variables — north, south, east, west and centre — to designate the corresponding sides of tiles. Additionally, we will have designated propositional variables for each colour in C, and for each tile $\tau_i \in \mathrm{T}$ there is a propositional variable p_i that represents this tile. Finally, we will use p^* for the special τ^*.

In our construction, we will represent each tile with (at least) five states: one for each of the four sides of a tile, and one for the centre. As for agents, we require only three of them for our construction. Agent s, for $s$quare, cannot distinguish states within the same tile. Agent v, for $v$ertical, cannot distinguish between the northern part of one tile and the southern part of the tile above. Similarly, the $h$orizontal agent h cannot distinguish between the eastern and western parts of adjacent tiles. See Figure 1 for the depiction of an intended grid-like model.

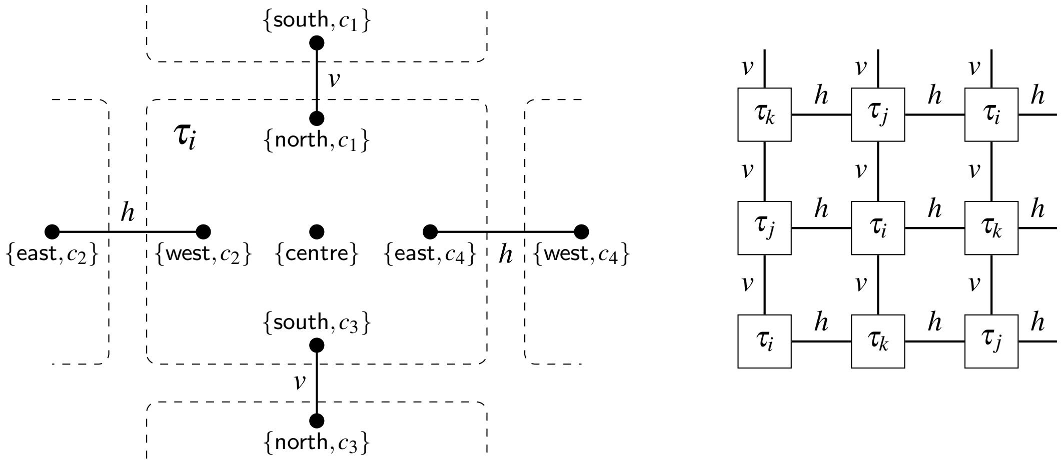


Figure 1: Left: a representation of a single tile τ_i, where agent s has the universal relation within the dashed square, relations h and v are equivalences, and reflexive arrows are omitted. Each state is labelled by a set of propositional variables that are true there. Right: an example of a grid-like model that we construct in our proof. Each tile τ has a similar structure as presented on the left of the figure.

Let an instance T of the recurring tiling problem be given. We start by construction of formula Ψ_{T} that will be satisfied in a given model if and only if the model is grid-like. We will build up Ψ_{T} step-by-step, defining useful subformulas along the way. Let Position be the following set Position := {north, south, east, west, centre}.

The first constraint, expressed by formula *one_colour*, is that each state is coloured by exactly one colour. To ensure that all five parts — north, south, east, west, and centre — are present in a current square, we state in *all_parts* that in all squares the square agent s has access to all five relevant states.

$$one\_colour := \bigvee_{c \in C} \left(c \wedge \bigwedge_{d \in C \setminus \{c\}} \neg d \right) \qquad all\_parts := \Box_s \bigvee_{q \in \mathsf{Position}} q \wedge \bigwedge_{q \in \mathsf{Position}} \Diamond_s q$$

The formulas *hor* and *vert* state that the relation h only allows us to move between east and west states, while v only allows movement between north and south states.

$$hor := \bigwedge_{q \in \{\mathsf{north},\mathsf{south},\mathsf{centre}\}} (q \rightarrow \Box_h q) \qquad vert := \bigwedge_{q \in \{\mathsf{east},\mathsf{west},\mathsf{centre}\}} (q \rightarrow \Box_v q)$$

With *one_pos* we force each state to satisfy exactly one propositional variable from Position, and with *one_tile* we ensure that all states within the same tile are labelled by the tile proposition.

$$one\_pos := \bigvee_{q \in \mathsf{Position}} \left(q \land \bigwedge_{q' \in \mathsf{Position} \setminus \{q\}} \neg q' \right) \qquad one\_tile := \bigvee_{\tau_i \in \mathrm{T}} \left(p_i \land \Box_s p_i \land \bigwedge_{\tau_j \in \mathrm{T} \setminus \{\tau_i\}} \neg p_j \right)$$

Next, we force each state in a tile to satisfy exactly one atom corresponding to their designated colour:

$$state\_col := \bigvee_{\tau_i \in \mathrm{T}} \left(p_i \to \bigwedge_{q \in \mathsf{Position} \setminus \{\mathsf{centre}\}} (q \to \tau_i(q)) \right),$$

where $\tau_i(q)$ is the colour of the tile τ_i on the side q (e.g. $\tau_i(\mathsf{south})$ is the bottom colour of tile τ_i).

All the formulas considered so far deal with the representation of a single tile. We will use the following abbreviation:

$$\psi_{tile} := one\_colour \land all\_parts \land hor \land vert \land one\_pos \land one\_tile \land state\_col$$

Adjoining tiles are required to have the same colour on the sides facing each other, we simulate this by requiring that agents h and v consider a current colour in the top and right directions. In such a way we also ensure that the grid is infinite in the positive quadrant.

$$adj\_tiles := \bigwedge_{c \in C} ((\mathsf{north} \land c \to \Diamond_v \mathsf{south} \land \Box_v c) \land (\mathsf{east} \land c \to \Diamond_h \mathsf{west} \land \Box_h c))$$

We are concerned with the reduction from the $\mathbb{N} \times \mathbb{N}$ recurring tiling problem, i.e. our grid will have left and bottom edges. We force the existence of a tile at position $(0,0)$ with the following formula:

$$init := \blacklozenge_{\{h,v,s\}} (\blacksquare_{\{v,s\}} (\mathsf{west} \to \Box_h \mathsf{west}) \land \blacksquare_{\{h,s\}} (\mathsf{south} \to \Box_v \mathsf{south}))$$

For the remaining formulas, it is useful to define two abbreviations. We use $\Box_{up}\varphi$ to denote $\Box_s(\mathsf{north} \to \Box_v(\mathsf{south} \to \varphi))$, i.e., we first move, by agent s, to the state representing the northern quadrant of the tile, then we move, by agent v, to southern quadrant of the tile above, where we evaluate φ. Similarly, we use $\Box_{right}\varphi$ to denote $\Box_s(\mathsf{east} \to \Box_h(\mathsf{west} \to \varphi))$. The duals $\Diamond_{up}$ and $\Diamond_{right}$ are defined as usual.

The next two formulas are used to guarantee that for every tile there are unique tiles, up to PALC-indistinguishability, above it and to its right.

$$up := [!](\Diamond_{up}\Diamond_s \mathsf{centre} \to \Box_{up}\Diamond_s \mathsf{centre})$$
$$right := [!](\Diamond_{right}\Diamond_s \mathsf{centre} \to \Box_{right}\Diamond_s \mathsf{centre})$$

Additionally, we use the following two formulas to establish a commutative property: going *up* and then *right* results in a state that is PALC-indistinguishable from going *right* and then *up*.

$$right\&up := [!](\Diamond_{right}\Diamond_{up}\Diamond_s \mathsf{centre} \to \Box_{up}\Box_{right}\Diamond_s \mathsf{centre})$$
$$up\&right := [!](\Diamond_{up}\Diamond_{right}\Diamond_s \mathsf{centre} \to \Box_{right}\Box_{up}\Diamond_s \mathsf{centre})$$

Finally, we make sure that any two states that are h or v related and that are in the same position are parts of indistinguishable tiles.

$$no\_change := \bigwedge_{q,q' \in \mathsf{Position}} [!]((q \wedge \Diamond_s q') \to (\Box_h (q \to \Diamond_s q') \wedge \Box_v (q \to \Diamond_s q')))$$

The formula *hor* states that unless we are in a east or west position, we cannot go to a different position using h. Similarly, *vert* states that unless we are in a north or south position we can't use v to change position. The formula *no_change* then states that any move by relation h or v that does not change the position must lead to an indistinguishable tile.

We abbreviate formulas with quantifiers as

$$\psi_{x\&y} := up \wedge right \wedge right\&up \wedge up\&right \wedge no\_change$$

In our reduction, we are interested in grids where a special tile appears infinitely often in the first column of the grid. The following formula requires that the special tile appears only in the leftmost column:

$$tile\_left := p^* \to \Box_s(\mathsf{west} \to \Box_h \mathsf{west})$$

All of this completes the necessary requirements for the grid. Now, by adding a common knowledge modality for all agents, we force all of the aforementioned formulas to hold everywhere in the grid.

$$\Psi_{\mathrm{T}} := \blacksquare_{\{h,v,s\}} \left(\psi_{tile} \wedge adj\_tiles \wedge init \wedge \psi_{x\&y} \wedge tile\_left\right)$$

Observe that Ψ_{T} does not say anything about the special tile τ^* appearing infinitely often in the first column. The formula merely requires that if there is a special tile, then it should appear in the first column. We first show that Ψ_{T} forces a grid-like model, and only after that will we consider the (in)finite number of occurrences of the special tile.

Lemma 1. Let T be an instance of the recurring tiling problem. If T can tile $\mathbb{N} \times \mathbb{N}$, then Ψ_{T} is satisfiable.

Proof. Assume that there is a tiling of the $\mathbb{N} \times \mathbb{N}$ plane with a finite set of tiles T. We construct model $M = (S, \sim, V)$ satisfying Ψ_{T} directly from the given tiling. In particular,

- $S = \mathbb{N} \times \mathbb{N} \times \{\mathfrak{n}, \mathfrak{s}, \mathfrak{e}, \mathfrak{w}, \mathfrak{c}\}$,
- $\sim_s = \{(i, j, \mathfrak{l}), (i', j', \mathfrak{l}') \mid i = i' \text{ and } j = j'\}$
- $\sim_v$ is the reflexive closure of $\{(i, j, \mathfrak{n}), (i, j+1, \mathfrak{s})\}$
- $\sim_h$ is the reflexive closure of $\{(i, j, \mathfrak{e}), (i+1, j, \mathfrak{w})\}$
- for all $\tau_k \in \mathrm{T}$, $V(p_k) = \{(i, j, \mathfrak{l}) \mid \tau_k \text{ is at } (i, j)\}$
- for all $c \in C$, $V(c) = \{(i, j, \mathfrak{l}) \mid \tau(\mathfrak{l}) = c\}$
- for all $l \in \mathsf{Position}$, $V(l) = \{(i, j, \mathfrak{l}) \mid l \text{ corresponds to } \mathfrak{l}\}$

To argue that $M_{(0,0,\mathfrak{c})} \models \Psi_{\mathrm{T}}$ we first notice that due to the fact that T tiles the $\mathbb{N} \times \mathbb{N}$ plane and by the construction of M, subformulas of Ψ_{T} that do not involve arbitrary announcements are straightforwardly satisfied.

Now, consider the formula *up*. For every $(i, j, \mathfrak{l})$, there is at most one $(i', j', \mathfrak{l}')$ that is reachable by taking an s-step to a north state followed by a v-step to a south state, namely $(i', j', \mathfrak{l}') = (i, j+1, \mathfrak{s})$. Furthermore, this property is retained in any submodel of M. As a consequence, in any state of any submodel of M, $\Diamond_{up}\chi$ implies $\Box_{up}\chi$, for every χ. In particular, it follows that $M_{(i,j,\mathfrak{l})} \models [!](\Diamond_{up}\Diamond_s \mathsf{centre} \to \Box_{up}\Diamond_s \mathsf{centre})$, i.e., $M_{(i,j,\mathfrak{l})} \models up$.

Similar reasoning shows that $(i, j, \mathfrak{l})$ satisfies the other conjuncts of $\psi_{x\&y}$. Hence $M_{(i,j,\mathfrak{l})} \models \psi_{tile} \wedge adj\_tiles \wedge init \wedge \psi_{x\&y} \wedge tile\_left$, for all $(i, j, \mathfrak{l})$, and thus $M_{(0,0,\mathfrak{c})} \models \Psi_{\mathrm{T}}$. □

The more complex part of the reduction is to show that if Ψ_T is satisfiable, then a tiling exists.

Lemma 2. Let T be an instance of the recurring tiling problem. If Ψ_T is satisfiable, then T can tile $\mathbb{N}\times\mathbb{N}$.

Proof. Let M be such that $M_s \models \Psi_\mathrm{T}$. The model M is partitioned by $\sim_s$, we refer to these partitions as grid points, and label these points as follows.

- The grid point containing s is labelled $(0,0)$.
- If A and B are grid points, A is labelled (i,j) and there is a north-state in A that is v-indistinguishable to a south-state in B, then B is labelled $(i,j+1)$.
- If A and B are grid points, A is labelled (i,j) and there is a east-state in A that is h-indistinguishable to a west-state in B, then B is labelled $(i+1,j)$.

Note that a single grid point might have multiple labels. We say that (i,j) is tiled with τ_i if there is some grid point labelled with (i,j) that contains a state where p_i holds. We start by noting that because the main connective of Ψ_T is $\blacksquare_{\{h,v,s\}}$, the formula holds in every labelled grid point. For every labelled grid point X and every $x\in X$, we therefore have $M_x \models \psi_{tile}$. So X contains states for every direction, each labelled with exactly one colour that corresponds to the tile that holds on X. We continue by proving the following claim.

Claim 1: Let X, A and B be grid points where X is labeled (i,j) while A and B are both labeled $(i,j+k)$ by virtue of being k-steps to the north of X. Then A and B are PALC-indistinguishable, in the sense that for every $\chi\in\mathsf{PALC}$, if there is an $a\in A$ such that $M_a\models\chi$ then there is a $b\in B$ such $M_b\models\chi$ (and vice versa).

Proof of Claim 1: By induction on k. As base case, let $k=1$ and suppose towards a contradiction that, for some $\chi\in\mathsf{PALC}$ and $a\in A$, $M_a\models\chi$ while for every $b\in B$, $M_b\not\models\chi$. Consider then the formula $\mathsf{centre}\to\Diamond_s\chi$. Every centre state in A satisfies this formula, while none of the centre states in B do. Hence, for every state $x\in X$, $M_x\models[\mathsf{centre}\to\Diamond_s\chi](\Diamond_{up}\Diamond_s\mathsf{centre}\wedge\neg\Box_{up}\Diamond_s\mathsf{centre})$. But that contradicts $M_x\models up$. From this contradiction, we prove the base case $k=1$.

Now, suppose as induction hypothesis that $k>1$ and that the claim holds for all $k'<k$. Again, suppose towards a contradiction that $M_a\models\chi$ while $M_b\not\models\chi$ for all $b\in B$. Let A' and B' be grid points that lie $k-1$ steps to the north of X and one step to the south of A and B, respectively. Then for every $a'\in A'$ and $b'\in B'$, $M_{a'}\models\Diamond_{up}\Diamond_s\chi$ and $M_{b'}\models\Diamond_{up}\neg\Diamond_s\chi$. By the induction hypothesis, A' and B' are indistinguishable, so $M_{a'}\models\Diamond_{up}\Diamond_s\chi\wedge\Diamond_{up}\neg\Diamond_s\chi$. But then there are distinguishable grid points one step to the north of A', contradicting the induction hypothesis. From this contradiction, we prove the induction step and thereby the claim.

Similar reasoning shows that any two grid points A,B that are labeled $(i+k,j)$ by virtue of being k steps to the right of the same grid point X are indistinguishable. Now, we can prove the next claim.

Claim 2: Let X,A and B be grid points, where X is labelled (i,j), A is labelled $(i+1,j+1)$ by virtue of being above A' which is to the right of X, and B is labelled $(i+1,j+1)$ by virtue of being to the right of B' which is above B. Then A and B are PALC-indistinguishable.

Proof of claim 2: Suppose towards a contradiction that for some $\chi\in\mathsf{PALC}$ and $a\in A$ we have $M_a\models\chi$, while $M_b\not\models\chi$ for all $b\in B$. Then for $x\in X$ we have $M_x\models[\mathsf{centre}\to\Diamond_s\chi](\Diamond_{right}\Diamond_{up}\Diamond_s\mathsf{centre}\wedge\Diamond_{up}\Diamond_{right}\neg\Diamond_s\mathsf{centre})$, contradicting $M_x\models right\&up$.

From Claim 1 it follows that any A and B that are labelled (i,j) by virtue of being i steps to the right and then j steps up from $(0,0)$ are PALC-indistinguishable. Claim 2 then lets us commute the “up” and "right" moves. Any path to (i, j) can be obtained from the path that first goes right i steps then up j steps by a finite sequence of such commutations. Hence any grid points A and B that are labelled (i, j) are PALC-indistinguishable.

The tile formulas p_i, for every $\tau_i \in \mathrm{T}$, are PALC-formulas, so there is exactly one tile τ_i that is assigned to the grid point (i, j). Furthermore, *state_col* then guarantees that each side of a grid point has the colour corresponding to the tile, and *adj_tiles* guaranteees that the tile colours match. This shows that if Ψ_T is satisfiable, then T can tile $\mathbb{N} \times \mathbb{N}$. □

3.2 Encoding the Recurring Tile

The final formula that is satisfied in a grid model if and only if a given tiling has a tile that occurs infinitely often in the first column would be

$$\Psi_\mathrm{T} \wedge \blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_\mathrm{T}.$$

In other words, the recurring tiling problem can be reduced to the APALC-satisfiability problem, where the reduction maps the instance (T, τ^*) of the recurring tiling problem to the satisfiability of $\Psi_\mathrm{T} \wedge \blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_\mathrm{T}$.

Intuitively, the formula states that if we remove all rows with the special tile, then our model is no longer a grid. See Figure 2, where on the left we have a grid with the special grey tile τ^* appearing infinitely often in the first column (every other tile in the first column is grey). Formula $\blacksquare_{\{h,s\}}\neg p^*$ holds only in those squares of the grid that lie on rows without the special tile. Thus, announcing $\blacksquare_{\{h,s\}}\neg p^*$ removes all rows that has the grey tile (see the right part of Figure 2). Since the grey tile appears infinitely often in the original grid, we have to remove an infinite number of rows after the announcement of $\blacksquare_{\{h,s\}}\neg p^*$, thus ensuring that what is left of the original model is not a grid.

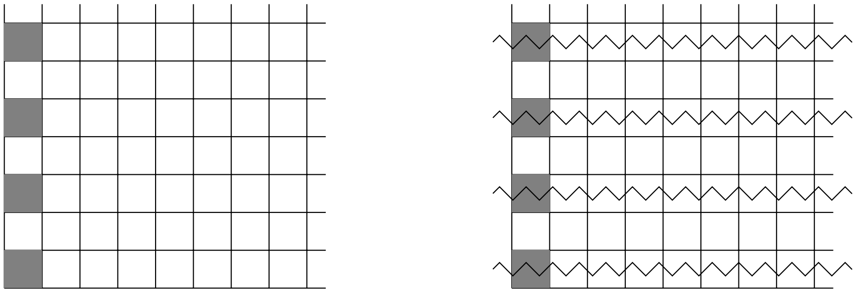


Figure 2: Left: An original grid with a special grey tile τ^* appearing infinitely often in the first column. Right: The grid after the public announcement of $\blacksquare_{\{h,s\}}\neg p^*$. Crossed-out rows are not preserved after the announcement.

Theorem 1. Let T be an instance of the tiling problem with a special tile $\tau^* \in \mathrm{T}$. Set T can tile $\mathbb{N} \times \mathbb{N}$ with τ^* appearing infinitely often in the first column if and only if $\Psi_\mathrm{T} \wedge \blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_\mathrm{T}$ is satisfiable.

Proof. First, let us can extend the labelling from the proof of Lemma 2 as follows:

- For every $q \in \mathsf{Position}$, if A and B are grid points, A is labeled (i, j) and there is a q state in A that is v or h-indistinguishable from a q state in B, then B is labeled (i, j).

It follows from *no_change* that this extended labelling retains the property that any two grid points with the same label are PALC-indistinguishable. Furthermore, from *hor* and *vert* it follows that every grid point that is reachable by h, v and s is now labelled with some coordinates (i,j). Hence we can identify the $\{h,v,s\}$-reachable grid points in any model of Ψ_{T} with $\mathbb{N}\times\mathbb{N}$.

Now, assume that set T cannot tile the $\mathbb{N}\times\mathbb{N}$ plane with a special tile $\tau' \in \mathrm{T}$ appearing infinitely often in the first column. We argue that in this case, $\Psi_{\mathrm{T}} \wedge \blacksquare_{\{v,s\}}([\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_{\mathrm{T}})$ is not satisfiable. The first conjunct is straightforward. If T cannot tile the $\mathbb{N}\times\mathbb{N}$ plane, then, by Lemma 2, Ψ_{T} is not satisfiable.

So suppose that T can tile the plane, but only in such a way that τ^* occurs finitely often. For every model $M_{(0,0,\mathfrak{l})}$ of Ψ_{T}, there is then some $k\in\mathbb{N}$ that is the last row in which p^* is true. The formula $\blacksquare_{\{h,s\}}\neg p^*$ holds exactly on those rows where p^* does not hold in the first column. As a result, the update $[\blacksquare_{\{h,s\}}\neg p^*]$ does not remove any rows past row k. The grid points $\mathbb{N}\times\mathbb{N}_{>k}$ then still form a grid that is isomorphic to $\mathbb{N}\times\mathbb{N}$, and that is tiled. See Figure 3 for a depiction of the situation.

It follows that $M_{(0,k,\mathfrak{l})} \not\models [\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_{\mathrm{T}}$, and therefore $M_{(0,0,\mathfrak{l})} \not\models \blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_{\mathrm{T}}$. This is true for every model of Ψ_{T}, so $\Psi_{\mathrm{T}}\wedge\blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_{\mathrm{T}}$ is not satisfiable.

If, on the other hand, T can tile the plane in such a way that τ^* occurs infinitely often in the first column, then there is a model of Ψ_{T} where the modality $[\blacksquare_{\{h,s\}}\neg p^*]$ removes infinitely many rows, and therefore does not leave any infinite grid. So $\Psi_{\mathrm{T}}\wedge\blacksquare_{\{v,s\}}[\blacksquare_{\{h,s\}}\neg p^*]\neg\Psi_{\mathrm{T}}$ is satisfiable. □

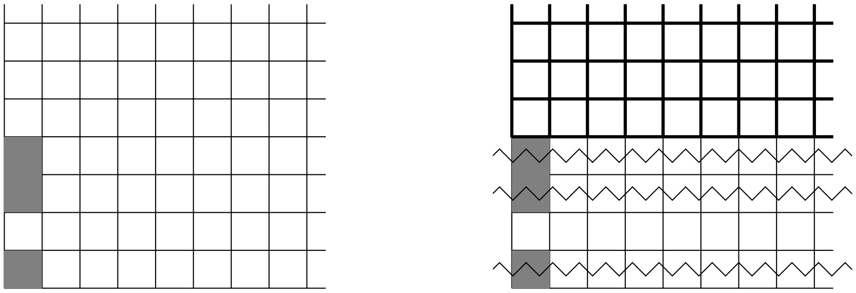


Figure 3: Left: An original grid with a special grey tile τ^* appearing finitely often in the first column. Right: The grid after the public announcement of $\blacksquare_{\{h,s\}}\neg p^*$. Crossed-out rows are not preserved after the announcement. A full $\mathbb{N}\times\mathbb{N}$ grid that is still available after the announcement is depicted with thick lines.

In the construction of Ψ_{T} and proofs of Lemmas 1 and 2, we used APALC quantifiers $[!]$. We can prove the similar results for GALC and CALC quantifers by substituting $[!]$ with $[\{h,v,s\}]$ and $[\langle\{h,v,s\}\rangle]$ correspondingly, and substituting PALC with $\mathsf{PALC}^{\{h,v,s\}}$. We get the hardness result from the Σ^1_1-completeness of the recurring tiling problem [18].

Theorem 2. The satisfiability problem of QPALCs is Σ^1_1-hard.

The Σ^1_1-hardness of the satisfiability problems of QPALCs together with the fact that the class of Σ^1_1 problems is strictly greater than the class of co-RE problems [23, Chapter 4] imply that the sets of validites of the logics are not RE, which, in turn, implies that QPALCs are not finitely axiomatisable.

Corollary 1. The set of valid formulas of QPALCs is neither RE nor co-RE.

Corollary 2. QPALCs do not have finitary axiomatisations.

4 Discussion

The existence of finitary axiomatisations of any of APAL, GAL, and CAL is a long-standing open problem. In this paper, we have showed that the satisfiability problem of the logics extended with common knowledge modality is Σ^1_1-hard, and thus they do not admit of finitary axiomatisations. Table 1 contains the overview of the known results, including those shown in this paper, and open questions.

	APAL	GAL	CAL	APALC	GALC	CALC
Finitary axiomatisation	?	?	?	✗ (Cor. 2)	✗ (Cor. 2)	✗ (Cor. 2)
Infinitary axiomatisation	✔[6]	✔[1]	?	✔[4]	✔[4]	?

Table 1: Overview of the known results and open problems.

It is important to point out that the use of common knowledge is instrumental in our construction. Arguments from [14, 3] did not rely on common knowledge to enforce local grid properties globally, and instead the authors used an agent with the universal relation over the set of states. This approach is good enough if one wants to demonstrate the existence of a grid-like model. However, if we also require that the grid satisfies some property, like a special tile occurring infinitely often in the first column, then the presence of the global agent makes it harder to ensure this. The problem is that such an unrestrained relation may access other grids within the same model, and thus we may end up in the situation when the property is satisfied by a set of grids taken together and not by any single grid.

Our construction is 'tighter' than those in [14, 3]. In particular, our *v*ertical and *h*orizontal agents can 'see' only one step ahead. This guarantees that we stay within the same grid. In order to force grid properties globally, we use common knowledge operators that allow us to traverse a given grid-like model in all directions. It is not yet clear how to have a 'tight' grid and still be able to traverse the model without common knowledge. With this work, apart from showing that QPALCs are Σ^1_1-hard, we also hope to have elucidated the exact obstacle one has to overcome in order to claim the same about QPALs.

Acknowledgements

We would like to thank the three anonymous reviewers for their encouraging comments and constructive suggestions, which helped us to improve the presentation of our result.

References

[1] Thomas Ågotnes, Philippe Balbiani, Hans van Ditmarsch & Pablo Seban (2010): *Group announcement logic*. *Journal of Applied Logic* 8(1), pp. 62–81, doi:10.1016/j.jal.2008.12.002.

[2] Thomas Ågotnes & Hans van Ditmarsch (2008): *Coalitions and announcements*. In Lin Padgham, David C. Parkes, Jörg P. Müller & Simon Parsons, editors: *Proceedings of the 7th AAMAS*, IFAAMAS, pp. 673–680.

[3] Thomas Ågotnes, Hans van Ditmarsch & Tim French (2016): *The undecidability of quantified announcements*. *Studia Logica* 104(4), pp. 597–640, doi:10.1007/s11225-016-9657-0.

[4] Thomas Ågotnes & Rustam Galimullin (2023): *Quantifying over information change with common knowledge*. *Autonomous Agents and Multi-Agent Systems* 37(19), p. 40, doi:10.1007/s10458-023-09600-1.

[5] Robert J. Aumann (1961): *The core of a cooperative game without side payments*. *Transactions of the American Mathematical Society* 98(3), pp. 539–552, doi:10.2307/1993348.

[6] Philippe Balbiani, Alexandru Baltag, Hans van Ditmarsch, Andreas Herzig, Tomohiro Hoshi & Tiago de Lima (2008): *'Knowable' as 'known after an announcement'*. Review of Symbolic Logic 1(3), pp. 305–334, doi:10.1017/S1755020308080210.

[7] Philippe Balbiani & Hans van Ditmarsch (2015): *A simple proof of the completeness of APAL*. Studies in Logic 8(2), pp. 65–78.

[8] Robert L. Berger (1966): *The undecidability of the domino problem*. Memoirs of the American Mathematical Society, pp. 1–38.

[9] Hans van Ditmarsch (2023): *To be announced*. Information and Computation 292, pp. 1–42, doi:10.1016/j.ic.2023.105026.

[10] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2008): *Dynamic epistemic logic*. Synthese Library 337, Springer, doi:10.1007/978-1-4020-5839-4.

[11] Hans van Ditmarsch, Wiebe van der Hoek, Barteld Kooi & Louwe B. Kuijer (2017): *Arbitrary arrow update logic*. Artificial Intelligence 242, pp. 80–106, doi:10.1016/j.artint.2016.10.003.

[12] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Vardi (1995): *Reasoning about knowledge*. The MIT Press.

[13] Jie Fan (2016): *Removing your ignorance by announcing group ignorance: a group announcement logic for ignorance*. Studies in Logic 9(4), pp. 4–33.

[14] Tim French & Hans van Ditmarsch (2008): *Undecidability for arbitrary public announcement logic*. In Carlos Areces & Robert Goldblatt, editors: Proceedings of the 7th AiML, College Publications, pp. 23–42.

[15] Rustam Galimullin (2019): *Coalition announcements*. Ph.D. thesis, University of Nottingham, UK.

[16] Rustam Galimullin (2021): *Coalition and relativised group announcement logic*. Journal of Logic, Language and Information 30(3), pp. 451–489, doi:10.1007/s10849-020-09327-2.

[17] David Harel (1985): *Recurring dominoes: making the highly undecidable highly understandable*. In Marek Karpinski & Jan van Leeuwen, editors: Topics in the Theory of Computation, North-Holland Mathematics Studies 102, North-Holland, pp. 51–71, doi:10.1016/S0304-0208(08)73075-5.

[18] David Harel (1986): *Effective transformations on infinite trees, with applications to high undecidability, dominoes, and fairness*. Journal of the ACM 33(1), pp. 224–248, doi:10.1145/4904.4993.

[19] Louwe B. Kuijer (2015): *Unsoundness of R(□)*. Manuscript. Available at https://personal.us.es/hvd/APAL_counterexample.pdf.

[20] Louwe B. Kuijer (2017): *Arbitrary arrow update logic with common knowledge is neither RE nor co-RE*. In Jérôme Lang, editor: Proceedings of the 16th TARK, EPTCS 251, pp. 373–381, doi:10.4204/EPTCS.251.27.

[21] John-Jules Ch. Meyer & Wiebe van der Hoek (1995): *Epistemic logic for AI and computer science*. Cambridge Tracts in Theoretical Computer Science 41, CUP, doi:10.1017/CBO9780511569852.

[22] Joseph S. Miller & Lawrence S. Moss (2005): *The undecidability of iterated modal relativization*. Studia Logica 79(3), pp. 373–407, doi:10.1007/s11225-005-3612-9.

[23] Piergiorgio Odifreddi (1989): *Classical recursion theory*. Studies in Logic and the Foundations of Mathematics 142, Elsevier.

[24] Marc Pauly (2002): *A modal logic for coalitional power in games*. Journal of Logic and Computation 12(1), pp. 149–166, doi:10.1093/logcom/12.1.149.

[25] Jan Plaza (1989): *Logics of public communications*. In: Proceedings of the 4th ISMIS, Oak Ridge National Laboratory, pp. 201–216.

Maximizing Social Welfare in Score-Based Social Distance Games

Robert Ganian
TU Wien
Vienna, Austria
rganian@gmail.com

Thekla Hamm
Utrecht University
Utrecht, Netherlands
thekla.hamm@gmail.com

Dušan Knop
Czech Technical University in Prague
Prague, Czech Republic
dusan.knop@fit.cvut.cz

Sanjukta Roy
Penn State University
Pennsylvania, USA
sanjukta@psu.edu

Šimon Schierreich
Czech Technical University in Prague
Prague, Czech Republic
schiesim@fit.cvut.cz

Ondřej Suchý
Czech Technical University in Prague
Prague, Czech Republic
ondrej.suchy@fit.cvut.cz

Social distance games have been extensively studied as a coalition formation model where the utilities of agents in each coalition were captured using a utility function u that took into account distances in a given social network. In this paper, we consider a non-normalized score-based definition of social distance games where the utility function $u^{\vec{s}}$ depends on a generic scoring vector $\vec{s}$, which may be customized to match the specifics of each individual application scenario.

As our main technical contribution, we establish the tractability of computing a welfare-maximizing partitioning of the agents into coalitions on tree-like networks, for every score-based function $u^{\vec{s}}$. We provide more efficient algorithms when dealing with specific choices of $u^{\vec{s}}$ or simpler networks, and also extend all of these results to computing coalitions that are Nash stable or individually rational. We view these results as a further strong indication of the usefulness of the proposed score-based utility function: even on very simple networks, the problem of computing a welfare-maximizing partitioning into coalitions remains open for the originally considered canonical function u.

1 Introduction

Coalition formation is a central research direction within the fields of algorithmic game theory and computational social choice. While there are many different scenarios where agents aggregate into coalitions, a pervasive property of such coalitions is that the participating agents exhibit *homophily*, meaning that they prefer to be in coalitions with other agents which are similar to them. It was this observation that motivated Brânzei and Larson to introduce the notion of *social distance games* (SDG) as a basic model capturing the homophilic behavior of agents in a social network [14].

Brânzei and Larson's SDG model consisted of a graph $G = (V, E)$ representing the social network, with V being the agents and E representing direct relationships or connections between the agents. To capture the utility of an agent v in a coalition $C \subseteq V$, the model considered a single function: $u(v,C) = \frac{1}{|C|} \cdot \sum_{w \in C \setminus \{v\}} \frac{1}{d_C(v,w)}$ where $d_C(v,w)$ is the distance between v and w inside C.

Social distance games with the aforementioned utility function u have been the focus of extensive study to date, with a number of research papers specifically targeting algorithmic and complexity-theoretic aspects of forming coalitions with maximum social welfare [1, 2, 3, 28]. Very recently, Flammini et al. [21, 22] considered a generalization of u via an adaptive real-valued scoring vector which weights the contributions to an agent's utility according to the distances of other agents in the coalition, and studied the price of anarchy and stability for non-negative scoring vectors. However, research to date has not revealed any polynomially tractable fragments for the problem of computing coalition structures

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 272–286, doi:10.4204/EPTCS.379.22

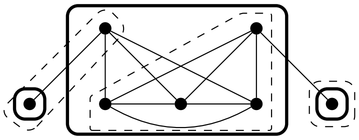


Figure 1: A social network illustrating the difference of maximising social welfare in our model compared to previous SDG models. (1) In Brânzei and Larson's SDG model, the welfare-maximum outcome is the grand coalition. (2) A welfare-maximum outcome in the normalized model of Flammini et al. with a scoring vector of $(1,0,0,0)$ is marked with dashed lines, while the same scoring vector in our non-normalized model produces the grand coalition. (3) A scoring vector of $\vec{s} = (1,0,-1)$ in our model produces the welfare-maximizing outcome marked with bold lines, with a welfare of 18. (4) A 'less welcoming' scoring vector of $\vec{s} = (1,-3)$ leads to the welfare maximizing dash-circled partition with a welfare of 14 (compared to only 12 for the bold-circled one).

with maximum social welfare (with or without stability-based restrictions on the behavior of individual agents), except for the trivial cases of complete (bipartite) graphs [14] and trees [35].

Our Contribution. The undisputable appeal of having an adaptive scoring vector—as opposed to using a single canonical utility function u—lies in the fact that it allows us to capture many different scenarios with different dynamics of coalition formation. However, it would also be useful for such a model to be able to assign negative scores to agents at certain (larger) distances in a coalition. For instance, guests at a gala event may be keen to accept the presence of friends-of-friends (i.e., agents at distance 2) at a table, while friends-of-friends may be less welcome in private user groups on social networks, and the presence of complete strangers in some scenarios may even be socially unacceptable.

Here, we propose the study of social distance games with a family of highly generic non-normalized score-based utility functions. Our aim here is twofold. First of all, these should allow us to better capture situations where agents at larger distances are unwelcome or even unacceptable for other agents. At the same time, we also want to obtain algorithms capable of computing welfare-maximizing coalition structures in such general settings, at least on well-structured networks.

Our model considers a graph G accompanied with an integer-valued, fixed but adaptive *scoring vector* $\vec{s}$ which captures how accepting agents are towards other agents based on their pairwise distance.[1] The utility function $u^{\vec{s}}(v,C)$ for an agent v in coalition C is then simply defined as $u^{\vec{s}}(v,C) = \sum_{w\in C\setminus\{v\}} \vec{s}(d_C(v,w))$; we explicitly remark that, unlike previous models, this is not normalized with respect to the coalition size. As one possible example, a scoring vector of $(1,0,-1)$ could be used in scenarios where agents are welcoming towards friends, indifferent to friends-of-friends, slightly unhappy about friends-of-friends-of-friends (i.e., agents at distance 3), and unwilling to group up with agents who are at distance greater than 3 in G. A concrete example which also illustrates the differences to previous SDG models is provided in Figure 1.

While non-normalized scoring functions have not previously been considered for social distance games, we view them a natural way of modeling agent utilities; in fact, similar ideas have been successfully used in models for a variety of other phenomena including, e.g., committee voting [20], resource allocation [13, 12] and Bayesian network structure learning [24, 36]. Crucially, it is not difficult to observe that many of the properties originally established by Brânzei and Larson for SDGs also hold for our non-normalized score-based model with every choice of $\vec{s}$, such as the small-world property [14, 27] and

[1] Formal definitions are provided in the Preliminaries.

the property that adding an agent with a close (distant) connection to a coalition positively (negatively) impacts the utilities of agents [14]. In addition, the proposed model can also directly capture the notion of *enemy aversion* with symmetric preferences [4, 34] by setting $\vec{\mathrm{s}} = (1)$.

Aside from the above, a notable benefit of the proposed model lies on the complexity-theoretic side of things. Indeed, a natural question that arises in the context of SDG is whether we can compute an outcome—a partitioning of the agents into coalitions—which maximizes the social welfare (defined as the sum of the utilities of all agents in the network). This question has been studied in several contexts, and depending on the setting one may also require the resulting coalitions to be stable under *individual rationality* (meaning that agents will not remain in coalitions if they have negative utility) or *Nash stability* (meaning that agents may leave to join a different coalition if it would improve their utility). But in spite of the significant advances in algorithmic aspects of other coalition formation problems in recent years [9, 10, 16, 23], we lack any efficient algorithm capable of producing such a welfare-optimal partitioning when using the utility function u even for the simplest types of networks.

To be more precise, when viewed through the refined lens of *parameterized complexity* [17, 19] that has recently become a go-to paradigm for such complexity-theoretic analysis, no tractable fragments of the problem are known. More precisely, the problem of computing a welfare-maximizing outcome under any of the previously considered models is not even known to admit an XP algorithm when parameterized by the minimum size of a vertex cover in the social network G—implying a significant gap towards potential fixed-parameter tractability. This means that the complexity of welfare-maximization under previous models remains wide open even under the strongest non-trivializing restriction of the network.

As our main technical contribution, we show that non-normalized score-based utility functions do not suffer from this drawback and can in fact be computed efficiently under fairly mild restrictions on G. Indeed, as our first algorithmic result we obtain an XP algorithm that computes a welfare-maximizing partitioning of the agents into coalitions parameterized by the treewidth of G, and we strengthen this algorithm to also handle additional restrictions on the coalitions in terms of individual rationality or Nash stability. As with numerous treewidth-based algorithms, we achieve this result via leaf-to-root dynamic programming along a tree-decomposition. However, the records we keep during the dynamic program are highly non-trivial and require an advanced branching step to correctly pre-computed the distances in the stored records. We remark that considering networks of small treewidth is motivated not only by the fundamental nature of this structural graph measure, but also by the fact that many real-world networks exhibit bounded treewidth [33].

In the next part of our investigation, we show that when dealing with simple scoring functions or bounded-degree networks, these results can be improved to fixed-parameter algorithms for welfare-maximization (including the cases where we require the coalitions to be individually rational or Nash stable). This is achieved by combining structural insights into the behavior of such coalitions with a different dynamic programming approach. Furthermore, we also use an entirely different technique based on quadratic programming to establish the fixed-parameter tractability of all 3 problems under consideration w.r.t. the minimum size of a vertex cover in G. Finally, we conclude with some interesting generalizations and special cases of our model and provide some preliminary results in these directions.

2 Preliminaries

We use $\mathbb{N}$ to denote the set of natural numbers, i.e., positive integers, and $\mathbb{Z}$ for the set of integers. For $i \in \mathbb{N}$, we let $[i] = \{1, \ldots, i\}$ and $[i]_0 = [i] \cup \{0\}$. We assume basic familiarity with graph-theoretic terminology [18].

Social Distance Games. A *social distance game* (SDG) consists of a set $N = \{1,\dots,n\}$ of *agents*, a simple undirected graph $G = (N,E)$ over the set of agents called a *social network*, and a non-increasing *scoring vector* $\vec{\mathrm{s}} = (s_1,\dots,s_\delta)$ where a) for each $a \in [\delta]$, $s_a \in \mathbb{Z}$ and b) for each $a \in [\delta-1]$, $s_{a+1} \leq s_a$.

In some cases, it will be useful to treat $\vec{\mathrm{s}}$ as a function from $\mathbb{N}$ rather than a vector; to this end, we set $\vec{\mathrm{s}}(a) = s_a$ for each $a \leq \delta$ and $\vec{\mathrm{s}}(a) = -\infty$ when $a > \delta$. The value "$-\infty$" here represents an inadmissible outcome, and formally we set $-\infty + z = -\infty$ and $-\infty < z$ for each $z \in \mathbb{Z}$.

A *coalition* is a subset $C \subseteq N$, and an outcome is a partitioning $\Pi = (C_1,\dots,C_\ell)$ of N into coalitions; formally, $\bigcup_{i=1}^{\ell} C_i = N$, every $C_i \in \Pi$ is a coalition, and all coalitions in Π are pairwise disjoint. We use Π_i to denote the coalition the agent $i \in N$ is part of in the outcome Π. The *utility* of an agent $i \in N$ for a coalition $\Pi_i \in \Pi$ is

$$\mathrm{u}^{\vec{\mathrm{s}}}(i,\Pi_i) = \sum_{j \in \Pi_i \setminus \{i\}} \vec{\mathrm{s}}(\mathrm{dist}_{\Pi_i}(i,j)),$$

where $\mathrm{dist}_{\Pi_i}(i,j)$ is the length of a shortest path between i and j in the graph $G[\Pi_i]$, i.e., the subgraph of G induced on the agents of Π_i. We explicitly note that if Π_i is a singleton coalition then $\mathrm{u}^{\vec{\mathrm{s}}}(i,\Pi_i) = 0$. Moreover, in line with previous work [14] we set $\mathrm{dist}_{\Pi_i}(i,j) := +\infty$ if there is no i-j path in $G[\Pi_i]$, meaning that $\mathrm{u}^{\vec{\mathrm{s}}}(i,\Pi_i) = -\infty$ whenever $G[\Pi_i]$ is not connected.

For brevity, we drop the superscript from $u^{\vec{\mathrm{s}}}$ whenever the scoring vector $\vec{\mathrm{s}}$ is clear from the context. To measure the satisfaction of the agents with a given outcome, we use the well-known notation of *social welfare*, which is the total utility of all agents for an outcome Π, that is,

$$\mathrm{SW}^{\vec{\mathrm{s}}}(\Pi) = \sum_{i \in N} \mathrm{u}^{\vec{\mathrm{s}}}(i,\Pi_i).$$

Here, too, we drop the superscript specifying the scoring vector whenever it is clear from the context.

We assume that all our agents are selfish, behave strategically, and their aim is to maximize their utility. To do so, they can perform *deviations* from the current outcome Π. We say that Π admits an *IR-deviation* if there is an agent $i \in N$ such that $\mathrm{u}(i,C) < 0$; in other words, agent i prefers to be in a singleton coalition over its current coalition. If no agent admits an IR-deviation, the outcome is called *individually rational* (IR). We say that Π admits an *NS-deviation* if there is an agent i and a coalition $C \in \Pi \cup \{\emptyset\}$ such that $\mathrm{u}(i,C \cup \{i\}) > \mathrm{u}(i,\Pi_i)$. Π is called *Nash stable* (NS) if no agent admits an NS-deviation. We remark that other notions of stability exist in the literature [13, Chapter 15], but Nash stability and individual rationality are the most basic notions used for stability based on individual choice [29, 38].

Having described all the components in our score-based SDG model, we are now ready to formalize the three classes of problems considered in this paper. We note that even though these are stated as decision problems for complexity-theoretic reasons, each of our algorithms for these problems can also output a suitable outcome as a witness. For an arbitrary fixed scoring vector $\vec{\mathrm{s}}$, we define:

$\vec{\mathrm{s}}$-SDG-WF
Input: A social network $G = (N,E)$, desired welfare $b \in \mathbb{N}$.
Question: Does the distance game given by G and $\vec{\mathrm{s}}$ admit an outcome with social welfare at least b?

$\vec{\mathrm{s}}$-SDG-WF-IR and $\vec{\mathrm{s}}$-SDG-WF-NASH are then defined analogously, but with the additional condition that the outcome must be individually rational or Nash stable, respectively.

We remark that for each of the three problems, one may assume w.l.o.g. that $s_1 > 0$; otherwise the trivial outcome consisting of $|N|$ singleton coalitions is both welfare-optimal and stable. Moreover, without loss of generality we assume G to be connected since an optimal outcome for a disconnected graph G can be obtained as a union of optimal outcomes in each connected component of G.

The last remark we provide to the definition of our model is that it trivially also supports the well-known *small world* property [27] that has been extensively studied on social networks. In their original work on SDGs, Brânzei and Larson showed that their model exhibits the small world property by establishing a diameter bound of 14 in each coalition in a so-called *core partition* [14]. Here, we observe that for each choice of $\vec{s}$, a welfare-maximizing coalition will always have diameter at most δ.

Parameterized Complexity. The *parameterized complexity* framework [17, 19] provides the ideal tools for the fine-grained analysis of computational problems which are NP-hard and hence intractable from the perspective of classical complexity theory. Within this framework, we analyze the running times of algorithms not only with respect to the input size n, but also with respect to a numerical parameter $k \in \mathbb{N}$ that describes a well-defined structural property of the instance; the central question is then whether the superpolynomial component of the running time can be confined by a function of this parameter alone.

The most favorable complexity class in this respect is FPT (short for "fixed-parameter tractable") and contains all problems solvable in $f(k) \cdot n^{\mathcal{O}(1)}$ time, where f is a computable function. Algorithms with this running time are called *fixed-parameter algorithms*. A less favorable, but still positive, outcome is an algorithm with running time of the form $n^{f(k)}$; problems admitting algorithms with such running times belong to the class XP.

Structural Parameters. Let $G = (V,E)$ be a graph. A set $U \subseteq V$ is a *vertex cover* if for every edge $e \in E$ it holds that $U \cap e \neq \emptyset$. The *vertex cover number* of G, denoted $\mathrm{vc}(G)$, is the minimum size of a vertex cover of G. A *nice tree-decomposition* of G is a pair $(\mathcal{T},\beta)$, where $\mathcal{T}$ is a tree rooted at a node $r \in V(\mathcal{T})$, $\beta\colon V(\mathcal{T}) \to 2^V$ is a function assigning each node x of $\mathcal{T}$ its *bag*, and the following conditions hold:

- for every edge $\{u,v\} \in E(G)$ there is a node $x \in V(\mathcal{T})$ such that $u,v \in \beta(x)$,
- for every vertex $v \in V$, the set of nodes x with $v \in \beta(x)$ induces a connected subtree of $\mathcal{T}$,
- $|\beta(r)| = |\beta(x)| = 0$ for every *leaf* $x \in V(\mathcal{T})$, and
- there are only tree kinds of internal nodes in $\mathcal{T}$:
  - x is an *introduce node* if it has exactly one child y such that $\beta(x) = \beta(y) \cup \{v\}$ for some $v \notin \beta(y)$,
  - x is a *join node* if it has exactly two children y and z such that $\beta(x) = \beta(y) = \beta(z)$, or
  - x is a *forget node* if it has exactly one child y such that $\beta(x) = \beta(y) \setminus \{v\}$ for some $v \in \beta(y)$.

The *width* of a nice tree-decomposition $(\mathcal{T},\beta)$ is $\max_{x \in V(\mathcal{T})} |\beta(x)| - 1$, and the treewidth $\mathrm{tw}(G)$ of a graph G is the minimum width of a nice tree-decomposition of G. Given a nice tree-decomposition and a node x, we denote by G^x the subgraph induced by the set $V^x = \bigcup_{y \text{ is a descendant of } x} \beta(y)$, where we suppose that x is a descendant of itself. It is well-known that optimal nice tree-decompositions can be computed efficiently [7, 30, 31].

Integer Quadratic Programming. Integer Quadratic Programming (IQP) over d dimensions can be formalized as the task of computing

$$\max\left\{x^T Qx \mid Ax \le b, x \ge 0, x \in \mathbb{Z}^d\right\}, \tag{IQP}$$

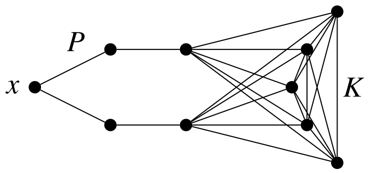


Figure 2: Social Network from Lemma 2.

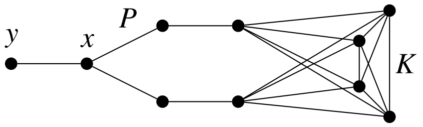


Figure 3: Social Network from Lemma 3.

where $Q \in \mathbb{Z}^{d\times d}$, $A \in \mathbb{Z}^{m\times d}$, $b \in \mathbb{Z}^m$. That is, IQP asks for an integral vector $x \in \mathbb{Z}^d$ which maximizes the value of a quadratic form subject to satisfying a set of linear constraints.

Proposition 1 ([32, 39], see also [25])**.** Integer Quadratic Programming *is fixed-parameter tractable when parameterized by* $d + \|A\|_\infty + \|Q\|_\infty$.

3 Structural Properties of Outcomes

As our first set of contributions, we establish some basic properties of our model and the associated problems that are studied within this paper. We begin by showcasing that the imposition of individual rationality or Nash stability as additional constraints on our outcomes does in fact have an impact on the maximum welfare that can be achieved (and hence it is indeed necessary to consider three distinct problems). We do not consider this to be obvious at first glance: intuitively, an agent i's own contribution to the social welfare can only improve if they perform an IR- or NS-deviation, and the fact that the distance function dist_{Π_i} is symmetric would seem to suggest that this can only increase the total social welfare.

Lemma 2. *There is a scoring vector $\vec{s}$ and a social network G such that the single outcome achieving the maximum social welfare is not individually rational.*

Proof. Consider a scoring function $\vec{s}$ such that $\vec{s} = (1,1,-1,-1,-1,-1)$. Consider the social network G in Figure 2 formed from a path P on 5 vertices and a clique K on 5 vertices by connecting the endpoints of P to all vertices of K. Let x be the central agent of P. Let C be the grand coalition in G. The graph can be viewed as a 6-cycle with K forming one "bold" agent. All vertices on the cycle contribute positively to the agent's utility, except for the one that is exactly opposite on the cycle. Hence, $\mathrm{u}(x,C) = 4-5 = -1$, while utility of all other agents is $8-1 = 7$ in C. This gives total social welfare of 62 for the grand coalition.

However, if x leaves the coalition to form its own one, their utility will improve from -1 to 0, whereas the total social welfare drops. Indeed, in $C \setminus \{x\}$ there are 2 agents with utility $6-2=4$, 2 agents with utility $7-1=6$ and 5 agents with utility $8-0$, giving total social welfare of 60. If any $y \neq x$ was to be excluded from C to form outcome $\{y\}, C\setminus\{y\}$, then y joining C improves social welfare, proving that it was not optimal. Finally, if the outcome consists of several coalitions with the largest one of size 8, then the welfare is at most $8\cdot 7+2\cdot 1 = 56$, if the largest size is 7, then we get at most $7\cdot 6+3\cdot 2 = 48$, for 6 it is $6\cdot 5+4\cdot 3 = 42$ and for 5 it is $5\cdot 4+5\cdot 4 = 40$.

Hence the grand coalition C is the only outcome with maximal social welfare, but it is not individually rational (and therefore not Nash stable), as $\mathrm{u}(x,C) = -1$. □

Lemma 3. *There is a scoring vector $\vec{s}$ and a social network G such that the single individually rational outcome achieving the maximum social welfare among such outcomes is not Nash stable.*

Proof. Consider again the scoring function $\vec{s} = (1,1,-1,-1,-1,-1)$. Similarly to previous lemma, consider the social network G in Figure 3 formed from a path P on 5 vertices and a clique K on 4 vertices by connecting the endpoints of P to all vertices of K and adding a agent y only connected to the central agent of P which we call x. Let C be the coalition containing all vertices of G except for y. As in the previous lemma, $G[C]$ can be viewed as a 6-cycle with K forming one "bold" agent. Hence, $\mathrm{u}_x(C) = 4-4 = 0$, while utility of other agents in C is $7-1 = 6$. Trivially $\mathrm{u}_y(\{y\}) = 0$, hence the outcome $(\{y\},C)$ is individually rational. It has total social welfare of 48. However, it is not Nash stable, as x wants to deviate to $\{x,y\}$ giving them utility 1.

However, the outcome $(\{x,y\},C\setminus\{x\})$, which is Nash stable, has total social welfare only 46. Note that $\mathrm{u}_z(C\setminus\{x\}) \geq 3$ for every agent $z \in C\setminus\{x\}$, so any outcome $(\{x,y,z\},C\setminus\{x,z\})$ cannot be Nash stable. While the total social welfare of the grand coalition is 46, the utility of y is $3-6 = -3$ in this coalition, so this outcome is not even individually rational. From the computations in the previous lemma, it follows, that to attain the social welfare of 48, the largest coalition in the outcome must be of size at least 7. Moreover, if it is of size exactly 7, then these 7 vertices must be at mutual distance at most 2. However, there are no 7 vertices in mutual distance at most 2 in G. Hence, in any outcome with social welfare 48 the largest coalition must be of size at least 8. Agent y has only 3 agents in distance at most 2 in G. Hence, for y to get a positive utility from some coalition, the coalition must be of size at most 7, i.e., y cannot be part of the largest coalition in any outcome with social welfare at least 48. However, for every $z \in C$, z joining the coalition $C\setminus\{z\}$ improves the social welfare of the outcome, proving that it was not optimal.

Hence the outcome $(\{y\},C)$ is the only individually rational outcome with maximal social welfare, but it is not Nash stable. □

It should be noted that Lemmas 2 and 3 also contrast many other models where outputs maximizing social welfare are stable for symmetric utilities [11, 6, 15].

As our next two structural results, we prove that on certain SDGs it is possible to bound not only the diameter but also the size of each coalition in a welfare-maximum outcome. Notably, we establish such bounds for SDGs on bounded-degree networks and SDGs which have a simple scoring vector on a tree-like network. While arguably interesting in their own right, these properties will be important for establishing the fixed-parameter tractability of computing welfare-optimal outcomes in the next section.

Lemma 4. *For every scoring vector $\vec{s} = (s_1,\dots,s_\delta)$, if G is a graph of maximum degree $\Delta(G)$ and C is a coalition of size more than $(s_1+1)\cdot\Delta(G)\cdot(\Delta(G)-1)^{\delta-1}$, then for every $i \in C$ we have $\mathrm{u}(i,C) < 0$.*

Proof. Let $i \in C$. There are at most $\Delta(G)\cdot(\Delta(G)-1)^{\delta-1}$ agents in distance at most δ from i. Each of these agents contributes at most s_1 to $\mathrm{u}(i,C)$. Every other agent contributes at most -1. Hence, if there are more than $(s_1+1)\cdot\Delta(G)\cdot(\Delta(G)-1)^{\delta-1}$ agents in C, then more than $s_1\cdot\Delta(G)\cdot(\Delta(G)-1)^{\delta-1}$ of them have a negative contribution to $\mathrm{u}(i,C)$ and

$$\mathrm{u}(i,C) < s_1\cdot\Delta(G)\cdot(\Delta(G)-1)^{\delta-1} - 1\cdot s_1\cdot\Delta(G)\cdot(\Delta(G)-1)^{\delta-1} = 0. \qquad \square$$

Lemma 5. *Let $\vec{s} = (s_1,\dots,s_\delta)$ be such that $s_2 < 0$. If G is a graph of treewidth tw and C is a coalition of size more than $2(s_1+1)\cdot\mathrm{tw}+1$, then $\sum_{i\in C}\mathrm{u}(i,C) < 0$.*

Proof. Each agent adjacent to i contributes s_1 to $\mathrm{u}(i,C)$, whereas all the other agents contribute at most -1. Since a graph of treewidth tw is tw-degenerate, there are $|E(G[C])| \leq |C|\cdot\mathrm{tw}$ pairs of adjacent agents and $\binom{|C|}{2} - |E(G[C])|$ pairs of non-adjacent agents. We have

$$\begin{aligned}
\sum_{i\in C} \mathrm{u}(i,C) &= \sum_{i,j\in C; i\neq j} \vec{\mathrm{s}}(\mathrm{dist}(i,j)) \\
&\leq 2\left(s_1 \cdot |E(G[C])| - \left(\binom{|C|}{2} - |E(G[C])|\right)\right) \\
&= 2\left((s_1+1)\cdot |E(G[C])| - \binom{|C|}{2}\right) \\
&\leq 2(s_1+1)\cdot |C| \cdot \mathrm{tw} - |C|(|C|-1) \\
&= |C|\left(2(s_1+1)\cdot \mathrm{tw} - (|C|-1)\right) \\
&< |C|\left(2(s_1+1)\cdot \mathrm{tw} - (2(s_1+1)\cdot \mathrm{tw} + 1 - 1)\right) = 0.
\end{aligned}$$

□

4 Computing Optimal Outcomes

4.1 Intractability

As our first step towards an understanding of the complexity of computing a welfare-optimal outcome in an SDG, we establish the NP-hardness of $\vec{\mathrm{s}}$-SDG-WF, $\vec{\mathrm{s}}$-SDG-WF-IR and $\vec{\mathrm{s}}$-SDG-WF-NASH even for a very simple choice of $\vec{\mathrm{s}}$.

Theorem 6. *Let* $\vec{\mathrm{s}} = (s_1)$ *for any* $s_1 > 0$. *Then* $\vec{\mathrm{s}}$-SDG-WF, $\vec{\mathrm{s}}$-SDG-WF-IR *and* $\vec{\mathrm{s}}$-SDG-WF-NASH *are* NP-*hard.*

Proof Sketch. As our first step, we prove the NP-hardness of the intermediate problem called 3-COLORING TRIANGLE COVERED GRAPH (3CTCG) via an adaptation of a known reduction from NOTALL-EQUAL-3-SAT [37, Theorem 9.8]:

3-COLORING TRIANGLE COVERED GRAPH (3CTCG)
Input: An undirected graph $G = (V,E)$ with $|V| = 3n$ vertices such that G contains a collection of n mutually vertex disjoint triangles.
Question: Does G have a 3-coloring?

Next, we reduce 3CTCG to our three problems via a single construction. Let G be an instance of 3CTCG with $3n$ vertices and $T_1,\dots,T_n$ the corresponding collection of triangles. Let $\overline{G}$ be a complement of G, let $s_1 = s_1(\vec{\mathrm{s}})$ and let $b = 3ns_1 \cdot (n-1)$. To establish the NP-hardness of $\vec{\mathrm{s}}$-SDG-WF, it suffices to show that G is a Yes-instance of 3CTCG if and only if $\overline{G}$ admits an outcome with social welfare at least b; for the remaining two problems, we additionally show that such an outcome will furthermore be individually rational and Nash stable. □

4.2 An Algorithm for Tree-Like Networks

We complement Theorem 6 by establishing that all three problems under consideration can be solved in polynomial time on networks of bounded treewidth—in other words, we show that they are XP-tractable w.r.t. treewidth. We first describe the "baseline" algorithm for solving $\vec{\mathrm{s}}$-SDG-WF, and then prove that this may be adapted to also solve the other two problems by expanding on its records and procedures (see the appendix).

Theorem 7. *For every fixed scoring vector $\vec{s}$, the $\vec{s}$-SDG-WF, $\vec{s}$-SDG-WF-IR, and $\vec{s}$-SDG-WF-NASH problems are in* XP *when parameterized by the treewidth of the social network G.*

Proof Sketch. Our algorithm is based on leaf-to-root dynamic programming along a nice tree-decomposition of the input social network with rather complicated structure. In each node x of the tree-decomposition, we store a set $\mathscr{R}_x$ of partial solutions called *records*. Each record realizes a single *signature* which is a triple (C,S,T), where

- C is a partition of bag agents into parts of coalitions; there are at most $\mathrm{tw}+1$ different coalitions intersecting $\beta(x)$ and, thus, at most $tw^{\mathscr{O}(\mathrm{tw})}$ possible partitions of $\beta(x)$.
- S is a function assigning each pair of agents that are part of the same coalition according to C the shortest intra-coalitional path; recall that for fixed $\vec{s}$, the diameter of every coalition is bounded by a constant δ and, therefore, there are $n^{\mathscr{O}(\delta)} = n^{\mathscr{O}(1)}$ possible paths for each pair of agents which gives us $n^{\mathscr{O}(\mathrm{tw}^2)}$ combinations in total.
- T is a table storing for every coalition P and every possible vector of distances to bag agents that are in P the number of agents from P that were already forgotten in some node of the tree-decomposition; the number of possible coalitions is at most $\mathrm{tw}+1$, the number of potential distance vectors is $\delta^{\mathrm{tw}+1} = 2^{\mathscr{O}(\mathrm{tw})}$, and there are at most n values for every combination of coalition and distance vector which leads to at most $n^{2^{\mathscr{O}(\mathrm{tw})}}$ different tables T.

The value of every record is a pair (π,w), where π is a partition of V^x such that $\mathrm{SW}(\pi) = w$ and π witnesses that there is a partition of V^x corresponding to the signature of the record, as described above. We store only one record for every signature – the one with the highest social welfare. Therefore, in every node x, there are at most $n^{2^{\mathscr{O}(\mathrm{tw})}}$ different records.

Once the computation ends, we check the record in the root node r and based on the value of w, we return the answer; Yes if $w \geq b$ and No otherwise. Moreover, as $G^r = G$, the partition π is also an outcome admitting social-welfare w. □

4.3 Fixed-Parameter Tractability

A natural follow-up question to Theorem 7 is whether one can improve these results to fixed-parameter algorithms. As our final contribution, we show that this is possible at least when dealing with simple scoring vectors, or on networks with stronger structural restrictions. To obtain both of these results, we first show that to obtain fixed-parameter tractability it suffices to have a bound on the size of the largest coalition in a solution (i.e., a welfare-optimal outcome).

Theorem 8. *For every fixed scoring vector $\vec{s}$, the variants of $\vec{s}$-SDG-WF, $\vec{s}$-SDG-WF-IR, $\vec{s}$-SDG-WF-NASH where we only consider outcomes consisting of coalitions of at most a prescribed size are* FPT *parameterized by the treewidth of the network and the maximum coalition size combined.*

Proof Sketch. Similar to the previous ones, we design a dynamic programming (DP) on a nice tree decomposition, albeit the procedure and records are completely different.

Given a subset of agents $X \subseteq N$, let $\Pi = (\pi_1, \pi_2, \ldots, \pi_\ell)$ be a partition of a set containing X and some "anonymous" agents. We use $\mathsf{T}(\Pi)$ to denote a set of graph topologies on $\pi_1, \pi_2, \ldots, \pi_\ell$ given X. That is, $\mathsf{T}(\Pi) = \{\mathsf{T}(\pi_1), \ldots, \mathsf{T}(\pi_\ell)\}$ where $\mathsf{T}(\pi_i)$ is some graph on $|\pi_i|$ agents, namely $\pi_i \cap X$ and $|\pi_i \setminus X|$ "anonymous" agents, for each $i \in [\ell]$. The maximum coalition size of any welfare maximizing partition is denoted by sz. Table, M, contains an entry $\mathsf{M}[x, C, \mathsf{T}(\Pi)]$ for every node x of the tree decomposition, each partition C of $\beta(x)$, and each set of graph topologies $\mathsf{T}(\Pi)$ given $\beta(x)$ where Π is a partition of at most $\mathrm{sz} \cdot \mathrm{tw}$ agents. An entry of M stores the maximum welfare in G^x under the condition that the partition into coalitions satisfies the following properties. Recall that for a partition P of agents and an agent a, we use P_a to denote the coalition agent a is part of in P.

1. *C and Π are consistent*, i.e., the partition of the bag agents $\beta(x)$ in G^x is denoted by C and $C_a = \Pi_a \cap \beta(x)$ for each agent $a \in \beta(x)$.
2. The coalition of agent $a \in \beta(x)$ in the graph G^x is Π_a.
3. $\mathsf{T}(\Pi)$ *is consistent with G^x* i.e., the subgraph of G^x induced on the agents in coalition of a is $\mathsf{T}(\Pi_a)$, i.e., $G^x[\Pi_a] = \mathsf{T}(\Pi_a)$.

Observe that we do not store Π. We only store the topology of Π which is a graph on at most $\mathrm{sz} \cdot \mathrm{tw}$ agents.

We say an entry of $\mathsf{M}[x, C, \mathsf{T}(\Pi)]$ is *valid* if it holds that

1. *C and Π are consistent*, i.e., $C_a = \Pi_a \cap \beta(x)$ for each agent $a \in \beta(x)$,
2. Either $C_a = C_b$, or $C_a \cap C_b = \emptyset$ for each pair of agents $a, b \in \beta(x)$,
3. $\mathsf{T}(\Pi)$ *is consistent with G^x in $\beta(x)$*, i.e., for each pair of agents $a, b \in \beta(x)$ such that $\Pi_a = \Pi_b$, there is an edge $(a, b) \in \mathsf{T}(\Pi_a)$ if and only if (a, b) is an edge in G^x.

Once the table is computed correctly, the solution is given by the value stored in $\mathsf{M}[r, C, \mathsf{T}(\Pi)]$ where C is empty partition and $\mathsf{T}(\Pi)$ is empty. Roughly speaking, the basis corresponds to leaves (whose bags are empty), and are initialized to store 0. For each entry that is not valid we store $-\infty$. To complete the proof, it now suffices to describe the computation of the records at each of the three non-trivial types of nodes in the decomposition and prove correctness. □

Similarly to Theorem 7, we design a dynamic programming on a nice tree decomposition, albeit the procedure and records are completely different.

From Lemma 5 it follows that if $s_2 < 0$ and $\mathrm{tw}(G)$ is bounded, then the maximum coalition size of a welfare maximizing outcome is bounded. Hence, using Theorem 8 we get the following.

Corollary 9. $\vec{s}$-SDG-WF-NASH, $\vec{s}$-SDG-WF-IR, *and* $\vec{s}$-SDG-WF *are fixed-parameter tractable parameterized by the treewidth* $\mathrm{tw}(G)$ *if* $s_2 < 0$.

Turning back to general scoring vectors, we recall that Lemma 4 provided a bound on the size of the coalitions in a welfare-optimal outcome in terms of the maximum degree $\Delta(G)$ of the network G. Applying Theorem 8 again yields:

Corollary 10. $\vec{s}$-SDG-WF-NASH, $\vec{s}$-SDG-WF-IR, *and* $\vec{s}$-SDG-WF *are fixed-parameter tractable parameterized by the treewidth* $\mathrm{tw}(G)$ *and the maximum degree* $\Delta(G)$ *of the social network.*

As our final contribution, we provide fixed-parameter algorithms for computing welfare-optimal outcomes that can also deal with networks containing high-degree agents. To do so, we exploit a different structural parameter than the treewidth—namely the vertex cover number of G ($\mathrm{vc}(G)$). We note that while the vertex cover number is a significantly more "restrictive" graph parameter than treewidth, it has found numerous applications in the design of efficient algorithms in coalition formation, including for other types of coalition games [5, 8, 26].

Theorem 11. $\vec{s}$-SDG-WF-NASH, $\vec{s}$-SDG-WF-IR, *and* $\vec{s}$-SDG-WF *are fixed-parameter tractable parameterized by the vertex cover number* $\mathrm{vc}(G)$ *of the social network.*

Proof Sketch. Let $k = \mathrm{vc}(G)$ and let U be a vertex cover for G of size k. Observe that in each solution there are at most k non-singleton coalitions, since G has a vertex cover of size k and each coalition must be connected. Furthermore, the vertices of $G - U$ can be partitioned into at most 2^k groups according to their neighborhood in the set U. That is, there are n_W vertices in $G - U$ such that their neighborhood is W for some $W \subseteq U$; denote this set of vertices I_W.

We perform exhaustive branching to determine certain information about the structure of the coalitions in a solution—notably:

1. which vertices of U belong to each coalition (i.e., we partition the set U); note that there are at most k^k such partitions, and
2. if there is at least one agent of I_W in the coalition or not ; note that there are at most $(2^{2^k})^k$ such assignments of these sets to the coalitions.

We branch over all possible admissible options of the coalitional structure described above possessed by a hypothetical solution. The total number of branches is upper-bounded by a function of the parameter value k and thus for the problems to be in FPT it suffices to show that for each branch we can find a solution (if it exists) by a fixed-parameter subprocedure. To conclude the proof, we show that a welfare-maximum outcome (which furthermore satisfies the imposed stability constraints) with a given coalitional structure can be computed by modeling this as an Integer Quadratic Program where $d + \|A\|_\infty + \|Q\|_\infty$ are all upper-bounded by a function of k—such a program can be solved in FPT time using Proposition 1.

The (integer) variables of the program are x_W^C, which express the number of vertices from the set I_W in the coalition with $C \subseteq U$; thus, we have $x_W^C \in \mathbb{Z}$ and $x_W^C \geq 1$. Let $\mathscr{C}$ be the considered partitioning of the vertex cover U. We use $C \in \mathscr{C}$ for the set $C \subseteq U$ in the coalition and C^+ for the set C and the guessed groups having at least one agent in the coalition. We require that the vertices of $G - U$ are also partitioned in the solution, i.e.,

$$\sum_{C \in \mathscr{C}} \sum_{W \in C^+} x_W^C = n_W \qquad \forall W \subseteq U. \tag{1}$$

The quadratic objective expresses the welfare of the coalitions in the solution while the linear constraints ensure the stability of the outcome; for the latter, we rely on the fact that it is sufficient to verify the stability for a single agent from the group I_W in each coalition. □

5 Conclusions and Future Research Directions

In this work, we studied social distance games through the lens of an adaptable, non-normalized scoring vector which can capture the positive as well as negative dynamics of social interactions within coalitions. The main focus of this work was on welfare maximization, possibly in combination with individual-based stability notions—individual rationality and Nash stability. It is not surprising that these problems are intractable for general networks; we complement our model with algorithms that work well in tree-like environments.

Our work opens up a number of avenues for future research. One can consider other notions of individual-based stability such as individual stability [13, pp. 360–361][23], or various notions of group-based stability such as core stability [13, p. 360][14, 34]. Furthermore, our results do not settle the complexity of finding stable solutions (without simultaneous welfare maximization). Therefore, it remains open if one can find a Nash stable solution for a specific scoring vector. Also, a more complex open problem is to characterize those scoring vectors that guarantee the existence of a Nash (or individually) stable solution.

Finally, we remark that the proposed score-based SDG model can be generalized further, e.g., by allowing for a broader definition of the scoring vectors. For instance, it is easy to generalize all our algorithms to scoring vectors which are not monotone in their "positive part". One could also consider situations where the presence of an agent that is "far away" does not immediately set the utility of other agents in the coalition to $-\infty$. One way to model these settings would be to consider "*open*" scoring vectors, for which we set $\vec{s}(a) = \vec{s}(\delta)$ for all $a > \delta$—meaning that distances over δ are all treated uniformly but not necessarily as unacceptable.

Notice that if $\vec{s}(\delta) \geq 0$ for an open scoring vector $\vec{s}$, the grand coalition is always a social-welfare maximizing outcome for all three problems—hence here it is natural to focus on choices of $\vec{s}$ with at least one negative entry. We note that all of our fixed-parameter algorithms immediately carry over to this setting for arbitrary choices of open scoring vectors $\vec{s}$. The situation becomes more interesting when considering the small-world property: while the diameter of every welfare-maximizing outcome can be bounded in the case of Nash stable or individually rational coalitions (as we prove in our final Theorem 12 below), whether the same holds in the case of merely trying to maximize social welfare is open and seems to be a non-trivial question. Because of this, Theorem 7 can also be extended to the $\vec{s}$-SDG-WF-IR and $\vec{s}$-SDG-WF-NASH with open scoring vectors, but it is non-obvious for $\vec{s}$-SDG-WF.

Theorem 12. *Let $\vec{s} = (s_1, \dots, s_\delta)$ be an arbitrary open scoring vector and G be a social network. Every outcome Π containing a coalition $C \in \Pi$ with diameter exceeding $\ell = 2 \cdot s_1 \cdot \delta$ can be neither Nash-stable nor individually rational.*

Proof Sketch. Consider a shortest path P in C whose length exceeds ℓ. We identify a set of edge cuts along P and show that at least one such cut must be near an agent whose utility in C is negative, due to the presence of a large number of agents that must be distant from the chosen edge cut. □

Acknowledgements. All authors are grateful for support from the OeAD bilateral Czech-Austrian WTZ-funding Programme (Projects No. CZ 05/2021 and 8J21AT021). Robert Ganian acknowledges support from the Austrian Science Foundation (FWF, project Y1329). Thekla Hamm also acknowledges support from FWF, project J4651-N. Dušan Knop, Šimon Schierreich, and Ondřej Suchý acknowledge the support of the Czech Science Foundation Grant No. 22-19557S. Šimon Schierreich was additionally supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/205/OHK3/3T/18.

References

[1] Alkida Balliu, Michele Flammini, Giovanna Melideo & Dennis Olivetti (2017): *Nash Stability in Social Distance Games.* In Satinder Singh & Shaul Markovitch, editors: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI '17*, AAAI Press, pp. 342–348, doi:10.1609/aaai.v31i1.10608.

[2] Alkida Balliu, Michele Flammini, Giovanna Melideo & Dennis Olivetti (2019): *On Non-Cooperativeness in Social Distance Games.* *Journal of Artificial Intelligence Research* 66, pp. 625–653, doi:10.1613/jair.1.11808.

[3] Alkida Balliu, Michele Flammini, Giovanna Melideo & Dennis Olivetti (2022): *On Pareto optimality in social distance games.* *Artificial Intelligence* 312, p. 103768, doi:10.1016/j.artint.2022.103768.

[4] Nathanaël Barrot & Makoto Yokoo (2019): *Stable and Envy-free Partitions in Hedonic Games*. In Sarit Kraus, editor: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI '19*, ijcai.org, pp. 67–73, doi:10.24963/ijcai.2019/10.

[5] Vittorio Bilò, Angelo Fanelli, Michele Flammini, Gianpiero Monaco & Luca Moscardelli (2018): *Nash Stable Outcomes in Fractional Hedonic Games: Existence, Efficiency and Computation*. *Journal of Artificial Intelligence Research* 62, pp. 315–371, doi:10.1613/jair.1.11211.

[6] Vittorio Bilò, Gianpiero Monaco & Luca Moscardelli (2022): *Hedonic Games with Fixed-Size Coalitions*. In: *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI '22*, AAAI Press, pp. 9287–9295, doi:10.1609/aaai.v36i9.21156.

[7] Hans L. Bodlaender (1996): *A Linear-Time Algorithm for Finding Tree-Decompositions of Small Treewidth*. *SIAM Journal on Computing* 25(6), pp. 1305–1317, doi:10.1137/S0097539793251219.

[8] Hans L. Bodlaender, Tesshu Hanaka, Lars Jaffke, Hirotaka Ono, Yota Otachi & Tom C. van der Zanden (2020): *Hedonic Seat Arrangement Problems*. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, IFAAMAS, Richland, SC, p. 1777–1779. Available at https://dl.acm.org/doi/10.5555/3398761.3398979.

[9] Niclas Boehmer & Edith Elkind (2020): *Individual-Based Stability in Hedonic Diversity Games*. In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI '20*, AAAI Press, pp. 1822–1829, doi:10.1609/aaai.v34i02.5549.

[10] Niclas Boehmer & Edith Elkind (2020): *Stable Roommate Problem With Diversity Preferences*. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An & Neil Yorke-Smith, editors: *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20*, IFAAMAS, pp. 1780–1782. Available at https://dl.acm.org/doi/10.5555/3398761.3398980.

[11] Anna Bogomolnaia & Matthew O. Jackson (2002): *The Stability of Hedonic Coalition Structures*. *Games and Economic Behavior* 38(2), pp. 201–230, doi:10.1006/game.2001.0877.

[12] Sylvain Bouveret & Jérôme Lang (2008): *Efficiency and Envy-freeness in Fair Division of Indivisible Goods: Logical Representation and Complexity*. *Journal of Artificial Intelligence Research* 32, pp. 525–564, doi:10.1613/jair.2467.

[13] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D. Procaccia, editors (2016): *Handbook of Computational Social Choice*. Cambridge University Press, doi:10.1017/CBO9781107446984.

[14] Simina Brânzei & Kate Larson (2011): *Social Distance Games*. In Toby Walsh, editor: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI '11*, IJCAI/AAAI, pp. 91–96, doi:10.5591/978-1-57735-516-8/IJCAI11-027.

[15] Martin Bullinger & Warut Suksompong (2023): *Topological Distance Games*. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI '23*, AAAI Press.

[16] Jiehua Chen, Robert Ganian & Thekla Hamm (2020): *Stable Matchings with Diversity Constraints: Affirmative Action is beyond NP*. In Christian Bessiere, editor: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI '20*, ijcai.org, pp. 146–152, doi:10.24963/ijcai.2020/21.

[17] Marek Cygan, Fedor V. Fomin, Łukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk & Saket Saurabh (2015): *Parameterized Algorithms*. Springer, doi:10.1007/978-3-319-21275-3.

[18] Reinhard Diestel (2017): *Graph Theory*, 5th edition. Graduate Texts in Mathematics, Springer, Berlin, Heidelberg, doi:10.1007/978-3-662-53622-3.

[19] Rodney G. Downey & Michael R. Fellows (2013): *Fundamentals of Parameterized Complexity*. Texts in Computer Science, Springer, doi:10.1007/978-1-4471-5559-1.

[20] Edith Elkind & Anisse Ismaili (2015): *OWA-Based Extensions of the Chamberlin-Courant Rule*. In Toby Walsh, editor: *Proceedings of the 4th International Conference Algorithmic Decision Theory, ADT '15*, *Lecture Notes in Computer Science* 9346, Springer, pp. 486–502, doi:10.1007/978-3-319-23114-3_29.

[21] Michele Flammini, Bojana Kodric, Martin Olsen & Giovanna Varricchio (2020): *Distance Hedonic Games*. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An & Neil Yorke-Smith, editors: *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20*, IFAAMAS, pp. 1846–1848. Available at `https://dl.acm.org/doi/10.5555/3398761.3399002`.

[22] Michele Flammini, Bojana Kodric, Martin Olsen & Giovanna Varricchio (2021): *Distance Hedonic Games*. In Tomás Bures, Riccardo Dondi, Johann Gamper, Giovanna Guerrini, Tomasz Jurdzinski, Claus Pahl, Florian Sikora & Prudence W. H. Wong, editors: *Proceedings of the 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM '21*, *Lecture Notes in Computer Science* 12607, Springer, pp. 159–174, doi:`10.1007/978-3-030-67731-2_12`.

[23] Robert Ganian, Thekla Hamm, Dušan Knop, Šimon Schierreich & Ondřej Suchý (2022): *Hedonic Diversity Games: A Complexity Picture with More than Two Colors*. In: *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI '22*, AAAI Press, pp. 5034–5042, doi:`10.1609/aaai.v36i5.20435`.

[24] Robert Ganian & Viktoriia Korchemna (2021): *The Complexity of Bayesian Network Learning: Revisiting the Superstructure*. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang & Jennifer Wortman Vaughan, editors: *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems, NeurIPS '21*, Curran Associates, Inc., pp. 430–442. Available at `https://proceedings.neurips.cc/paper/2021/hash/040a99f23e8960763e680041c601acab-Abstract.html`.

[25] Tomáš Gavenčiak, Martin Koutecký & Dušan Knop (2022): *Integer programming in parameterized complexity: Five miniatures*. *Discrete Optimization* 44(Part 1), p. 100596, doi:`10.1016/j.disopt.2020.100596`.

[26] Tesshu Hanaka & Michael Lampis (2022): *Hedonic Games and Treewidth Revisited*. In Shiri Chechik, Gonzalo Navarro, Eva Rotenberg & Grzegorz Herman, editors: *Proceedings of the 30th Annual European Symposium on Algorithms, ESA '22*, *Leibniz International Proceedings in Informatics* 244, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 64:1–64:16, doi:`10.4230/LIPIcs.ESA.2022.64`.

[27] Matthew O. Jackson (2008): *Social and economic networks*. Princeton University Press, Princeton, NJ, doi:`10.1515/9781400833993`.

[28] Christos Kaklamanis, Panagiotis Kanellopoulos & Dimitris Patouchas (2018): *On the Price of Stability of Social Distance Games*. In Xiaotie Deng, editor: *Proceedings of the 11th International Symposium Algorithmic Game Theory, SAGT '18*, *Lecture Notes in Computer Science* 11059, Springer, pp. 125–136, doi:`10.1007/978-3-319-99660-8_12`.

[29] Mehmet Karakaya (2011): *Hedonic coalition formation games: A new stability notion*. *Mathematical Social Sciences* 61(3), pp. 157–165, doi:`10.1016/j.mathsocsci.2011.03.004`.

[30] Ton Kloks (1994): *Treewidth: Computations and Approximations*. *Lecture Notes in Computer Science* 842, Springer, Berlin, Heidelberg, doi:`10.1007/BFb0045375`.

[31] Tuukka Korhonen (2021): *A Single-Exponential Time 2-Approximation Algorithm for Treewidth*. In: *Proceedings of the 62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS '21*, IEEE, pp. 184–192, doi:`10.1109/FOCS52979.2021.00026`.

[32] Daniel Lokshtanov (2015): *Parameterized Integer Quadratic Programming: Variables and Coefficients*. *CoRR* abs/1511.00310, doi:`10.48550/arXiv.1511.00310`. arXiv:`1511.00310`.

[33] Silviu Maniu, Pierre Senellart & Suraj Jog (2019): *An Experimental Study of the Treewidth of Real-World Graph Data*. In Pablo Barceló & Marco Calautti, editors: *Proceedings of the 22nd International Conference on Database Theory, ICDT '19*, *Leibniz International Proceedings in Informatics* 127, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 12:1–12:18, doi:`10.4230/LIPIcs.ICDT.2019.12`.

[34] Kazunori Ohta, Nathanaël Barrot, Anisse Ismaili, Yuko Sakurai & Makoto Yokoo (2017): *Core Stability in Hedonic Games among Friends and Enemies: Impact of Neutrals*. In Carles Sierra, editor: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI '17*, ijcai.org, pp. 359–365, doi:`10.24963/ijcai.2017/51`.

[35] Masahiro Okubo, Tesshu Hanaka & Hirotaka Ono (2019): *Optimal Partition of a Tree with Social Distance*. In Gautam K. Das, Partha Sarathi Mandal, Krishnendu Mukhopadhyaya & Shin-Ichi Nakano, editors: *Proceedings of the 13th International Conference on Algorithms and Computation, WALCOM '19*, *Lecture Notes in Computer Science* 11355, Springer, pp. 121–132, doi:10.1007/978-3-030-10564-8_10.

[36] Sebastian Ordyniak & Stefan Szeider (2013): *Parameterized Complexity Results for Exact Bayesian Network Structure Learning*. *Journal of Artificial Intelligence Research* 46, pp. 263–302, doi:10.1613/jair.3744.

[37] Christos H. Papadimitriou (1994): *Computational complexity*. Addison-Wesley.

[38] Shao Chin Sung & Dinko Dimitrov (2007): *On Myopic Stability Concepts for Hedonic Games*. *Theory and Decision* 62(1), pp. 31–45, doi:10.1007/s11238-006-9022-2.

[39] Kevin Zemmer (2017): *Integer Polynomial Optimization in Fixed Dimension*. Doctoral thesis, ETH Zurich, Zurich, doi:10.3929/ethz-b-000241796.

System of Spheres-based Two Level Credibility-limited Revisions

Marco Garapa
Universidade da Madeira
CIMA - Centro de Investigação
em Matemática e Aplicações
mgarapa@staff.uma.pt

Eduardo Fermé
Universidade da Madeira
NOVA Laboratory for Computer Science
and Informatics (NOVA LINCS)
eduardo.ferme@staff.uma.pt

Maurício D. L. Reis
Universidade da Madeira
CIMA - Centro de Investigação
em Matemática e Aplicações
m_reis@staff.uma.pt

Two level credibility-limited revision is a non-prioritized revision operation. When revising by a two level credibility-limited revision, two levels of credibility and one level of incredibility are considered. When revising by a sentence at the highest level of credibility, the operator behaves as a standard revision, if the sentence is at the second level of credibility, then the outcome of the revision process coincides with a standard contraction by the negation of that sentence. If the sentence is not credible, then the original belief set remains unchanged. In this paper, we propose a construction for two level credibility-limited revision operators based on Grove's systems of spheres and present an axiomatic characterization for these operators.

1 Introduction

Belief Change (also called *Belief Revision*) is an area that studies the dynamics of belief. One of the main goals underlying this area is to model how a rational agent updates her set of beliefs when confronted with new information. The main model of belief change is the AGM model [1]. In that model, each *belief* of an agent is represented by a sentence and the *belief state* of an agent is represented by a logically closed set of (belief-representing) sentences. These sets are called *belief sets*. A change consists in adding or removing a specific sentence from a belief set to obtain a new belief set. The AGM model considers three kinds of belief change operators, namely *expansion, contraction* and *revision*. An expansion occurs when new information is added to the set of the beliefs of an agent. The expansion of a belief set $\mathbf{K}$ by a sentence α (denoted by $\mathbf{K}+\alpha$) is the logical closure of $\mathbf{K}\cup\{\alpha\}$. A contraction occurs when information is removed from the set of beliefs of an agent. A revision occurs when new information is added to the set of the beliefs of an agent while retaining consistency if the new information is itself consistent. From the three operations, expansion is the only one that can be univocally defined. The other two operations are characterized by a set of postulates that determine the behaviour of each one of these functions, establishing conditions or constrains that they must satisfy.

Although the AGM model has acquired the status of standard model of belief change, several researchers (for an overview see [5, 6]) have pointed out its inadequateness in several contexts and proposed several extensions and generalizations to that framework. One of the criticisms to the AGM model that appears in the belief change literature is the total acceptance of the new information, which is characterized by the *success* postulate for revision. “The AGM model always accepts the new information. This feature appears, in general, to be unrealistic, since rational agents, when confronted with information that contradicts previous beliefs, often reject it altogether or accept only parts of it” ([7]). This may happen for various reasons. For example, the new information may lack on credibility or it may contradict previous highly entrenched beliefs.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 287–298, doi:10.4204/EPTCS.379.23

Models in which the belief change operators considered do not satisfy the *success* postulate are designated by *non-prioritized belief change operators* ([17]). The output of a non-prioritized revision may not contain the new belief that has motivated that revision.

Two level credibility-limited revision operators (two level CL revision operators for short) are non-prioritized revision operators that were proposed (independently) in [8] and [3]. When revising by means of a two level CL revision operator two levels of credibility and one level of incredibility are considered. When revising by a sentence at the highest level of credibility, the operator behaves as a standard revision. In this case the new information is incorporated in the agent's belief set. If the sentence is at the second level of credibility, then the outcome of the revision process coincides with a standard contraction by the negation of that sentence. In this case, the new information is not accepted but all the beliefs that are inconsistent with it are removed. The intuition underlying this behaviour is that, the belief is not credible enough to be incorporated in the agent's belief set, but creates some doubt in the agent's mind making her remove all the beliefs that are inconsistent with it.

In this paper, we propose a construction for *two level CL revision operators* based on Grove's systems of spheres and present an axiomatic characterization for these operators. The rest of the paper is organized as follows: In Section 2 we introduce the notations and recall the main background concepts and results that will be needed throughout this article. In Section 3 we present the two level CL revision operators and an axiomatic characterization for a class of these operators. In Section 4 we propose a construction for *two level CL revision operators* based on Grove's systems of spheres and present an axiomatic characterization for these operators. In Section 5, we present a brief survey of related works. In Section 6, we summarize the main contributions of the paper.

2 Background

2.1 Formal Preliminaries

We will assume a propositional language $\mathscr{L}$ that contains the usual truth functional connectives: $\neg$ (negation), $\wedge$ (conjunction), $\vee$ (disjunction), $\rightarrow$ (implication) and $\leftrightarrow$ (equivalence). We will also use $\mathscr{L}$ to denote the set of all formulas of the language. We shall make use of a consequence operation Cn that takes sets of sentences to sets of sentences and which satisfies the standard Tarskian properties, namely *inclusion, monotony* and *iteration*. Furthermore, we will assume that Cn satisfies *supraclassicality, compactness* and *deduction*. We will sometimes use $Cn(\alpha)$ for $Cn(\{\alpha\})$, $A \vdash \alpha$ for $\alpha \in Cn(A)$, $\vdash \alpha$ for $\alpha \in Cn(\emptyset)$, $A \not\vdash \alpha$ for $\alpha \notin Cn(A)$, $\not\vdash \alpha$ for $\alpha \notin Cn(\emptyset)$. The letters $\alpha, \beta, \ldots$ will be used to denote sentences of $\mathscr{L}$. $A, B, \ldots$ shall denote sets of sentences of $\mathscr{L}$. $\mathbf{K}$ is reserved to represent a set of sentences that is closed under logical consequence (i.e. $\mathbf{K} = Cn(\mathbf{K})$) — such a set is called a *belief set* or *theory*. Given a belief set $\mathbf{K}$ we will denote $Cn(\mathbf{K} \cup \{\alpha\})$ by $\mathbf{K} + \alpha$. We will use the symbol $\top$ to represent an arbitrary tautology and the symbol $\perp$ to represent an arbitrary contradiction. A possible world is a maximal consistent subset of $\mathscr{L}$. The set of all possible worlds will be denoted by $\mathscr{M}_{\mathscr{L}}$. Sets of possible worlds are called propositions. The set of possible worlds that contain $R \subseteq \mathscr{L}$ is denoted by $\|R\|$, i.e., $\|R\| = \{M \in \mathscr{M}_{\mathscr{L}} : R \subseteq M\}$. If R is inconsistent, then $\|R\| = \emptyset$. The elements of R are designated by $R-$worlds. For any sentence α, $\|\alpha\|$ is an abbreviation of $\|Cn(\{\alpha\}\|)$ and its elements are designated by α-worlds.

2.2 AGM Revisions

The operation of revision of a belief set consists of the incorporation of new beliefs in that set. In a revision process, some previous beliefs may be retracted in order to obtain, as output, a consistent belief set. The following postulates, which were originally presented in [12, 13, 14], are commonly known as *AGM postulates for revision*:[1]

($\star$1) $\mathbf{K}\star\alpha = Cn(\mathbf{K}\star\alpha)$ (*i.e.* $\mathbf{K}\star\alpha$ is a belief set). (Closure)
($\star$2) $\alpha \in \mathbf{K}\star\alpha$. (Success)
($\star$3) $\mathbf{K}\star\alpha \subseteq \mathbf{K}+\alpha$. (Inclusion)
($\star$4) If $\neg\alpha \notin \mathbf{K}$, then $\mathbf{K}+\alpha \subseteq \mathbf{K}\star\alpha$. (Vacuity)
($\star$5) If α is consistent, then $\mathbf{K}\star\alpha$ is consistent. (Consistency)
($\star$6) If $\vdash \alpha \leftrightarrow \beta$, then $\mathbf{K}\star\alpha = \mathbf{K}\star\beta$. (Extensionality)
($\star$7) $\mathbf{K}\star\alpha \cap \mathbf{K}\star\beta \subseteq \mathbf{K}\star(\alpha\vee\beta)$. (Disjunctive overlap)
($\star$8) If $\neg\alpha \notin \mathbf{K}\star(\alpha\vee\beta)$, then $\mathbf{K}\star(\alpha\vee\beta) \subseteq \mathbf{K}\star\alpha$. (Disjunctive inclusion)

Definition 1 ([1]). *An operator $\star$ for a belief set $\mathbf{K}$ is a basic AGM revision if and only if it satisfies postulates* $(\star1)$ *to* $(\star6)$. *It is an AGM revision if and only if it satisfies postulates* $(\star1)$ *to* $(\star8)$.

2.3 AGM Contractions

A contraction of a belief set occurs when some beliefs are removed from it (and no new beliefs are added). The following postulates, which were presented in [1] (following [12, 13]), are commonly known as *AGM postulates for contraction*:

($\div$1) $\mathbf{K}\div\alpha = Cn(\mathbf{K}\div\alpha)$ (*i.e.* $\mathbf{K}\div\alpha$ is a belief set). (Closure)
($\div$2) $\mathbf{K}\div\alpha \subseteq \mathbf{K}$. (Inclusion)
($\div$3) If $\alpha \notin \mathbf{K}$, then $\mathbf{K} \subseteq \mathbf{K}\div\alpha$. (Vacuity)
($\div$4) If $\nvdash \alpha$, then $\alpha \notin \mathbf{K}\div\alpha$. (Success)
($\div$5) $\mathbf{K} \subseteq (\mathbf{K}\div\alpha)+\alpha$. (Recovery)
($\div$6) If $\vdash \alpha \leftrightarrow \beta$, then $\mathbf{K}\div\alpha = \mathbf{K}\div\beta$. (Extensionality)
($\div$7) $\mathbf{K}\div\alpha \cap \mathbf{K}\div\beta \subseteq \mathbf{K}\div(\alpha\wedge\beta)$. (Conjunctive overlap)
($\div$8) $\mathbf{K}\div(\alpha\wedge\beta) \subseteq \mathbf{K}\div\alpha$ whenever $\alpha \notin \mathbf{K}\div(\alpha\wedge\beta)$. (Conjunctive inclusion)

Definition 2 ([1]). *An operator $\div$ for a belief set $\mathbf{K}$ is a basic AGM contraction if and only if it satisfies postulates* $(\div1)$ *to* $(\div6)$. *It is an AGM contraction if and only if it satisfies postulates* $(\div1)$ *to* $(\div8)$.

There are several contraction operators that are exactly characterized by the postulates $(\div1)$ to $(\div8)$, namely the *(transitively relational) partial meet contractions* [1], *safe contraction* [2, 25], *system of spheres-based contraction* [16] and *epistemic entrenchment-based contraction* [14, 15].

The Levi and Harper identities[2] make contraction and revision interchangeable. These identities allow us to define the revision and the contraction operators in terms of each other. The Levi (respectively Harper) identity enable the use of contraction (resp. revision) as primitive function and treat revision (resp. contraction) as defined in terms of contraction (resp. revision).

[1] These postulates were previously presented in [1] but with slightly different formulations.

[2] **Harper identity:** [20] $\mathbf{K}\div\alpha = (\mathbf{K}\star\neg\alpha)\cap\mathbf{K}$.
Levi identity: [22] $\mathbf{K}\star\alpha = (\mathbf{K}\div\neg\alpha)+\alpha$.

2.4 Sphere-based Operations of Belief Change

Grove ([16]), inspired by the semantics for counterfactuals ([23]) proposed a structure called *system of spheres* to be used for defining revision functions. Figuratively, the distance between a possible world and the innermost sphere reflects its plausibility towards $\|\mathbf{K}\|$. The closer a possible world is to $\|\mathbf{K}\|$, the more plausible it is.

Definition 3 ([16])**.** *Let* $\mathbf{K}$ *be a belief set. A system of spheres, or spheres' system, centred on* $\|\mathbf{K}\|$ *is a collection* $\mathbb{S}$ *of subsets of* $\mathscr{M}_{\mathscr{L}}$*, i.e.,* $\mathbb{S} \subseteq \mathscr{P}(\mathscr{M}_{\mathscr{L}})$*, that satisfies the following conditions:*

($\mathbb{S}$1) $\mathbb{S}$ *is totally ordered with respect to set inclusion; that is, if* $U,V \in \mathbb{S}$*, then* $U \subseteq V$ *or* $V \subseteq U$.

($\mathbb{S}$2) $\|\mathbf{K}\| \in \mathbb{S}$*, and if* $U \in \mathbb{S}$*, then* $\|\mathbf{K}\| \subseteq U$ *(*$\|\mathbf{K}\|$ *is the* $\subseteq$*-minimum of* $\mathbb{S}$*).*

($\mathbb{S}$3) $\mathscr{M}_{\mathscr{L}} \in \mathbb{S}$ *(*$\mathscr{M}_{\mathscr{L}}$ *is the largest element of* $\mathbb{S}$*).*

($\mathbb{S}$4) *For every* $\alpha \in \mathscr{L}$*, if there is any element in* $\mathbb{S}$ *intersecting* $\|\alpha\|$ *then there is also a smallest element in* $\mathbb{S}$ *intersecting* $\|\alpha\|$.

The elements of $\mathbb{S}$ *are called spheres. For any consistent sentence* $\alpha \in \mathscr{L}$*, the smallest sphere in* $\mathbb{S}$ *intersecting* $\|\alpha\|$ *is denoted by* $\mathbb{S}_\alpha$.

Given a system of spheres $\mathbb{S}$ centered on $\|\mathbf{K}\|$ it is possible to define expansion, revision and contraction operators based on $\mathbb{S}$.

Definition 4 ([16])**.** *Let* $\mathbf{K}$ *be a belief set.*

(a) An operation $+$ *on* $\mathbf{K}$ *is a system of spheres-based expansion operator if and only if there exists a system of spheres* $\mathbb{S}$ *centered on* $\|\mathbf{K}\|$ *such that for all* α *it holds that:*

$$\mathbf{K}+\alpha = \bigcap(\|\mathbf{K}\| \cap \|\alpha\|).$$

(b) An operation $\div$ *on* $\mathbf{K}$ *is a system of spheres-based contraction operator if and only if there exists a system of spheres* $\mathbb{S}$ *centered on* $\|\mathbf{K}\|$ *such that for all* α *it holds that:*

$$\mathbf{K} \div \alpha = \begin{cases} \bigcap((S_{\neg\alpha} \cap \|\neg\alpha\|) \cup \|\mathbf{K}\|) & \text{if } \|\neg\alpha\| \neq \emptyset \\ \mathbf{K} & \text{otherwise} \end{cases}$$

(c) An operation $\star$ *on* $\mathbf{K}$ *is a system of spheres-based revision operator if and only if there exists a system of spheres* $\mathbb{S}$ *centered on* $\|\mathbf{K}\|$ *such that for all* α *it holds that:*

$$\mathbf{K}\star\alpha = \begin{cases} \bigcap(S_\alpha \cap \|\alpha\|) & \text{if } \|\alpha\| \neq \emptyset \\ \mathscr{L} & \text{otherwise} \end{cases}$$

It holds that sphere-based revision and contraction operators are characterized, by the (eight) AGM postulates for revision and contraction, respectively ([16]).

3 Two Level Credibility-limited Revisions

The *two level CL revisions* are operators of non-prioritized revision. When revising a belief set by a sentence α, we first need to analyse the degree of credibility of that sentence. When revising by a sentence that is considered to be at the highest level of credibility, the operator works as a standard revision operator. If it is considered to be at the second level of credibility, then that sentence is not incorporated in the revision process but its negation is removed from the original belief set. When revising by a non-credible sentence, the operator leaves the original belief set unchanged. The following definition formalizes this concept:

Definition 5 ([8, 3]). *Let* $\mathbf{K}$ *be a belief set,* $\star$ *be a basic AGM revision operator on* $\mathbf{K}$ *and* C_H *and* C_L *be subsets of* $\mathscr{L}$. *Then* $\odot$ *is a two level CL revision operator induced by* $\star$, C_H *and* C_L *if and only if:*

$$\mathbf{K}\odot\alpha = \begin{cases} \mathbf{K}\star\alpha & \textit{if } \alpha\in C_H \\ (\mathbf{K}\star\alpha)\cap\mathbf{K} & \textit{if } \alpha\in C_L \\ \mathbf{K} & \textit{if } \alpha\notin(C_L\cup C_H) \end{cases}$$

In the previous definition $C_H\cup C_L$ represent the sentences that are considered to have some degree of credibility. C_H and C_L represent respectively the set of sentences that are considered to be at the first (highest) and at the second level of credibility. Note that if $\alpha\in C_L$, then $\mathbf{K}\odot\alpha=(\mathbf{K}\star\alpha)\cap\mathbf{K}$. According to the Harper identity $(\mathbf{K}\star\alpha)\cap\mathbf{K}$ coincides with the contraction of $\mathbf{K}$ by $\neg\alpha$.

This construction can be further specified by adding constraints to the structure of the set(s) of credible sentences. In [19, 9], the following properties for a given set of credible sentences C were proposed:
Credibility of Logical Equivalents: If $\vdash\alpha\leftrightarrow\beta$, then $\alpha\in C$ if and only if $\beta\in C$.[3]
Single Sentence Closure: If $\alpha\in C$, then $Cn(\alpha)\subseteq C$.
Element Consistency: If $\alpha\in C$, then $\alpha\nvdash\bot$.
Credibility lower bounding: If $\mathbf{K}$ is consistent, then $\mathbf{K}\subseteq C$.

Additionally, in [8] the following condition that relates a set of credible sentences C with a revision function $\star$ was introduced. This condition, designated by *condition* (**C - $\star$**), states that if a sentence α is not credible, then any possible outcome of revising the belief set $\mathbf{K}$ through $\star$ by a credible sentence contains $\neg\alpha$. The intuition underlying this property is that if α is not credible then its negation cannot be removed. Thus its negation should still be in the outcome of the revision by any credible sentence.

If $\alpha\notin C$ and $\beta\in C$, then $\neg\alpha\in\mathbf{K}\star\beta$. (**C - $\star$**)

3.1 Two level credibility-limited revision postulates

We now recall from [8] some of the postulates proposed to express properties of the two level CL revision operators. The first postulate was originally proposed in [24], the second in [21], the following three in [19] and the remaining ones in [8].
(Consistency Preservation) If $\mathbf{K}$ is consistent, then $\mathbf{K}\odot\alpha$ is consistent.
(Confirmation) If $\alpha\in\mathbf{K}$, then $\mathbf{K}\odot\alpha=\mathbf{K}$.
(Strict Improvement) If $\alpha\in\mathbf{K}\odot\alpha$ and $\vdash\alpha\rightarrow\beta$, then $\beta\in\mathbf{K}\odot\beta$.
(Regularity) If $\beta\in\mathbf{K}\odot\alpha$, then $\beta\in\mathbf{K}\odot\beta$.
(Disjunctive Distribution) If $\alpha\vee\beta\in\mathbf{K}\odot(\alpha\vee\beta)$, then $\alpha\in\mathbf{K}\odot\alpha$ or $\beta\in\mathbf{K}\odot\beta$.
(N-Recovery) $\mathbf{K}\subseteq\mathbf{K}\odot\alpha+\neg\alpha$.
(N-Relative success) If $\neg\alpha\in\mathbf{K}\odot\alpha$, then $\mathbf{K}\odot\alpha=\mathbf{K}$.
(N-Persistence) If $\neg\beta\in\mathbf{K}\odot\beta$, then $\neg\beta\in\mathbf{K}\odot\alpha$.
(N-Success Propagation) If $\neg\alpha\in\mathbf{K}\odot\alpha$ and $\vdash\beta\rightarrow\alpha$, then $\neg\beta\in\mathbf{K}\odot\beta$.
(Weak Relative Success) $\alpha\in\mathbf{K}\odot\alpha$ or $\mathbf{K}\odot\alpha\subseteq\mathbf{K}$.
(Weak Vacuity) If $\neg\alpha\notin\mathbf{K}$, then $\mathbf{K}\subseteq\mathbf{K}\odot\alpha$.
(Weak Disjunctive Inclusion) If $\neg\alpha\notin\mathbf{K}\odot(\alpha\vee\beta)$, then $\mathbf{K}\odot(\alpha\vee\beta)+(\alpha\vee\beta)\subseteq\mathbf{K}\odot\alpha+\alpha$.
(Containment) If $\mathbf{K}$ is consistent, then $\mathbf{K}\cap((\mathbf{K}\odot\alpha)+\alpha)\subseteq\mathbf{K}\odot\alpha$.

The following observations relate some of the postulates presented above.

[3] In [19] this property was designated by *closure under logical equivalence* and was formulated as follows: If $\vdash\alpha\leftrightarrow\beta$, and $\alpha\in C$, then $\beta\in C$.

Observation 1 ([8])**.** *Let* $\mathbf{K}$ *be a consistent and logically closed set and* $\odot$ *be an operator on* $\mathbf{K}$*.*
(a) If $\odot$ *satisfies closure, consistency preservation, weak relative success and N-Recovery, then it satisfies N-Relative success.*
(b) If $\odot$ *satisfies weak vacuity and inclusion, then it satisfies confirmation.*

Observation 2. *Let* $\mathbf{K}$ *be a consistent and logically closed set and* $\odot$ *be an operator on* $\mathbf{K}$*.*
(a) If $\odot$ *satisfies consistency preservation, closure, vacuity, inclusion, strict improvement, disjunctive inclusion, disjunctive overlap and N-recovery, then it satisfies regularity.*
(b) If $\odot$ *satisfies consistency preservation, closure, vacuity, weak relative success and disjunctive inclusion, then it satisfies disjunctive distribution.*
(c) If $\odot$ *satisfies N-recovery and closure, then it satisfies containment.*

In the following theorem we recall from [8] an axiomatic characterization for a two level CL revision operator induced by an AGM revision and sets C_H and C_L satisfying some given properties.[4]

Observation 3 ([8])**.** *Let* $\mathbf{K}$ *be a consistent and logically closed set and* $\odot$ *be an operator on* $\mathbf{K}$*. Then the following conditions are equivalent:*

1. $\odot$ *satisfies weak relative success, closure, inclusion, consistency preservation, weak vacuity, extensionality, strict improvement, N-persistence, N-recovery, disjunctive overlap and weak disjunctive inclusion.*

2. $\odot$ *is a two level CL revision operator induced by an AGM revision operator* $\star$ *for* $\mathbf{K}$ *and sets* $C_H, C_L \subseteq \mathscr{L}$ *such that:* C_L *satisfy credibility of logical equivalents and element consistency,* $C_H \cap C_L = \emptyset$*,* C_H *satisfies element consistency, credibility lower bounding and single sentence closure and condition* $(C_H \cup C_L$ - $\star)$ *holds.*

4 System of Spheres-based Two Level Credibility-limited Revisions

In this section we present the definition of a system of spheres-based two level CL revision operator. We start by presenting the notion of two level system of spheres, centred on $\|\mathbf{K}\|$.

Definition 6. *Let* $\mathbf{K}$ *be a belief set. A two level system of spheres centred on* $\|\mathbf{K}\|$ *is a pair* $(\mathbb{S}_i, \mathbb{S})$ *whose elements are subsets of* $\mathscr{M}_{\mathscr{L}}$*, i.e.,* $\mathbb{S} \subseteq \mathscr{P}(\mathscr{M}_{\mathscr{L}})$ *and* $\mathbb{S}_i \subseteq \mathscr{P}(\mathscr{M}_{\mathscr{L}})$*, such that:*
(a) $\mathbb{S}$ *and* $\mathbb{S}_i$ *satisfy conditions* $(\mathbb{S}1)$*,* $(\mathbb{S}2)$ *and* $(\mathbb{S}4)$ *of Definition 3;*
(b) $\mathbb{S}_i \subseteq \mathbb{S}$*;*
(c) If $X \in \mathbb{S}_i$*, then* $X \subseteq Y$ *for all* $Y \in \mathbb{S} \setminus \mathbb{S}_i$*.*

Intuitively, a two level system of spheres $(\mathbb{S}_i, \mathbb{S})$, centered on $\|\mathbf{K}\|$ is a system composed by two systems of spheres $\mathbb{S}_i$ and $\mathbb{S}$, both centered on $\|\mathbf{K}\|$, where $\mathbb{S}_i \subseteq \mathbb{S}$ and in which the condition $(\mathbb{S}3)$ of Definition 3 is relaxed for $\mathbb{S}_i$ and $\mathbb{S}$, allowing the existence of possible worlds outside the union of all spheres of $\mathbb{S}_i$ and of $\mathbb{S}$.[5] Conditions (b) and (c) impose that the spheres of $\mathbb{S}_i$ are the innermost ones (see Figure 1).

The following observation is a direct consequence of condition (c). It states that all spheres contained in a given sphere of $\mathbb{S}_i$ belong to $\mathbb{S}_i$.

Observation 4. *If* $\mathbb{S}_i$ *and* $\mathbb{S}$ *satisfy condition (c) of Definition 6, then it holds that:*
If $X \in \mathbb{S}$ *and* $Y \in \mathbb{S}_i$ *are such that* $X \subseteq Y$*, then* $X \in \mathbb{S}_i$*.*

[4] Actually, the containment postulate was also included in the list of postulates of the representation theorem presented in [8], however as Observation 2 illustrates, containment follows from closure and N-recovery.

[5] Condition $(\mathbb{S}3)$ of Definition 3 was also relaxed in [19] when constructing a (modified) system of spheres for credibility-limited revision operators.

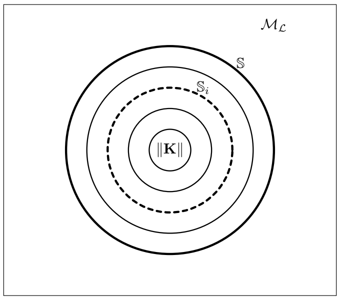


Figure 1: Schematic representation of a two level system of spheres $(\mathbb{S}_i, \mathbb{S})$, centred on $\|\mathbf{K}\|$. The dashed circle establishes the boundary between the spheres of $\mathbb{S}_i$ and those of $\mathbb{S} \setminus \mathbb{S}_i$. Worlds outside the thickest line are not elements of any sphere of $\mathbb{S}$.

In a system of spheres centered on $\|\mathbf{K}\|$, the worlds considered most plausible are those that lie in the innermost sphere (i.e. in $\|\mathbf{K}\|$), and the closer a possible world is to the center, the more plausible it is considered to be. Similarly, the worlds lying in the spheres of $\mathbb{S}_i$ have a higher degree of plausibility than those in the spheres of $\mathbb{S} \setminus \mathbb{S}_i$. Intuitively, a two level system of spheres $(\mathbb{S}_i, \mathbb{S})$, centered on $\|\mathbf{K}\|$ defines three clusters. The first cluster is formed by the worlds in the spheres of $\mathbb{S}_i$. These worlds are the ones to which a higher degree of plausibility is assigned (relatively to those outside the spheres of $\mathbb{S}_i$). The second cluster is formed by the worlds in the spheres of $\mathbb{S} \setminus \mathbb{S}_i$, which are assigned some (lower) degree of plausibility. Finally, the third cluster is formed by the worlds outside the spheres of $\mathbb{S}$, which are considered to be not plausible.

We are now in conditions to present the definition of a system of spheres-based two level CL revision operator. The outcome of the revision by means of a system of spheres-based two level CL revision operator of a belief set $\mathbf{K}$ by a sentence α (see Figure 2) is:
- the intersection of the most plausible α-worlds, if these are α-worlds in the cluster of the most plausible worlds.[6]
- the intersection of all the worlds contained in the union of the set of $\mathbf{K}$-worlds with the set of the most plausible α-worlds, if the α-worlds are considered to be plausible, but are not in the cluster of the most plausible ones.
- $\mathbf{K}$ if the α-worlds are not plausible, i.e, in this case the belief set remains unchanged.

Definition 7. *Let $\mathbf{K}$ be a belief set and $(\mathbb{S}_i, \mathbb{S})$ be a two level system of spheres centered on $\|\mathbf{K}\|$. The system of spheres-based two level CL revision operator induced by $(\mathbb{S}_i, \mathbb{S})$ is the operator $\odot_{(\mathbb{S}_i, \mathbb{S})}$ such that, for all α:*

$$\mathbf{K} \odot_{(\mathbb{S}_i,\mathbb{S})} \alpha = \begin{cases} \bigcap(S_\alpha \cap \|\alpha\|) & \textit{if } S_\alpha \in \mathbb{S}_i \\ \bigcap(\|\mathbf{K}\| \cup (S_\alpha \cap \|\alpha\|)) & \textit{if } S_\alpha \in \mathbb{S} \setminus \mathbb{S}_i \\ \mathbf{K} & \textit{if } X \cap \|\alpha\| = \emptyset, \textit{for all } X \in \mathbb{S} \end{cases}$$

An operator $\odot$ on $\mathbf{K}$ is a system of spheres-based two level CL revision operator if and only if there exists a two levels system of spheres $(\mathbb{S}_i, \mathbb{S})$ centred on $\|\mathbf{K}\|$ such that $\mathbf{K} \odot \alpha = \mathbf{K} \odot_{(\mathbb{S}_i,\mathbb{S})} \alpha$ holds for all α.

[6] Note that being X a set of possible worlds $\bigcap X$ is a belief set.

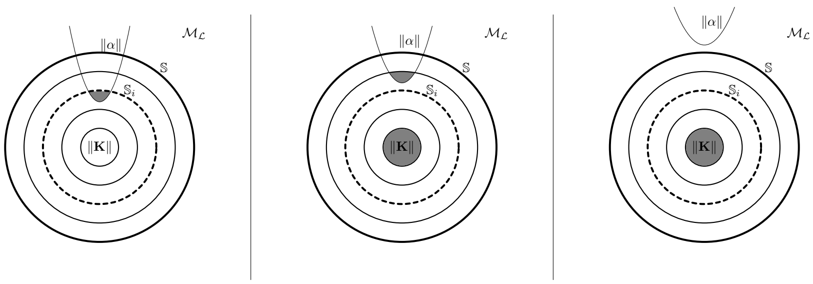


Figure 2: Schematic representation of the worlds of the outcome of the system of spheres-based two level CL revision operator induced by a two level system of spheres $(\mathbb{S}_i, \mathbb{S})$ centred on $\|\mathbf{K}\|$ by a sentence α. In the first case, it holds that $S_\alpha \in \mathbb{S}_i$ and in the second that $S_\alpha \in \mathbb{S} \setminus \mathbb{S}_i$. In the third case, all the α-worlds are outside the spheres of $\mathbb{S}$.

4.1 Representation theorems

We now present a representation theorem for system of spheres-based two level CL revision operators. It also relates these operators with the two level CL revision operators induced by AGM revision operators and sets $C_H, C_L \subseteq \mathscr{L}$ satisfying some given properties. Considering the axiomatic characterization for the latter, presented in Observation 3, we note that we only need to ensure that the Condition (C_H - $\star$) holds, to guarantee that the class of these operators coincides with the class of system of spheres-based two level CL revision operators.

Theorem 1. *Let* $\mathbf{K}$ *be a consistent and logically closed set and* $\odot$ *be an operator on* $\mathbf{K}$*. Then the following conditions are equivalent:*

1. $\odot$ *satisfies weak relative success, closure, inclusion, consistency preservation, vacuity, extensionality, strict improvement, N-persistence, N-recovery, disjunctive overlap and disjunctive inclusion.*

2. $\odot$ *is a system of spheres-based two level CL revision operator.*

3. $\odot$ *is a two level CL revision operator induced by an AGM revision operator* $\star$ *for* $\mathbf{K}$ *and sets* $C_H, C_L \subseteq \mathscr{L}$ *such that:* C_L *satisfy credibility of logical equivalents and element consistency,* $C_H \cap C_L = \emptyset$*,* C_H *satisfies element consistency, credibility lower bounding and single sentence closure and conditions* ($C_H \cup C_L$ - $\star$) *and* (C_H - $\star$) *hold.*

5 Related Works

In this section we will mention other approaches related with the present paper.

- In [8], the two level CL revision operators were defined in terms of a basic AGM revision operator and sets C_H and C_L of credible sentences. Several properties have been proposed for these sets. Postulates to characterize two level CL revision operators were proposed. Results exposing the relation between the postulates and the properties of C_H and C_L were presented. Axiomatic characterizations for several classes of two level CL revision operators were presented (namely for two level CL revision operators induced by basic AGM revisions and by AGM revisions in which the associated sets of credible sentences satisfy certain properties).

- In [3], the operators of two CL revision were introduced in terms of basic AGM belief revisions operators (in that paper these operators are designated by *Filtered belief revision*). The possibility that an item of information could still be "taken" seriously, even if it is not accepted as being fully credible (this type

of information is there called *allowable*) was discussed. A syntactic analysis of filtered belief revision was provided.
- In [4], the works presented in [3] and [8] were extended by introducing the notion of partial belief revision structure, providing a characterization of filtered belief revision in terms of properties of these structures. There it is considered the notion of rationalizability of a choice structure in terms of a plausibility order and established a correspondence between rationalizability and AGM consistency in terms of the eight AGM postulates for revision. An interpretation of credibility, allowability and rejection of information in terms of the degree of implausibility of the information was provided.
- In [19] credibility-limited revision operators were presented. When revising a belief set by a sentence by means of a credibility-limited revision operator, we need first to analyse whether that sentence is credible or not. When revising by a credible sentence, the operator works as a basic AGM revision operator, otherwise it leaves the original belief set unchanged. Two level credibility-limited revisions operators can be seen as a generalization of credibility-limited revision operators. In fact, in the case that $C_L = \emptyset$ both types of operators coincide. In [19] several properties were prosed for C (the set of credible sentences) and this model was developed in terms of possible world models. Representations theorems for different classes of Credibility-limited revisions operators were presented. The extension of credibility-limited revision operators to the belief bases setting was studied in [7, 9, 10, 11].

6 Conclusion

The model of credibility-limited revision ([19]) is essentially a generalization of the AGM framework ([1]) of belief revision, which addresses one of the main shortcomings pointed out to that framework, namely the fact that it assumes that any new information has priority over the original beliefs. In the model of credibility-limited revisions two classes of sentences are considered. Some sentences —the so-called *credible sentences*— are accepted in the process of revision by them, while the remaining sentences are such that the process of revising by them has no effect at all in the original belief set.

In its turn, the model of two level CL revision ([3, 8]) generalizes credibility-limited revision by considering an additional class of sentences. A sentence of this class is such that, although a revision by it does not lead to its acceptance, it causes the removal of its negation from the original belief set.

The present paper offers a semantic approach to the two level CL revision operators. More precisely, it introduces a class of two-level CL revision operators whose definition is based on a structure called two level system of spheres, which generalizes the well-known systems of spheres proposed by Grove ([16]). This semantic definition provides some additional insight on the intuition that underlays the notion of two-level CL revisions.

Acknowledgements This paper was partially supported by FCT-Fundação para a Ciência e a Tecnologia, Portugal through project PTDC/CCI-COM/4464/2020. M.G. and M.R. were partially supported by the Centro de Investigação em Matemática e Aplicações (CIMA), through the grant UIDB/04674/2020 of FCT. E.F. was partially supported by FCT through project UIDB/04516/2020 (NOVA LINCS).

References

[1] Carlos Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985. doi:10.1007/BF00370430

[2] Carlos Alchourrón and David Makinson. On the logic of theory change: Safe contraction. *Studia Logica*, 44:405–422, 1985. doi:10.1007/BF00370430

[3] Giacomo Bonanno. Credible information, allowable information and belief revision - extended abstract. In Lawrence S. Moss, editor, *Proceedings Seventeenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2019, Toulouse, France, 17-19 July 2019*, volume 297 of *EPTCS*, pages 82–90, 2019. doi:10.4204/eptcs.297.6

[4] Giacomo Bonanno. Filtered belief revision: Syntax and semantics. *Journal of Logic, Language and Information*, 31:645–675, 2022. doi:10.1007/s10849-022-09374-x

[5] Eduardo Fermé and Sven Ove Hansson. AGM 25 years: Twenty-five years of research in belief change. *Journal of Philosophical Logic*, 40:295–331, 2011. doi:10.1007/s10992-011-9171-9

[6] Eduardo Fermé and Sven Ove Hansson. *Belief Change: Introduction and Overview*. Springer Briefs in Computer Science Series. Springer, 2018. doi:10.1007/978-3-319-60535-7

[7] Eduardo Fermé, Juan Mikalef, and Jorge Taboada. Credibility-limited functions for belief bases. *Journal of Logic and Computation*, 13:1:99–110, 2003. doi:10.1093/logcom/13.1.99

[8] Marco Garapa. Two level credibility-limited revisions. *The Review of Symbolic Logic*, 15(2):388–408, 2022. doi:10.1017/S1755020320000283

[9] Marco Garapa, Eduardo Fermé, and Maurício Reis. Studies in credibility-limited base revision. In *Proceedings of the Sixteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2018)*, pages 240–247, 2018. https://aaai.org/papers/28-studies-in-credibility-limited-base-revision/

[10] Marco Garapa, Eduardo Fermé, and Maurício D. L. Reis. Levi and Harper identities for non-prioritized belief base change. *Artificial Intelligence, 2023*. doi:10.1016/j.artint.2023.103907

[11] Marco Garapa, Eduardo Fermé, and Maurício D.L. Reis. Credibility-limited base revision: New classes and their characterizations. *Journal of Artificial Intelligence Research*, 69:1023 – 1075, 2020. doi:10.1613/jair.1.12298

[12] Peter Gärdenfors. Conditionals and changes of belief. *Acta Philosophica Fennica*, 30:381–404, 1978.

[13] Peter Gärdenfors. Rules for rational changes of belief. In Tom Pauli, editor, *Philosophical Essays dedicated to Lennart Åqvist on his fiftieth birthday*, number 34 in Philosophical Studies, pages 88–101, 1982.

[14] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, 1988.

[15] Peter Gärdenfors and David Makinson. Revisions of knowledge systems using epistemic entrenchment. In Moshe Y. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 83–95, Los Altos, 1988. Morgan Kaufmann. http://www.tark.org/proceedings/tark_mar7_88/p83-gardenfors.pdf

[16] Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988. doi:10.1007/BF00247909

[17] Sven Ove Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50:413–427, 1999. doi:10.1023/A:1005534223776

[18] Sven Ove Hansson. *A Textbook of Belief Dynamics. Theory Change and Database Updating.* Applied Logic Series. Kluwer Academic Publishers, Dordrecht, 1999.

[19] Sven Ove Hansson, Eduardo Fermé, John Cantwell, and Marcelo Falappa. Credibility-limited revision. *Journal of Symbolic Logic*, 66(4):1581–1596, 2001. doi:10.2307/2694963

[20] William L. Harper. Rational conceptual change. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1976:462–494, 1976. doi:10.1086/psaprocbienmeetp.1976.2.192397

[21] Hirofumi Katsuno and Alberto Mendelzon. Propositional knowledge base revision and minimal change. *Journal of Artificial Intelligence*, 52:263–294, 1991. doi:10.1016/0004-3702(91)90069-V

[22] Isaac Levi. Subjunctives, dispositions, and chances. *Synthèse*, 34:423–455, 1977. doi:10.1007/BF00485649

[23] David Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[24] David Makinson. Screened revision. *Theoria*, 63:14–23, 1997. doi:10.1111/j.1755-2567.1997.tb00737.x

[25] Hans Rott and Sven Ove Hansson. Safe contraction revisited. In Sven Ove Hansson, editor, *David Makinson on Classical Methods for Non-Classical Problems*, volume 3 of *Outstanding Contributions to Logic*, pages 35–70. Springer Netherlands, 2014. doi:10.1007/978-94-007-7759-0_4

7 Appendix

In this appendix we provide a sketch proof for the main result presented in this paper.

Proof sketch of Theorem 1:

(2) to (1):

Let $\odot$ be a system of spheres-based two level credibility limited revision operator induced by a two levels system of spheres $(\mathbb{S}_i, \mathbb{S})$. We need to prove that $\odot$ satisfies all the postulates present in statement (1) .

(1) to (2):

Assume that $\odot$ satisfies all the postulates listed in statement (1) and consider the following constructions for $\mathbb{S}$ and $\mathbb{S}_i$:

$S \in \mathbb{S}_i$ iff:

(a) $S = \|\mathbf{K}\|$;

(b) $\emptyset \neq S \subseteq \{w : w \in \|\mathbf{K} \odot \alpha\|$, for some α such that $\|\mathbf{K} \odot \alpha\| \subseteq \|\alpha\|\}$ and $\|\mathbf{K} \odot \alpha\| \subseteq S$ for all α such that $S \cap \|\alpha\| \neq \emptyset$.

$S \in \mathbb{S}$ iff:

(a) $S = \|\mathbf{K}\|$;

(b) $\emptyset \neq S \subseteq \{w : w \in \|\mathbf{K} \odot \alpha\|$, for some α such that $\|\mathbf{K} \odot \alpha\| \cap \|\alpha\| \neq \emptyset\}$, $\|\mathbf{K} \odot \alpha\| \subseteq S$ for all α such that $S \cap \|\alpha\| \neq \emptyset$ and if $S \cap \|\alpha\| = \emptyset$ and $S \notin \mathbb{S}_i$, then $\|\mathbf{K} \odot \alpha\| \cap S = \|\mathbf{K}\|$.

We need to show that:

1. $(\mathbb{S}_i, \mathbb{S})$ is a two level system of spheres centred on $\|\mathbf{K}\|$. To do so, it is necessary to prove that:
   i. $\mathbb{S}$ and $\mathbb{S}_i$ satisfy conditions $(\mathbb{S}1)$, $(\mathbb{S}2)$ and $(\mathbb{S}4)$, of Definition 3;
   ii. $\mathbb{S}_i \subseteq \mathbb{S}$;
   iii. If $X \in \mathbb{S}_i$, then $X \subseteq Y$ for all $Y \in \mathbb{S} \setminus \mathbb{S}_i$.
2. If $\|\alpha\| = \emptyset$, then $\mathbf{K} \odot \alpha = \mathbf{K}$;
3. For α such that $\mathbf{K} \odot \alpha \nvdash \neg\alpha$ and $S(\alpha) = \bigcup\{\|\mathbf{K} \odot \delta\| : \|\alpha\| \subseteq \|\delta\|\}$, it holds that:
   i. $S(\alpha) \in \mathbb{S}$
   ii. $S(\alpha) = S_\alpha$ (i.e. $S(\alpha)$ is the minimal sphere that intersects with $\|\alpha\|$).
   iii.
$$\mathbf{K} \odot \alpha = \begin{cases} \bigcap(S_\alpha \cap \|\alpha\|) & \text{if } S_\alpha \in \mathbb{S}_i \\ \bigcap(\|\mathbf{K}\| \cup (S_\alpha \cap \|\alpha\|)) & \text{if } S_\alpha \in \mathbb{S} \setminus \mathbb{S}_i \\ \mathbf{K} & \text{if } X \cap \|\alpha\| = \emptyset, \text{for all } X \in \mathbb{S} \end{cases},$$

   where $S_\alpha = S(\alpha)$.

(1) to (3):

Let $\odot$ be an operator satisfying the postulates listed in statement (1). Let $\star$ be the operation such that:

i. If $\neg\alpha \notin \mathbf{K} \odot \alpha$, then $\mathbf{K} \star \alpha = \mathbf{K} \odot \alpha + \alpha$;

ii. If $\neg\alpha \in \mathbf{K} \odot \alpha$, then $\mathbf{K} \star \alpha = Cn(\alpha)$.

Furthermore let $C_H = \{\alpha : \alpha \in \mathbf{K} \odot \alpha\}$ and $C_L = \{\alpha : \neg\alpha \notin \mathbf{K} \odot \alpha\} \setminus C_H$.
These are the same construction that were used in the corresponding part of Observation 3. Then, regarding this proof, it remains only to show that condition (**C_H - $\star$**) holds.

(3) to (1):
By Observation 3 it only remains to prove that $\odot$ satisfies vacuity and disjunctive inclusion. □

A “Game of Like” : Online Social Network Sharing As Strategic Interaction

Emmanuel J. Genot*
Department of Philosophy
Lund University
Lund, Sweden
emmanuel.genot@fil.lu.se

We argue that behavioral science models of online content-sharing overlook the role of strategic interactions between users. Borrowing from accuracy-nudges studies decision-theoretic models, we propose a basic game model and explore special cases with idealized parameter settings to identify refinements necessary to capture real-world online social network behavior. Anticipating those refinements, we sketch a strategic analysis of content amplification and draw a connection between Keynes’ “beauty contest” analogy and recent social-epistemological work on echo chambers. We conclude on the model’s prospects from analytical and empirical perspectives.

1 Motivations

Online search engines garnered attention from social epistemologists in the early days of the commercial Internet, when A. Goldman analyzed them as retrieval systems in [5]. Later, T.A. Simpson extended Goldman’s analysis into a model of surrogate expertise in [21] in direct response to Google Search personalization algorithms. Recently, epistemologists have turned to online social networks (hereafter OSN), fulfilling a similar function of online information sources, with even greater personalization. Notably, T.C. Nguyen [12] provided a much-needed conceptual analysis of OSN epistemic bubbles and echo chambers, and C. O’Connor and J.O. Weatherall [14] proposed that applying network epistemology to OSN could address limitations of contagion models of online information spread. At the same time, behavioral scientists have independently addressed the limitations of contagion models by looking at OSN-sharing through a rational choice lens. Particularly, studies that shaped the field and its public perception have manifested a Bayesian influence. Widely publicized studies like [23] (a *Science* cover story: “How lies spread–On social media, fake news beats the truth”) and [16] (a *Nature* cover story: “Misinformation–A prompt to check accuracy cuts online sharing of fake news”) appealed to Bayesian decision theory and expected utility theory to rationalize OSN content-sharing and interpret diffusion-model data analyses.[1]

*I wish to thank Erik Mohlin, Jens Ulrik Hansen, Justine Jacot, and Patricia Rich, for their invaluable help at the various stages of this paper’s development; three anonymous referees, whose comments and suggestions brought about some major changes and (hopefully) improvements; and Rineke Verbrugge, who reviewed those changes, and suggested further improvements. Any mistakes left are on me.

[1] “[U]ser characteristics and network structure could not explain the differential diffusion of truth and falsity, we sought alternative explanations for the differences in their diffusion dynamics. One alternative explanation emerges from information theory and Bayesian decision theory. Novelty attracts human attention, contributes to productive decision-making, and encourages information sharing because novelty updates our understanding of the world.” [23, p. 1149]. Similarly, “people do care more about accuracy than other content dimensions but accuracy nonetheless often has little effect on sharing, because (ii) the social media context focuses [users’] attention on other factors such as the desire to attract and please followers/friends or to signal one’s group membership. In the language of utility theory, an ‘attentional spotlight’ is shone upon certain terms in the decider’s utility function, such that only those terms are weighed when making a decision” [16, p. 591]. The framework of [16]

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 299–307, doi:10.4204/EPTCS.379.24

Decision theory best models decisions under uncertainty about the state of nature, but OSN-sharing outcomes depend on reactions from a community of users. The preferred model for decisions *under uncertainty about other agents' decisions* is game theory, and while the formalisms are inter-translatable, decision theory is less expressive. As pointed out by J. Harsanyi, the game-to-decision direction loses in translation the explicit expression of mutual expectations of rationality (via solution concepts [6, 7]). Compounding this issue, decision-theoretic models from behaviral science studies (such as [23, 16]) were not proposed as translations for games and thus did not explicitly translate mutual expectations into constraints on decision-makers priors (as per the games-to-decision direction, cf. [8, 6]), leaving their role almost entirely unanalyzed. Unfortunately, social epistemology offers no ready-made solution. Nguyen's analysis is strategic but informal and cannot bear on the data without a formal reconstruction, while network epistemology does not address strategic expectations formally.

The absence of a strategic analysis of OSN-sharing motivated the approach presented in the remainder of this paper. Section 2 builds upon behavioral science decision-theoretic models to propose a simplified game model for OSN-sharing, differentiating between content-based and engagement-based preferences. Section 3 examines special cases that highlight the model's salient features and limitations and identifies extensions necessary to reconstruct real-life OSN users' behavior. Section 4 extrapolates informally and proposes that special cases of OSN-sharing elicit strategy selection akin to reasoning in guessing games and could illuminate content amplification scenarios, including Nguyen's epistemic bubbles and echo chambers. We conclude with the analytic prospects of a strategic re-interpretation of extant data, and a suggestion for the design of new studies.

2 A Game of Like

Behavioral science studies of OSN-sharing often acknowledge the role of strategic interactions between users but have so far fallen short of factoring in their contribution. Pennycook *et al.* (2021) is a paradigmatic example: the authors note that "the desire to attract and please followers/friends or to signal one's group membership" [16, p. 591] contributes to content-sharing decisions, but propose a utility function limited to personal preferences for content having such-and-such characteristics.[2] A natural first step toward a strategic model is thus to introduce the missing terms, then specify a game based on this completed picture. For simplicity, we can let $u_{p_i}(\cdot)$ denote i's *personal utility*, expressing how some content aligns with i's personal preferences for content having such-and-such characteristics, with the understanding that this alignment could be further analyzed along multiple dimensions (as in [16], cf. n. 2). To that, we add a term that we denote $u_{s_i}(\cdot)$, for the *social utility*, expressing how reactions to the content shared—'likes,' re-shares, comments, etc.—satisfy i's preferences for social validation or, more generally, engagement. Finally, we introduce a parameter, that we denote γ, to represent the relative weights of i's personal preferences for content and social preferences for engagement. In the decision-theoretic model of [16], the only action being 'sharing,' actions and content are indiscernible, and the

is implicitly decision-theoretic, as utilities take as argument proxies for individual choices (content shared) rather than strategic profiles (cf. Section 2).

[2] "Consider a piece of content x which is defined by k different characteristic dimensions [including] whether the content is false/misleading $F(x)$, and other $k-1$ dimensions [that] are non-accuracy-related (e.g. partisan alignment, humorousness, etc) defined as $C_2(x)\dots C_k(x)$. In our model, the utility [...] from sharing content x is given by: $U(x) = -\alpha_1\beta_F F(x) + \sum_{i=2}^{k} \alpha_i\beta_i C_i(x)$ where β_F indicates how much they dislike sharing misleading content and $\beta_2 \dots \beta_k$ indicate how much they care about each of the other dimensions (i.e. βs indicate preferences); while α_1 indicates how much the person is paying attention to accuracy, and $\alpha_2 \dots \alpha_k$ indicate how much the person is paying attention to each of the other dimensions." [16, Supplementary Information: S9-10]

utility function can range over the content. In a game-theoretic model, the utility function ranges over *strategy profiles*, and we must distinguish content from actions.

As a basic model, we consider an n-player repeated game G in strategic form with a set P of players, where any $i \in P$ can, at each round, 'like' or 'share' content. As simplifications, we assume that players only share new content at round $r = 0$, so any 'share' action at round $r \geq 1$ is a 'reshare.' Under this simplification, we can specify a set of (original) *content* $C_G = \{c_1, \ldots, c_n\}$ for G, where c_i is the content introduced at $r = 0$ by agent i. The *action set* for some player i at some round r is $A_i = \{\mathsf{like}_i(x,y), \mathsf{share}_i(x,y) : x \in C_G, y \in P\}$, where y is a player who shared x at some round $r' < r$, and from whom i is re-sharing x. Note that, under our simplifying assumption, at $r = 0$, there is nothing to 'like.' If all actions are visible to all players, no restriction is imposed on x or y. Explicitly: any content shared by some player at round r can be reshared by any other player at round $(r+1)$. This amounts to a game with perfect information, adequate for demonstrating the strategic standpoint's fruitfulness but insufficient to model real-world OSNs (see below). Our earlier discussion of personal and social preferences yields a utility function, as below.

$$U_i(\cdot) \quad = \quad \gamma_i u_{p_i}(\cdot) + (1-\gamma_i) u_{s_i}(\cdot) \tag{1}$$

Intuitively, in the decision-theoretic approach, $u_{p_i}(c)$ expresses i's preferences as a function of the distance between c and i's ideal content located in a multi-dimensional space whose dimensions correspond to i's criteria of evaluation. In a round of G, the argument of (1) is a strategy profile $\overline{\sigma} = (\sigma_1, \ldots \sigma_n)$ where σ_i is player i's strategy at that round. Following the same intuition, $u_{p_i}(\overline{\sigma})$ can be thought of as a function of the relative distances between i's ideal content and the 'community content' $c_1, \ldots, c_n$, or some weighed sum thereof, representing how close C_G is to i's ideal.[3] So construed, and under our simplification, $u_{p_i}(\overline{\sigma})$ remains constant after $r = 0$. Again, this personal preference model is sufficient for our purposes. Still, in a real-world OSN, overall engagement could indirectly affect $u_{p_i}(\overline{\sigma})$ (i may care for overall visibility, and in a model with incomplete information, visibility would depend on engagement, see below).

In a decision-theoretic model (e.g., extending that of [16]) $u_{s_i}(c)$ would be a function of the (accumulated) engagement from users other than i, with c (when shared by i). In G, $u_{s_i}(\overline{\sigma})$ at round r is, in part, a function of how other players have engaged in r with the content i shared at some $r' > r$; and in part, of the accrued social utility inherited from earlier rounds. The candidate functions for computing either component are too numerous to review here, and which one applies to particular cases may be empirically constrained by algorithms. Still, it suffices for our purposes to note that, at some round r, $u_{s_i}(\overline{\sigma})$ does not 'reset' i's social utility; that the contribution of 'likes' and (re)shares may vary; and that evaluations may depend on players' knowledge.[4] For definiteness, we can assume a function $u_{s_i}(\overline{\sigma})$ that ranks higher strategy profiles where content i shared (or reshared) is both liked and reshared rather than liked or shared (alone)—i.e., a function that takes some weighed sum of 'likes' and (re)shares, rather than an average (or an argmax). This justifies the shorthand "game of like"—as a nod to J. Conway's "game of life" [4]—since the preferred social outcome, over repetitions, is like-and-reshare, a strengthened form of 'like' ("game of share" would be equally justified, but the homage and homophony would be lost).

[3] Note that i may be indifferent to others' strategies, in which case $u_{p_i}(\overline{\sigma}) = u_{p_i}(\overline{\sigma}')$ whenever i's strategy σ_i is the same in $\overline{\sigma}$ and $\overline{\sigma}'$.

[4] For a concrete example, Twitter's ranking algorithm weighs 'like' reactions more than re-tweets (reshares) when determining which content should appear in users' feeds. A knowledgeable user may prioritize sharing content they believe would receive 'likes' to optimize the chances that other users are exposed to their content later, whether they value engagement as social validation or as a means to increase content visibility.

Let us conclude this section with a few words on our model's (self-imposed) limitations. In real-world OSN, new content can be introduced at any time, and players have only a partial picture of the content they can reshare. A more realistic "game of like" would have imperfect information (e.g., as a model of bounded attention): any content c would be available to a player i to react to at r with a certain probability, depending on overall engagement with c prior to r. In such a model, i could be aware of some c, close to i's ideal content, and care for its visibility (the probability of c being available to other players) and thus for other players' engagement with c. Conversely, i might not worry much about some c, far removed from their ideal, as long as c's probability of being available to other players would remain low. Still, a simplified model with perfect information already acknowledges the relevance of overall interaction by virtue of the argument of $\gamma_i u_{p_i}(\cdot)$ being a strategy profile, and thus furthers goal of identifying strategic components of OSN-sharing. Hence, our "game of like" with perfect information is a proof-of-concept and a foundation for future developments. The next section considers special cases, varying players' γ types, to determine which refinements would be necessary to turn the proof-of-concept model into a model for real-world OSNs.

3 Strategy Selection

Let us begin with the limit case where, for all $i \in P$, $\gamma_i = 1$, denoted $G_{\gamma=1}$ for later reference. We could distinguish *a priori* between a variety of subcases, depending on whether players have non-equivocal prior beliefs about other players' personal preferences; and/or whether they have non-equivocal prior beliefs about other players' γ. However, the differences between those subcases are inconsequential. To see this, assume an arbitrary player i in $G_{\gamma=1}$ who *does have* non-equivocal prior beliefs about other players' personal preferences for content and γ-type (say, following a round of cheap talk). *Ex hypothesis*, at any round r of $G_{\gamma=1}$, for any $i \in P$, $U_i(\overline{\sigma})$=$u_{p_i}(\overline{\sigma})$. Hence, i's best strategy at round $r = 0$ is to share whatever content c_i available to them that is closest to their ideal content (according to their dimensions of evaluation). Beliefs about other players' preferences and γ type do not affect that choice. Hence, i would choose the same content *without* any information about other players. Since the only assumption we made about i is that $\gamma_i = 1$, this generalizes to any $i \in P$ for $G_{\gamma=1}$ (and yields an equilibrium solution in the basic model for $r = 0$ in $G_{\gamma=1}$). Under the assumption that content is only introduced at round $r = 0$, the distance between the 'community content' and any player i's ideal content remains constant across repetitions, whatever their strategy at $(r \geq 1)$. Relaxing this simplifying assumption is one way to model how players can attempt to drive community content closer to their preferences by sharing more content closer to their ideal at any new round $(r \geq 1)$. But this would not bring the model closer to real-world OSN, as "spamming" content is only efficient if the content is visible, bringing us back to a version of the "game of like" with imperfect information. Conversely, a "game of like" *without* content introduced at round $r > 0$, and with $(\gamma = 1)$-players only, would be susceptible to manipulations by coalitions of like-minded players, who would want to see some content promoted. Therefore, relaxing the assumption that no new content is introduced past $r = 0$ would not be especially illuminating without an explicit topological model of content distances and auxiliary assumptions about how variable availability of content correlates with engagement.

In a second limit case, denoted $G_{\gamma=0}$, all $i \in P$ are such that $\gamma_i = 0$. Unlike $G_{\gamma=1}$, player priors about others can significantly impact the game. To see this, consider the limit subcase where players' γ type is common knowledge. Then, $G_{\gamma=0}$ becomes a game of reciprocation-or-retaliation or *quid pro quo*, where players either trade reciprocal 'likes' and re-shares, or ignore one another, and where content becomes inconsequential (so that it matters little whether new content can be introduced after $r = 0$ or not). To see this, consider a simple $G_{\gamma=0}$ case with two players i and j, content introduction restricted to $r = 0$, and (as a simplification) no marginal utility for 'liking' or re-sharing one's content. Hence, the only utility i and j can get is from the other player's liking or re-sharing their content. At $r = 0$, they share (resp.) c_i and c_j. At $r = 1$, i (j) can like or re-share c_j (c_i), or do nothing (for definiteness: repeating their move from $r = 0$). If either does nothing at $r = 1$, the other can retaliate at $r = 2$ by playing nothing; otherwise, they can reciprocate and play the remaining action (like, or reshare) they did not play at $r = 1$. With no introduction of new content, they can repeat the cycle over c_i and c_j. If new content is allowed, they can repeat cycles of three rounds (introduction, like, or re-share, then reciprocation or retaliation) to accrue utility. The strategy just described turning the "game of like" into a game of reciprocation-or-retaliation, and resembles the *tit-for-tat* strategy in the repeated Prisonner's Dilemma.

As extreme as it is, this case suggests that when $(\gamma = 0)$-players have non-equivocal beliefs about one another's γ type, the closer the players are to having correct beliefs, the closer $G_{\gamma=0}$ resembles a *quid pro quo* game. Assume now a subcase of $G_{\gamma=0}$ where players have equivocal beliefs about γ types—i.e., do not *know* that other players are $(\gamma = 0)$-players. If they also have equivocal beliefs about other players' personal preferences for content, the rational choice (for any i) is a mixed strategy assigning equal weight to any content i has access to at $r = 0$ and hope for the best. Lifting the restriction on content introduction is more consequential than in the $G_{\gamma=1}$ case, as repeated observations of others' sharing behavior are necessary to infer their personal preferences for content from their actions or their preferences for engagement. Since, *ex hypothesis*, no player in $G_{\gamma=0}$ actually cares for content (as long as they receive engagement), inferences from sharing behavior to personal preferences could result in 'false consensus' situations if players gradually amplify a salient type of content, leading to an echo chamber (in the sense of [12]; cf. Section 4). However, even without lifting the assumption, we can form a picture of a repeated game with new content by assuming a round of cheap talk prior to $r = 0$, during which players can form priors (or update equivocal priors) about other players' preferences based on observed behavior. Suppose that some candidate content appears salient for eliciting positive reactions—say, pictures of cats in precarious positions. Then, upon engaging in $G_{\gamma=0}$, players could anticipate similar pictures to elicit 'like' and 'share' reactions, skewing the content shared at $r = 0$ toward pictures of cats in precarious positions. Thus, it would appear that a majority of players favor cat pictures. Even without the introduction of new content, this could lead to cat pictures being increasingly reshared at every $r \geq 1$ without (*ex hypothesis*) any player selecting their strategy out of personal preference for that type of content, resulting in a 'false consensus.' Again, as with $G_{\gamma=1}$, how engagement could impact visibility appears more critical than whether or not content may be repeatedly introduced. Subsequently, the need to accommodate $(\gamma < 1)$-players does not require further refinements beyond those suggested by $G_{\gamma=1}$: imperfect information and an explicit content evaluation and comparison model. The latter would, in particular, suffice for representing how $(\gamma < 1)$-players form (and revise) beliefs about the majority's opinion, instrumental in selecting strategies for eliciting engagement.

4 Mutual Expectations and Social Influence

Our remark about the majority's opinion being of import to $(\gamma < 1)$-players may remind the reader of J.M. Keynes' "Beauty Contest" analogy for professional investment, quoted below.

> [P]rofessional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole. [...] It is not a case of choosing those [faces] that,

> to the best of one's judgment, are really the prettiest, nor even those that average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. [9, p. 156]

The parallel is intentional: we propose that Keynes' "third degree" describes the reasoning of a $(\gamma < 1)$-player selecting a strategy that could elicit (re)share reactions from other $(\gamma < 1)$-players who would want to elicit 'like' reactions. More generally, a "game of like" with some proportion of $(\gamma < 1)$-players relates to *guessing games*, proposed as a generalization of J.M. Keynes' beauty contest by R. Nagel (first, in [11]; see [1] for an overview of empirical studies). A formal reconstruction of this suggestion would require an explicit model of preference distances (already identified as a necessary refinement for our basic model to capture real-world OSNs), but we can offer an informal sketch.

Asssume the standpoint of a player of type $\gamma = 0$, that we will denote γ_0, reasoning about other players of a "game of like."[5] When choosing between multiple options for content to share, when γ_0's goal is accruing "like" reactions, γ_0 is equally well-off: (*i*) choosing based on their own preferences for content, or: (*ii*) choosing based on the majority's preference (e.g., as inferred following a round of cheap talk) when preferences agree; and: (*iii*) possibly worse off, when preferences disagree. In case (*iii*), γ_0 would be better off switching to an option that agrees with the majority's (displayed) preferences. Thus, options based on γ_0 preferences are *weakly dominated* by options based on the majority's preferences (as inferred by γ_0). Consider now how γ_0 would approach selecting a strategy for eliciting "share" reactions; as simplification, assume that γ_0 believes that most players are like him and care more for engagement than for content. Then, γ_0 expects that most players would (re-)share content to elicit (at least) 'like' reactions. If γ_0 assumes that those players are rational, they expect those players to reason to (*i–iii*) above. From there, γ_0 can conclude that selecting an option based on their own preferences for content would yield the same payoff as choosing based on the majority's opinion of the majority's (displayed) preferences for content (if in agreement); and possibly a worse payoff (if in disagreement). In the latter case, γ_0 would be better off switching options. Hence, a selection based on the majority's opinion of the majority's (displayed) preferences *weakly dominates* a selection based on γ_0's preferences for content alone.

The argument just sketched guesstimates too many important parameters to be general—e.g., the respective distribution of γ types among the players, the cost of seeking social feedback with contrary-to-personal preferences for other players that γ_0, how γ_0 would arrive at estimates for those, etc. However, it suffices to motivate a comparison between a subclass of "game of like," Keynes' beauty contest, and Nagel's guessing games. And empirical motivation for this comparison would be the reconstruction of the real-world OSN behavior colloquially called 'signal boosting,' whereby users of an OSN leverage the influence of public figures ("influencers") with a larger following base, tagging them in hope to be re-shared. A well-reported example is a November 13, 2020 Twitter video featuring actor R. Quaid reading aloud an earlier tweet from then-US president D.J. Trump under a stroboscopic light, with an over-dramatic voice. Trump (unsurprisingly) reshared Quaid's video, which then accrued millions of views from Trump's followers, reaching beyond Quaid's following. In fact, we have already encountered in Section 3 a variant of (involuntary) signal-boosting behavior, as a pathway to amplification (false consensus) when discussing $G_{\gamma=0}$. This seems grounds enough to suggest that a "game of like" model of OSN with influencers could contribute to a formal theory of online amplification, echoing Keynes'

[5] We assume that the agent is a $(\gamma = 0)$-player rather than a weaker $(\gamma < 1)$ to avoid dealing with correlations between personal preferences for content and preferences for engagement. Otherwise, we would have to factor in the cost of sharing contrary-to-preference content, which could offset the benefit of engagement.

motivations for the beauty-contest analogy (speculative asset bubbles).

Another possible contribution that circles back to social epistemology is a possible formal reconstruction of Nguyen's conceptual analysis [12]. Nguyen proposes that *epistemic bubbles* occur when individuals receive limited exposure to information sources challenging their pre-existing beliefs, in contrast to *echo chambers*, which emerge when individuals receive extensive exposure to information sources that align with their pre-existing beliefs. Epistemic bubbles result from combined personal choice and algorithmic curation, particularly when online platforms tailor content to individual preferences, thereby restricting the information individuals encounter. In an echo chamber, people reinforce their views and are shielded from diverse perspectives and alternative information, leading to the exclusion of dissenting opinions. Nguyen notes that epistemic bubbles are easy to burst with the presentation of contrary evidence, while echo chambers are self-reinforcing, with social interaction actively fostering distrust of outside sources. Nguyen's analysis of echo chambers invites a formal reconstruction in a "game of like" model with imperfect information, bringing it closer to the methodological frameworks of behavioral science (*modulo* a game-to-decision translation).

5 Concluding Remarks

We argued in Section 1 that, while OSN-sharing is a strategic interaction, behavioral science models overlook the contribution of strategic anticipations. We extrapolated from behavioral science decision-theoretic models a basic game model of OSN-sharing (Section 2) and explored some limit cases to determine refinements necessary to capture real-world OSN-sharing (Section 3). A connection with Keynes' Beauty contest (and, more generally, guessing games) allowed us to sketch a strategic analysis of content amplification in the presence of influencers and users leveraging influence and suggested a direction for the model's development (Section 4). Still, a "game of like" model may not contribute to conceptual analysis beyond a formal reconstruction of Nguyen's framework. And Nguyen's informal analysis has already done the heavy lifting of rigorously ordering concepts inherited from unsystematic public discourse, such as "echo chambers" and "filter bubbles" (introduced, resp., in [22] and [15]), whose previously heterogeneous use had prevented consensus among researchers (see [20]). Rather, the litmus test for a "game of like" model would be a contribution to the critical re-evaluation of empirical data assessed from a decision-theoretic standpoint; and a suggestion of empirical investigations that a decision-theoretic standpoint would have neglected. To conclude, we want to suggest that, as incomplete as it is, our "game of like" model already achieves that.

As for critical re-evaluation, consider the widely-publicized study by Pennycook *et al.* [16], in which the intervention condition proceeds from the auxiliary hypothesis that accuracy competes for attention with social incentives.[6] From a strategic standpoint, the authors' other auxiliary hypothesis—that "people do care more about accuracy than other content dimensions" (p. 591)—could characterize common knowledge of one dimension of users' preferences. If it does, having "the concept of accuracy more [...] salient in [one's] mind" (*ibid*) could *prime* engagement-based expectations, rather than shutting them down; in a game-to-decision translation, a Bayesian decision-maker would then anticipate a better prospect of eliciting other users' reactions conditional on being perceived as accurate (compared to

[6] "In the control condition of each experiment, participants were shown 24 news headlines (balanced on veracity and partisanship) and asked how likely they would be to share each headline on Facebook. In the treatment condition, participants were asked to rate the accuracy of a single non-partisan news headline at the outset of the study (ostensibly as part of a pretest for stimuli for another study). They then went on to complete the same sharing intentions task as in the control condition, but with the concept of accuracy more likely to be salient in their minds." [16, p. 591]

conditional on being perceived as inaccurate). Compare this with the intervention condition from the more recent study by Ren *et al.* [18], which socially incentivized both accuracy and engagement.[7] As for the design of new studies, consider the question of whether differences in intervention conditions between [16] and [18] translate into differences in reasoning about other users' strategies is an interesting question. A positive answer would partition "accuracy nudges" into two classes (engagement-based and non-engagement-based). A negative answer would invalidate the auxiliary hypothesis that accuracy competes with the social dimension. The connection we drew with Nagel's work on guessing games suggest an empirical approach to answering this question, with following the methodology of [3], which established neural correlates of lower- and higher-order "Keynes degree" reasoning in guessing games.

References

[1] Antoni Bosch-Domenech, Jose G. Montalvo, Rosemarie Nagel & Albert Satorra (2002): *One, two, (three), infinity,...: Newspaper and lab beauty-contest experiments*. American Economic Review 92(5), pp. 1687–1701, doi:10.1257/000282802762024737.

[2] Nick Chater & Mike Oaksford (2008): *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA, New York, NY, doi:10.1093/acprof:oso/9780199216093.001.0001.

[3] Giorgio Coricelli & Rosemarie Nagel (2009): *Neural correlates of depth of strategic reasoning in medial prefrontal cortex*. Proceedings of the National Academy of Sciences 106(23), pp. 9163–9168, doi:10.1073/pnas.0807721106.

[4] Martin Gardner (1970): *The fantastic combinations of John Conway's new solitaire game "life"*. Scientific American 223(4), pp. 120–123, doi:10.1038/scientificamerican1070-120.

[5] Alvin Goldman (1999): *Knowledge in a Social World*. Oxford University Press, Oxford, doi:10.1093/0198238207.001.0001.

[6] John C. Harsanyi (1982): *Comment—Subjective probability and the theory of games: Comments on Kadane and Larkey's paper*. Management Science 28(2), pp. 120–124, doi:10.1287/mnsc.28.2.120.

[7] John C. Harsanyi (1982): *Rejoinder to professors Kadane and Larkey*. Management Science 28(2), pp. 124–125, doi:10.1287/mnsc.28.2.124a.

[8] Joseph B. Kadane & Patrick D. Larkey (1982): *Subjective probability and the theory of games*. Management Science 28(2), pp. 113–120, doi:10.1287/mnsc.28.2.113.

[9] John M. Keynes (1978): *The General Theory of Employment, Interest and Money (1936)*. In Johnson E. & Moggridge D., editors: The Collected Writings of John Maynard Keynes, 7, Royal Economic Society, doi:10.1017/UPO9781139524278.

[10] Hause Lin, Gordon Pennycook & David G. Rand (2023): *Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing*. Cognition 230, 105312, doi:10.1016/j.cognition.2022.105312.

[11] Rosemarie Nagel (1995): *Unraveling in guessing games: An experimental study*. The American Economic Review 85(5), pp. 1313–1326. Available at http://www.jstor.org/stable/2950991.

[12] C. Thi Nguyen (2020): *Echo chambers and epistemic bubbles*. Episteme 17(2), pp. 141–161, doi:10.1017/epi.2018.32.

[7] "In each incentive condition, we told participants that they would be entered into a lottery for a $50 prize, and we manipulated how they would earn tickets to increase their odds of winning the prize. In the *Accuracy* condition, we told participants that they would earn a ticket if the post they shared was validated to be true by a professional fact-checker. In the *Like* condition, we told participants that they would earn a ticket for each "like" they received from others. In the *Comment* condition, we told participants that they would earn a ticket for each comment they received from others. In the *Control* condition, we did not provide additional incentives. [18, 104421:4]

[13] Mike Oaksford & Nick Chater (1994): *A rational analysis of the selection task as optimal data selection*. Psychological Review 101(4), pp. 608–631, doi:10.1037/0033-295X.101.4.608.

[14] Cailin O'Connor & James O. Weatherall (2019): *The Misinformation Age: How False Beliefs Spread*. Yale University Press, New Haven, CT, doi:10.2307/j.ctv8jp0hk.

[15] Eli Pariser (2011): *The Filter Bubble: What the Internet is Hiding From You*. Penguin UK, London.

[16] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles & David G. Rand (2021): *Shifting attention to accuracy can reduce misinformation online*. Nature 592(7855), pp. 590–595, doi:10.1038/s41586-021-03344-2.

[17] Steve Rathje, Jon Roozenbeek, Jay J. Van Bavel & Sander van der Linden (2023): *Accuracy and social motivations shape judgments of (mis)information*. Nature Human Behaviour, pp. 1–12, doi:10.1038/s41562-023-01540-w.

[18] Zhiying (Bella) Ren, Eugen Dimant & Maurice Schweitzer (2023): *Beyond belief: How social engagement motives influence the spread of conspiracy theories*. Journal of Experimental Social Psychology 104, 104421, doi:10.1016/j.jesp.2022.104421.

[19] Jon Roozenbeek, Alexandra L. J. Freeman & Sander van der Linden (2021): *How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020)*. Psychological Science 32(7), pp. 1169–1178, doi:10.1177/09567976211024535.

[20] Amy Ross Arguedas, Craig T. Robertson, Richard Fletcher & Rasmus K. Nielsen (2022): *Echo chambers, filter bubbles, and polarisation: a literature review*. University of Oxford, Reuters Institute for the Study of Journalism. Available at https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review.

[21] Thomas W. Simpson (2012): *Evaluating Google as an epistemic tool*. Metaphilosophy 43(4), pp. 426–445, doi:10.1111/j.1467-9973.2012.01759.x.

[22] Cass R. Sunstein (2001): *Echo chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton University Press Princeton, NJ.

[23] Soroush Vosoughi, Deb Roy & Sinan Aral (2018): *The spread of true and false news online*. Science 359(6380), pp. 1146–1151, doi:10.1126/science.aap9559.

Belief Revision from Probability

Jeremy Goodman
School of Philosophy
University of Southern California, USA
jeremy.goodman@usc.edu

Bernhard Salow
Faculty of Philosophy
University of Oxford, UK
bernhard.salow@philosophy.ox.ac.uk

In previous work ([5, 6]), we develop a question-relative, probabilistic account of belief. On this account, what someone believes relative to a given question is (i) closed under entailment, (ii) sufficiently probable given their evidence, and (iii) sensitive to the relative probabilities of the answers to the question. Here we explore the implications of this account for the dynamics of belief. We show that the principles it validates are much weaker than those of orthodox theories of belief revision like AGM [1], but still stronger than those valid according to the popular Lockean theory of belief [4], which equates belief with high subjective probability. We then consider a restricted class of models, suitable for many but not all applications, and identify some further natural principles valid on this class. We conclude by arguing that the present framework compares favorably to the rival probabilistic accounts of belief developed by Leitgeb [13, 14] and Lin and Kelly [17].

1 Probability Structures

We will work with the following simplification of the models in [5]:

Definition 1.1. A *probability structure* is a tuple $\langle S, \mathscr{E}, Q, Pr, t\rangle$ such that:

1. S is a non-empty set (of *states*),
2. $\mathscr{E} \subseteq \mathscr{P}(S)\backslash\{\emptyset\}$ (the *possible bodies of evidence*),
3. Q (the *question*) is a partition of S,
4. Pr (the *prior*) is a probability distribution over S, and
5. $t \in [0,1]$ (the *threshold*)

Propositions are modeled as subsets of S, where p is true in s if and only if $s \in p$. We say that $E' \in \mathscr{E}$ is *the result of discovering* p in $E \in \mathscr{E}$ just in case $E' = E \cap p$; this will allow us to talk about how beliefs evolve in response to changes in one's evidence.

Which propositions an agent believes is a function of their evidence and is also given by a set of states, so that an agent with evidence E believes p if and only if $B(E) \subseteq p$. This ensures that their beliefs are closed under entailment, and thus already marks a departure from popular 'Lockean' accounts of belief [4], according to which one believes a proposition if and only if its probability exceeds a particular threshold. But it is compatible with the more plausible direction of Lockeanism, namely:

> THRESHOLD: You believe p only if p is sufficiently probable given your evidence.
> If $B(E) \subseteq p$, then $Pr(p|E) \geq t$.

We can think of the members of the question Q as its *answers*; we write $[s]_Q$ for the member of Q containing s. The proposal in [5] then boils down to claiming that $s \in B(E)$ if and only if $s \in E$ and the answers to Q that are more probable than $[s]_Q$ have total probability less than the threshold t. Writing Pr_E for $Pr(\cdot|E)$, this can be formalized as follows:

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 308–317, doi:10.4204/EPTCS.379.25

Definition 1.2. $B(E) = \{s \in E : Pr_E(\{s' : Pr_E([s']_Q) > Pr_E([s]_Q)\}) < t\}$

This means that one believes as much as possible subject to two constraints: (i) THRESHOLD, and (ii) that the totality of one's beliefs corresponds to the conjunction of one's evidence with a disjunction of answers to Q that includes any answer at least as probable (given one's evidence) as any other it includes. One notable attraction of this proposal is that what one believes corresponds to the discrete analogue of the highest posterior-density region typically used to define 'credible intervals' from probability density functions in Bayesian statistics. A logically significant feature of the proposal, to which we will return later, is that it involves not only local probability comparisons between different answers to Q, but also a global probability comparison between a collection of such answers and the threshold t.

2 Principles and Results

A core idea behind the orthodox AGM [1] theory of belief revision is that belief revisions are trivial whenever what you learn is compatible with your initial beliefs: you should simply add the discovery to your beliefs, draw out the logical consequences of these beliefs, and leave everything else unchanged. Here we will focus on five principles that encode various aspects of this idea. Exploring when and how these principles can fail will be a useful way of exploring the extent to which our account of belief requires departing from orthodoxy when it comes to belief dynamics. These principles are:[1]

$\Diamond-$ If you don't believe not-p and then discover p, you shouldn't give up any beliefs.

If $B(E) \cap p \neq \emptyset$, then $B(E \cap p) \subseteq B(E)$.

$\Diamond R$ If you don't believe not-p and then discover p, you shouldn't reverse any of your beliefs (i.e. go from believing something to believing its negation).

If $B(E) \cap p \neq \emptyset$, then $B(E) \cap B(E \cap p) \neq \emptyset$.

$\Box+$ If you believe p and then discover p, you shouldn't form any new beliefs.

If $B(E) \subseteq p$, then $B(E) \subseteq B(E \cap p)$.

$\Box-$ If you believe p and then discover p, you shouldn't give up any beliefs.

If $B(E) \subseteq p$, then $B(E \cap p) \subseteq B(E)$

$\Box R$ If you believe p and then discover p, you shouldn't reverse any of your beliefs.

If $B(E) \subseteq p$, then $B(E) \cap B(E \cap p) \neq \emptyset$.

These principles are not logically independent: the $\Diamond$ principles entail the corresponding $\Box$ principles, and the $+$ and $-$ principles each entail the corresponding R principles. All of them are valid according to AGM. By contrast, only $\Box R$ is valid according to Lockean theories that equate believing a proposition with assigning it a sufficiently high probability (for some probability threshold less than 1), and it is valid only if this probability threshold is above $\frac{\sqrt{5}-1}{2} \approx .62$ (as discussed in [19]).

The present account falls in between these extremes:

Proposition 1. *$\Box-$ and $\Box R$ are valid on the class of probability structures.*

Proposition 2. *$\Diamond-$, $\Diamond R$, and $\Box+$ can all fail in probability structures.*

[1] The $\Diamond$ indicates that the discovery is compatible with your initial beliefs, while the $\Box$ indicate that it is something you initially believe. $\Diamond-$ is often referred to as 'preservation'; [19] call $\Box-$ 'weak preservation' and $\Box R$ 'very weak preservation'. If we interpret the non-monotonic consequence relation $p \mid\!\sim q$ as saying that $B(p) \subseteq q$, then $\Diamond-$ corresponds to 'rational monotony', $\Box+$ to 'cut', and $\Box-$ to 'cautious monotony' in the standard terminology from [12].

We will illustrate Proposition 2 with two examples. Consider a much discussed thought experiment:

> **Flipping for Heads**
> A coin flipper will flip a fair coin until it lands heads.

A natural model of this case is as follows:

$$\begin{array}{ll} S = \{s_1, s_2, \dots\} & \mathscr{E} = \{\{s_i, s_{i+1}, s_{i+2}, \dots\} : s_i \in S\} \\ Q = \{\{s_i\} : s_i \in S\} & Pr(\{s_i\}) = \frac{1}{2^i} \\ t = .99 & \end{array}$$

Here s_i is the state in which the coin lands heads on the ith flip, and $\{s_i, s_{i+1}, s_{i+2}, \dots\}$ is your evidence if you have just observed the coin land tails on the first $i-1$ flips. The question Q is maximally fine-grained, and the probabilities match the known objective chances.

In this model, $B(\{s_i, s_{i+1}, s_{i+2}, \dots\}) = \{s_i, s_{i+1}, \dots, s_{i+6}\}$: you always believe that the coin will land heads within the next seven flips. $\Diamond-$ is violated whenever you observe the coin land tails. For example, let $p = \{s_2, s_3, \dots\}$. Then $B(S) \cap p \neq \emptyset$, but $B(S \cap p) = \{s_2, \dots, s_8\} \not\subseteq \{s_1, \dots, s_7\} = B(S)$. We think this is exactly the right prediction.

To turn this into a counterexample to $\Box+$, we add new body of evidence $E' = \{s_1, s_2, \dots, s_7\}$ to $\mathscr{E}$. Intuitively, we can think of this as the evidence you receive if you walk away from the experiment before the first flip, and are later told that the coin landed heads within the first seven flips. It is easy to verify that $B(E') = \{s_1, \dots, s_6\}$. So $B(S) \not\subseteq B(S \cap E')$, even though $B(S) \subseteq E'$. That this can happen should be unsurprising in a framework like ours in which agents have 'inductive' beliefs that go beyond what is strictly entailed by their evidence: discovering something that you previously believed only inductively will strengthen your evidence, putting you in a position to draw further inductive conclusions.

Counterexamples to $\Diamond R$ are subtler, for reasons we will explain in the next section. But here is one:

> **Drawing a Card**
> You are holding a deck of cards, which is either a fair deck consisting of 52 different cards or a trick deck consisting of 52 Aces of Spades. Your background evidence makes it 90% likely that the deck is fair. You draw a card at random; it is an Ace of Spades.

Here is a possible model of the example:

$$\begin{array}{ll} S = \{F_1, F_2, \dots, F_{52}, T\} & \mathscr{E} = \{S, \{F_1\}, \dots, \{F_{51}\}, \{F_{52}, T\}\} \\ Q = \{\{F_1, F_2, \dots, F_{52}\}, \{T\}\} & Pr(\{F_i\}) = \frac{.9}{52} \approx .017, Pr(\{T\}) = .1 \\ t = .85 & \end{array}$$

The states F_i are all states in which the deck is fair; they are distinguished only by which card you will draw, with F_{52} being the one where you draw the Ace of Spades. State T is the state in which the deck is the trick deck (and you thus draw an Ace of Spades). Your evidence settles all and only what card you drew; so when you draw an Ace of Spades, it leaves open both that you did so by chance and that you did so because it is a trick deck. The question is simply whether the deck is fair. It is easy to see that, according to this model, you should initially believe only that the deck is fair. Your initial beliefs are thus compatible with it being fair and you drawing the Ace of Spades by chance. Yet when you discover that you drew an Ace of Spades, you should reverse your opinion and conclude that you're holding the trick deck, since $Pr(\{T\}|\{F_{52}, T\}) \approx \frac{.1}{.1+.017} \approx .855 > t$.

Note that, in this model, your discovery is not a disjunction of answers to the question Q. If we changed the question to a more fine-grained one, so that your discovery was a disjunction of its answers, then the case would no longer yield a counterexample to $\Diamond R$. For example, relative to the question *is the deck fair and will I draw an Ace of Spades* – i.e. relative to $Q' = \{\{F_1, F_2, \dots, F_{51}\}, \{F_{52}\}, \{T\}\}$ – you will initially believe that you won't draw an Ace of Spades, in which case your subsequent discovery isn't compatible with your initial beliefs. And relative to the question *is the deck fair and what will I draw* – i.e. relative to the maximally fine-grained $Q'' = \{\{s\} : s \in S\}$ – you will initially have no non-trivial beliefs, and in particular you won't start out believing that the deck is fair. In the next section, we will see that this is part of a more general pattern about counterexamples to $\Diamond R$.

3 Orthogonality

In the previous section, we saw that some of the surprising belief dynamics in probability structures depended on discoveries that cross-cut the question. Notice that structures in which this cannot happen, because every member of $\mathscr{E}$ is the union of some subset of Q, satisfy the following constraint:

$$\textsc{orthogonality}: \frac{Pr([s]_Q)}{Pr([s']_Q)} = \frac{Pr([s]_Q|E)}{Pr([s']_Q|E)} \text{ for all } s, s' \in E \in \mathscr{E} \text{ s.t. } Pr([s']_Q|E) > 0$$

This says that the only way that getting new evidence can change the relative probability of two answers to Q is by completely ruling out one of those answers. While we can ensure ORTHOGONALITY by making the question fine-grained enough to capture all possible discoveries, this isn't always necessary. For example, we could fine-grain the states and bodies of evidence in our model of **Flipping for Heads** to capture the fact that you discover where on the table the coin lands. The bodies of evidence in such a fine-grained model will cross-cut the question *how many time will the coin be flipped*; but, plausibly, ORTHOGONALITY will still hold for this question, since the added information about where the coin lands is probabilistically independent from how many times it will be flipped.

ORTHOGONALITY is interesting because it leads to a stronger logic of belief revision. Firstly,

Proposition 3. *$\Diamond R$ is valid on the class of probability structures satisfying* ORTHOGONALITY.

Secondly, consider the following principle. It says (roughly) that if you're sure that, whatever you're about to discover, you won't believe a given proposition afterwards, then you already don't believe it:

$\Pi-$ If Π is a partition any member of which you could discover, then there is a $p \in \Pi$ such that you shouldn't give up any beliefs upon discovering p.

If $\Pi \subseteq \mathscr{E}$ is a partition of E, then $B(E \cap p) \subseteq B(E)$ for some $p \in \Pi$.

We then have the following result:

Proposition 4. $\Pi-$ *can fail in probability structures. But it is valid on the class of probability structures satisfying* ORTHOGONALITY.

It is also worth noting that $\Diamond-$ and $\Box+$ can still fail in structures satisfying ORTHOGONALITY. In particular, ORTHOGONALITY holds in the structures we used in the last section to argue that **Flipping for Heads** yields counterexamples to $\Diamond-$ and $\Box+$.

In our view, a good deal of ordinary talk about what people believe is well-modelled by structures satisfying ORTHOGONALITY. This is because we think that the question Q, to which attributions of belief are implicitly relativized, typically coincides with the question under discussion in the conversational context in which those attributions are made. Moreover, when a discovery is salient, it is natural to consider a question that is sufficiently fine-grained to capture all the aspects of this discovery that are relevant to its answers. Counterexamples to ORTHOGONALITY (and thus to $\Diamond R$ and $\Pi-$) therefore tend to be 'elusive' in Lewis's [16] sense: attending to these cases often changes the context in such a way that they can no longer be described as counterexamples.

That being said, we do not think that ORTHOGONALITY is plausible as a general constraint. This is because, very often, the only way to ensure ORTHOGONALITY is to adopt a very fine-grained question;

and, often, such fine-grained questions make overly skeptical predictions about what we can believe. Consider, for example, the following case:

> **One Hundred Flips**
> You will flip a fair coin 100 times and watch how it lands each time.

There are natural contexts in which you can be correctly described as initially believing that the coin will not land heads more than 90 times. Our theory predicts this for various natural questions, even for very high thresholds t – for example, the polar question *will the coin land heads more than 90 times* or the slightly more fine-grained question *how often will the coin lands heads*. But neither of these questions satisfies ORTHOGONALITY. For example, discovering that coin lands tails on the first flip will favor 'no' over 'yes' for the first question, and '51' over '49' for the second question, without ruling out any of these answers. In fact, the only natural question that satisfies ORTHOGONALITY is the maximally fine-grained question *what will the exact sequence of heads and tails be*. But all answers to this question are equally likely, and so this question prevents you from having any non-trivial beliefs about what will happen.

We conclude that ORTHOGONALITY should be rejected as a general constraint, even if it will often hold when we are considering a particular case with a limited range of discoveries. $\Diamond R$ and $\Pi-$ are thus not fully general principles of belief revision; but counterexamples are likely to be difficult to pin down.

ORTHOGONALITY is also a fruitful principle in that it helps to facilitate comparisons between our framework and other probabilistic theories of belief. Let us now turn to these.

4 Comparisons

In this section, we consider two influential probabilistic accounts from the literature, and compare them with our own account. The first can be seen as a version of our account with an additional constraint imposed on probability structures, and validates $\Diamond-$ but not $\Box+$; the second can be seen as defining belief from probability structures in a related but different way, and validates $\Box+$ but not $\Diamond-$.

4.1 A Stability Theory

The first theory we want to consider is inspired by Leitgeb's *stability theory of belief* [13, 14]. The guiding idea behind this theory is a probabilistic analogue of $\Diamond-$ that Leitgeb calls the *Humean thesis*. However, despite the 'stablity' moniker, the constraints imposed by Leitgeb's theory are *synchronic* ones relating probabilities, partitions, and thresholds at a single time. So both to facilitate comparison with our framework, and to be (in our view) more faithful to its motivating idea, we will consider a strengthening of Leitgeb's theory according to which the requirements it imposes on one's beliefs prior to a discovery continue to hold after one has made that discovery. We can then interpret the view as proposing the following constraint on probability structures:[2]

STABILITY: For all $E \in \mathscr{E}$ and $X \subseteq Q$, if $Pr(\bigcup X) \geq t$ and $E \cap \bigcup X \neq \emptyset$, then $Pr(\bigcup X|E) \geq t$.

We then have the following results:

Proposition 5. $\Diamond-$ *is valid in probability structures satisfying* STABILITY *and* ORTHOGONALITY. *But* $\Box+$ *can still fail; and* $\Diamond-$ *can fail in structures satisfying* STABILITY *but not* ORTHOGONALITY.

[2] Our STABILITY strengthens Leitgeb's theory in two ways: first, by identifying the threshold that characterizes the minimal probability of anything one believes with the threshold in terms of which stability is defined, and, second, by not allowing this threshold to be different for different possible bodies of evidence. It also departs from his formulation in quantifying over $\mathscr{E}$ rather than $\{\bigcup Y : Y \subseteq Q\}$; however, we read him as identifying $\mathscr{E}$ with $\{\bigcup Y : \emptyset \neq Y \subseteq Q\}$, so this is not a substantive departure.

This illustrates how a kind of qualitative stability of belief can be secured by a kind of probabilistic stability (given ORTHOGONALITY), without entailing the full strength of AGM.[3]

We reject STABILITY because we reject $\Diamond-$ (even in cases where ORTHOGONALITY holds), and along with it the informal idea that rational belief should be stable in anything like the way that Leitgeb claims it should be. STABILITY also places implausible constraints on what agents can believe at a given time. For example, [11] show, in effect, that in **Flipping for Heads** STABILITY entails that the only way to have any non-trivial beliefs about how many times the coin will be flipped is to believe that it will be flipped only once. (This argument depends only on the symmetries of the example, and doesn't depend on whether the coin is fair, biased towards heads, or biased towards tails.) See also [18] and [3].

4.2 The Tracking Theory

Lin and Kelly [17] defend a theory which (for reasons we can't explain here) they call the 'tracking theory' of belief. This theory can be seen as an alternative way of defining belief in probability structures, with the parameter t playing a rather different role. Put informally, a state s is compatible with your LK-beliefs if there is no answer to Q that is more than $\frac{1}{t}$ times more likely than $[s]_Q$. Formally:

Definition 4.1. $B_{LK}(E) = \{s \in E : (\forall q \in Q)(Pr_E([s]_Q) \geq t \times Pr_E(q))\}$

In many cases – such as **Flipping for Heads** – the subject will have similar beliefs according to our theory and according to Lin and Kelly's (provided t is chosen judiciously: low values of t for Lin and Kelly correspond to high values of t for us). However, there are important structural differences between the theories. In particular, LK-beliefs are sensitive only to local comparisons of probability between particular answers, while beliefs as we understand them depend also the probabilities of sets of answers. A consequence of this locality is that, as Lin and Kelly note, their theory validates a reasonably strong theory of belief revision (assuming ORTHOGONALITY, which they essentially build in):

Proposition 6. *$\Box+$, $\Box-$, $\Box R$, $\Diamond R$, and $\Pi-$ are all valid for LK-belief on the class of probability structures satisfying* ORTHOGONALITY.

The major shortcoming of the tracking theory, in our view, is that it fails to entail THRESHOLD. Consider a case like **Drawing a Card**, in which one state initially has very low probability (.1) but every other state has even lower probability (.017). Then relative to a fine-grained question such as *is the deck fair and which card will you draw*, you will LK-believe that the deck is a trick deck even for reasonably low values of t (such as .2). But this belief is only .1 likely on your evidence! And we can, of course, make the case more extreme by increasing the number of distinct cards in the fair deck; so the believed proposition can be arbitrarily improbable for any fixed value of t.

One might defend the tracking theory against such cases by insisting that we choose a more coarse-grained question; while the theory still fails to entail THRESHOLD, this response at least prevents it from recommending the extreme violations just discussed. However, moving to coarser-grained questions is often in conflict with ORTHOGONALITY. Moreover, the reasons we gave previously for rejecting ORTHOGONALITY as a general constraint applies to the tracking theory as well: just like our theory, the tracking theory will make implausibly skeptical predictions in **One Hundred Flips** unless combined with an ORTHOGONALITY-violating question such as *how many heads will there be*.

Without ORTHOGONALITY, the dynamics of LK-belief are substantially less constrained:

Proposition 7. *$\Box+$ and $\Box-$ are valid for LK-belief on the class probability structures. $\Diamond R$ and $\Pi-$ can both fail in such structures.*

[3] Leitgeb [14, chapter 4] describes his theory as compatible with AGM (and thus with $\Box+$) since, upon getting new evidence, one may adopt a different, higher threshold than before. But doing so is in no way required by the demands of stability.

$\Box+$ is then the only principle valid for LK-beliefs but not for beliefs as we understand them.

Moreover, without ORTHOGONALITY, the tracking theory invalidates a new principle that holds for belief as we understand it (assuming we restrict to probability structures with $t > .5$). Consider the following variant of **Drawing a Card** (taken from [8], who also makes parallel observations as an objection to Levi's [15] account of belief):

> **Drawing a Card v.2**
> You are holding a deck which could be either a 'fair' deck of 52 different cards, or one of 52 different 'trick' decks that just contain the same card 52 times. Given your background evidence, the probability that you are holding the fair deck is $\frac{1}{5}$, with the remaining $\frac{4}{5}$ distributed evenly across the 52 trick decks. You are about to draw and turn over one card from your deck.

Let us assume that Q is *which of the 53 possible decks am I holding* and $t > .25$. According to the tracking theory, you initially believe that you hold the fair deck, but after drawing a card you believe that you are holding the relevant trick deck. So we have a failure of the following principle, which says (roughly) that if you're sure that, whatever you're about to discover, you'll believe that a given proposition is false, then don't currently believe that the proposition is true:

ΠR If Π is a partition any member of which you could discover, there is a $p \in \Pi$ such that you shouldn't reverse any of your beliefs upon discovering p.

$$\text{If } \Pi \subseteq \mathscr{E} \text{ is a partition of } E\text{, then } B(E) \cap B(E \cap p) \neq \emptyset \text{ for some } p \in \Pi.$$

By contrast, if belief requires probability over a threshold greater than .5 (as it does on our account), this principle cannot fail.[4]

Overall, then, we see few advantages for the tracking theory over our own. Given ORTHOGONALITY, which Lin and Kelly essentially build into their formalism, the tracking theory offers a stronger theory of belief revision. However, the theory violates THRESHOLD, often in dramatic ways. Moreover, to make reasonable predictions in cases like **One Hundred Flips**, both theories need to appeal to coarse-grained questions that conflict with ORTHOGONALITY. Having done so, both theories invalidate many principles of belief revision, although the details differ slightly (with our theory invalidating $\Box+$ and Lin and Kelly's invalidating ΠR).

5 Further work

We conclude with three directions for further work. One concerns nonmonotonic consequence, where $p \mid\!\sim q$ is interpreted as $B(p) \subseteq q$. We think that distinguishing one's evidence from one's beliefs that go beyond one's evidence offers a productive way of thinking about nonmonotonic consequence, and that the logic resulting from our framework contrasts in interesting ways with the one resulting from Lockean theories of belief (explored in [9]).

The second direction concerns constraints on $\mathscr{E}$. Consider, for example, the Monty Hall problem, in which it is crucial that when one gets new evidence about one's environment, one also gets evidence that one has gotten such evidence. We argue in [7] that such cases motivate a *nestedness* requirement on $\mathscr{E}$: if two possible bodies of evidence are mutually consistent, then one entails the other. This requirement induce new subtleties in the resulting nonmonotonic logic.

[4] Failures of ΠR are to be expected for certain notions of belief that are weaker than the one we are operating with here. For example, your 'best guess' about what deck you are holding plausibly does change no matter what card you draw; and arguably what we 'believe' (in ordinary English) often aligns with our best guesses. See [10] and [2] for discussion.

A third question for future work concerns what happens when probability structures are generalized by making the relevant question a function of one's evidence. [5, Appendix C] motivate this generalization, in order to vindicate certain judgments about a family of examples discussed in [6]. We hope to explore these models in future work; one notable feature is that they invalidate $\Box-$ but still validate ΠR.

Acknowledgements

We thank Kevin Dorst, Josh Pearson, and three anonymous referees for TARK for very helpful comments on earlier versions of this material.

A Proofs

Proposition 1. *$\Box-$ and $\Box R$ are valid on the class of probability structures.*

Proof. Since $\Box-$ entails $\Box R$, it's sufficient to prove the former. We suppose that $B(E) \subseteq p$, and show that $B(E \cap p) \subseteq B(E)$.

Note that if $s \in B(E)$, $[s]_Q \subseteq B(E) \subseteq p$. So for any $q \in Q$, if $Pr([s]_Q|E) \geq Pr(q|E)$, then also $Pr([s]_Q|E \cap p) \geq Pr(q|E \cap p)$. Contraposing, this means that if $Pr(q|E \cap p) \geq Pr([s]_Q|E \cap p)$ and $s \in B(E)$, then $Pr(q|E) \geq Pr([s]_Q|E)$, and so $q \subseteq B(E)$.

Moreover, since $B(E) \subseteq p$, $Pr(B(E)|E \cap p) \geq Pr(B(E)|E) \geq t$.

Now, note that $B(E \cap p)$ is the minimal $X \subseteq E \cap p$ such that (i) if $s \in X$ and $Pr(q|E \cap p) \geq Pr([s]_Q|E \cap p)$ for $q \in Q$, then $q \subseteq X$, and (ii) $Pr(X|E \cap p) \geq t$. By the above, $B(E)$ satisfies both (i) and (ii); so it contains the minimal such X as a subset. So $B(E \cap p) \subseteq B(E)$, as required.

□

Proposition 2. *$\Diamond-$, $\Diamond R$, and $\Box+$ can all fail in probability structures.*

Proof. Counter-models are given in the main text. □

Proposition 3. *$\Diamond R$ is valid in probability structures satisfying* ORTHOGONALITY.

Proof. Suppose that $B(E) \cap p \neq \emptyset$. Let $s \in B(E) \cap p$ be such that $Pr_E([s]_Q) \geq Pr_E([s']_Q)$ for any $s' \in B(E) \cap p$. We will show that, given ORTHOGONALITY, there can be no $q \in Q$ such that $Pr_{E\cap p}(q) > Pr_{E\cap p}([s]_Q)$. It follows that $s \in B(E \cap p)$, thus establishing $B(E) \cap B(E \cap p) \neq \emptyset$.

By ORTHOGONALITY, if $q \in Q$ and $Pr_{E\cap p}(q) > Pr_{E\cap p}([s]_Q)$, then either $Pr_E(q) > Pr_E([s]_Q)$ or else $Pr_{E\cap p}([s]_Q) = 0$. But $s \in E \cap p$, so $Pr_{E\cap p}([s]_Q) \neq 0$. So suppose $Pr_E(q) > Pr_E([s]_Q)$. By the way s was chosen, it follows that $q \cap (B(E) \cap p) = \emptyset$. But $q \cap p \neq \emptyset$, since $Pr_{E\cap p}(q) > 0$. So $q \cap B(E) = \emptyset$. But since $s \in B(E)$, this contradicts the assumption that $Pr_E(q) > Pr_E([s]_Q)$.

□

Proposition 4. *$\Pi-$ can fail in probability structures. But it is valid in probability structures satisfying* ORTHOGONALITY.

Proof. To see that $\Pi-$ can fail, consider

$$\begin{array}{ll} S = \{s_1, s_2, s_3, s_4, s_5, s_6\} & \mathscr{E} = \{S, \{s_1, s_3, s_5\}, \{s_2, s_4, s_6\}\} \\ Q = \{\{s_1, s_2\}, \{s_3, s_4\}, \{s_5\}, \{s_6\}\} & Pr \text{ is uniform} \\ t = .65 & \end{array}$$

Let $p = \{s_1, s_3, s_5\}$ and $\Pi = \{p, S \setminus p\}$. Then $B(S \cap p) = \{s_1, s_3, s_5\} \not\subseteq \{s_1, s_2, s_3, s_4\} = B(S)$ and $B(S \cap S \setminus p) = \{s_2, s_4, s_6\} \not\subseteq B(S)$.

Now suppose $\langle S, \mathscr{E}, Q, Pr, t\rangle$ satisfies ORTHOGONALITY. To show that $\Pi-$ holds, we suppose that $B(E \cap p_i) \not\subseteq B(E)$ for each $p_i \in \Pi$, and deduce a contradiction.

For each i, let $s_i \in B(E \cap p_i) \setminus B(E)$ be such that $Pr_{E\cap p_i}([s_i]_Q) \geq Pr_{E\cap p_i}([s]_Q)$ for every $s \in B(E \cap p_i) \setminus B(E)$. Since $s_i \in B(E \cap p_i)$, $Pr_{E\cap p_i}(\{s : Pr_{E\cap p_i}([s]_Q) \geq Pr_{E\cap p_i}([s_i]_Q)\} < t$. By ORTHOGONALITY, $Pr_E([s]_Q) \geq Pr_E([s_i]_Q)$ entails that either $Pr_{E\cap p_i}([s]_Q) \geq Pr_{E\cap p_i}([s_i]_Q)$ or $Pr_{E\cap p_i}([s]_Q) = 0$. So $Pr_{E\cap p_i}(\{s : Pr_E([s]_Q) \geq Pr_E([s_i]_Q)\}) = Pr_{E\cap p_i}(\{s : Pr_{E\cap p_i}([s]_Q) \geq Pr_{E\cap p_i}([s_i]_Q)\}) < t$.

Now let k be such that, for every i, $Pr_E([s_k]_Q) \geq Pr_E([s_i]_Q)$. Then $\{s : Pr_E([s]_Q) \geq Pr_E([s_k]_Q)\} \subseteq \{s : Pr_E([s]_Q) \geq Pr_E([s_i]_Q)\}$, and so $Pr_{E\cap p_i}(\{s : Pr_E([s]_Q) \geq Pr_E([s_k]_Q)\}) \leq Pr_{E\cap p_i}(\{s : Pr_E([s]_Q) \geq Pr_E([s_i]_Q)\}) < t$ for every i. But then by the law of total probability, $Pr_E(\{s : Pr_E([s]_Q) \geq Pr_E([s_k]_Q)\}) < t$, contradicting the assumption that $s_k \notin B(E)$.

□

Proposition 5. *$\Diamond-$ is valid in probability structures satisfying* STABILITY *and* ORTHOGONALITY*; but $\Box+$ can fail in such structures. Moreover, $\Diamond-$ can fail in probability structures satisfying* STABILITY *in which* ORTHOGONALITY *fails.*

Proof. To see that $\Diamond-$ holds, note that $B(E \cap p)$ is the minimal $X \subseteq E \cap p$ such that (i) if $s \in X$ and $Pr(q|E \cap p) \geq Pr([s]_Q|E \cap p)$ for $q \in Q$, then $q \subseteq X$, and (ii) $Pr(X|E \cap p) \geq t$. Then if $B(E) \cap p \neq \emptyset$, $Pr(B(E) \cap p|E \cap p) = Pr(B(E)|E \cap p) \geq t$ by STABILITY, so $B(E) \cap p$ meets condition (ii). Moreover, it meets condition (i) by ORTHOGONALITY. So $B(E) \cap p$ contains the minimal X meeting (i) and (ii) as a subset. So $B(E \cap p) \subseteq B(E) \cap p \subseteq B(E)$, as required.

To see how $\Box+$ can fail, let $S = \{a, b, c\}$, $\mathscr{E} = \{S, \{a, b\}\}$, $Q = \{\{s\} : s \in S\}$, $Pr(\{a\}) = .9$, $Pr(\{b\}) = .09$, $Pr(\{c\}) = .01$, and $t = .9001$. This structure satisfies STABILITY. $\Box+$ fails, since $B(S) = \{a, b\} \not\subseteq B(\{a, b\}) = \{a\}$.

To see how $\Diamond-$ can fail in the absence of ORTHOGONALITY, consider a probability structure in which $Q = \{A, B, C\}$, $\mathscr{E} = \{S, E\}$, $Pr(A) = \frac{1}{2}, Pr(B) = \frac{1}{4} + \varepsilon, Pr(C) = \frac{1}{4} - \varepsilon$, $Pr_E(A) = Pr_E(B) = Pr_E(C) = \frac{1}{3}$, and $t = \frac{1}{2} + \varepsilon$. STABILITY hold, but $\Diamond-$ fails: $B(S) \cap E \neq \emptyset$, but $B(E) = E \not\subseteq B(S) = A \cup B$. □

Proposition 6. *$\Box+$, $\Box-$, $\Box R$, $\Diamond R$, and $\Pi-$ are valid for LK-belief on the class of probability structures satisfying* ORTHOGONALITY

Proof. For $\Diamond R$, see the proof of Proposition 3; For $\Box+$ and $\Box-$ (and hence $\Box R$), see the proof of Proposition 7; for $\Pi-$, see [7]. □

Proposition 7. *$\Box+$ and $\Box-$ are valid for LK-Belief on the class of probability structures. $\Diamond R$ and $\Pi-$ can both fail in such structures.*

Proof. The failures of $\Diamond R$ and $\Pi-$ follow from the failure of ΠR described in the main text.

Suppose that $B_{LK}(E) \subseteq p$. We will show that $B_{LK}(E \cap p) = B_{LK}(E)$, thus establishing $\Box+$ and $\Box-$.

Since $B_{LK}(E) \subseteq p$, we have that if $s \in B_{LK}(E)$, then $[s]_Q \subseteq p$. So if $s \in B_{LK}(E)$ then $\frac{Pr([s]_Q|E\cap p)}{Pr(q|E\cap p)} \geq \frac{Pr([s]_Q|E)}{Pr(q|E)}$ for any $q \in Q$ such that $Pr(q|E \cap p) > 0$. So if $s \in B_{LK}(E)$, then for any $q \in Q$ with $Pr(q|E \cap p) > 0$, $\frac{Pr([s]_Q|E\cap p)}{Pr(q|E\cap p)} \geq \frac{Pr([s]_Q|E)}{Pr(q|E)} \geq t$. And if $Pr(q|E \cap p) = 0$, then trivially $Pr([s]_Q|E) \geq t \times Pr(q|E \cap p)$. So $s \in B_{LK}(E \cap p)$.

Moreover, if $s \notin B_{LK}(E)$, then $t > 0$ and there is a $q \in Q$ such that (i) $Pr(q|E) \times t > Pr([s]_Q|E)$ and (ii) $q \cap B_{LK}(E) \neq \emptyset$. Assuming $Pr([s]_Q|E \cap p) > 0$ then, by the above, $\frac{Pr(q|E\cap p)}{Pr([s]_Q|E\cap p)} \geq \frac{Pr(q|E)}{Pr([s]_Q|E)} > \frac{1}{t}$. In that case $s \notin B_{LK}(E \cap p)$. And if $Pr([s]_Q|E \cap p) > 0$ then also $t \times Pr(q|E \cap p) > \times Pr([s]_Q|E \cap p)$. So in that case also $s \notin B_{LK}(E \cap p)$.

So $B_{LK}(E \cap p) = B_{LK}(E)$, as required.

□

References

[1] Carlos Alchourrón, Peter Gärdenfors & David Makinson (1985): *On the Logic of Theory Change: Partial Meet Contraction and Revision Functions*. Journal of Symbolic Logic 50, pp. 510–530, doi:10.2307/2274239.

[2] Kevin Dorst & Matthew Mandelkern (2023): *Good Guesses*. Philosophy and Phenomenological Research 105(3), pp. 581–618, doi:10.1111/phpr.12831.

[3] Igor Douven & Hans Rott (2018): *From Probabilities to Categorical Beliefs: Going Beyond Toy Models*. Journal of Logic and Computation 28(6), pp. 1099–1124, doi:10.1093/logcom/exy017.

[4] Richard Foley (1993): *Working without a net: A study of egocentric epistemology*. Oxford University Press.

[5] Jeremy Goodman & Bernhard Salow (2021): *Knowledge from Probability*. In Joseph Y. Halpern & Andrés Perea, editors: Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2021, Beijing, China, June 25-27, 2021, EPTCS 335, pp. 171–186, doi:10.4204/EPTCS.335.15.

[6] Jeremy Goodman & Bernhard Salow (2023): *Epistemology Normalized*. Philosophical Review 132(1), pp. 89–145, doi:10.1215/00318108-10123787.

[7] Jeremy Goodman & Bernhard Salow (ms): *Belief revision normalized*. Unpublished manuscript.

[8] Ian Hacking (1967): *Isaac Levi, "Gambling with Truth: An Essay on Induction and the Aims of Science" (Book Review)*. Synthese 17, pp. 444–448, doi:10.2307/20114579.

[9] James Hawthorne (1996): *On the Logic of Nonmonotonic Conditionals and Conditional Probabilities*. Journal of Philosophical Logic 25(2), pp. 185–218, doi:10.1007/bf00247003.

[10] Ben Holguín (2022): *Thinking, Guessing, and Believing*. Philosophers' Imprint 22(1), pp. 1–34, doi:10.3998/phimp.2123.

[11] Kevin Kelly & Hanti Lin (2021): *Beliefs, Probabilities, and Their Coherent Correspondence*. In: Lotteries, Knowledge, and Rational Belief: Essays on the Lottery Paradox, Cambridge University Press, pp. 185–222, doi:10.1017/9781108379755.011.

[12] Sarit Kraus, Daniel Lehmann & Menachem Magidor (1990): *Nonmonotonic reasoning, preferential models, and cumulative logics*. Artificial Intelligence 44, pp. 167–207, doi:10.1016/0004-3702(90)90101-5.

[13] Hannes Leitgeb (2014): *The Stability Theory of Belief*. Philosophical Review 123, pp. 131–171, doi:10.1215/00318108-2400575.

[14] Hannes Leitgeb (2017): *The Stability of Belief: How Rational Belief Coheres with Probability*. Oxford University Press, doi:10.1093/acprof:oso/9780198732631.001.0001.

[15] Isaac Levi (1967): *Gambling with Truth: An Essay on Induction and the Aims of Science*. MIT Press.

[16] David Lewis (1996): *Elusive Knowledge*. Australasian Journal of Philosophy 74, pp. 549–567, doi:10.1080/00048409612347521.

[17] Hanti Lin & Kevin Kelly (2012): *Propositional Reasoning That Tracks Probabilistic Reasoning*. Journal of Philosophical Logic 41(6), pp. 957–981, doi:10.1007/s10992-012-9237-3.

[18] Hans Rott (2017): *Stability and scepticism in the modelling of doxastic states: Probabilities and plain beliefs*. Minds and Machines 27, pp. 167–197, doi:10.1007/s11023-016-9415-0.

[19] Ted Shear & Branden Fitelson (2019): *Two Approaches to Belief Revision*. Erkenntnis 84(3), pp. 487–518, doi:10.1007/s10670-017-9968-1.

Selling Data to a Competitor

(Extended Abstract)

Ronen Gradwohl
Department of Economics and Business Administration
Ariel University
roneng@ariel.ac.il

Moshe Tennenholtz
Faculty of Data and Decision Sciences
The Technion – Israel Institute of Technology
moshet@ie.technion.ac.il

We study the costs and benefits of selling data to a competitor. Although selling all consumers' data may decrease total firm profits, there exist other selling mechanisms—in which only some consumers' data is sold—that render both firms better off. We identify the profit-maximizing mechanism, and show that the benefit to firms comes at a cost to consumers. We then construct Pareto-improving mechanisms, in which each consumers' welfare, as well as both firms' profits, increase. Finally, we show that consumer opt-in can serve as an instrument to induce firms to choose a Pareto-improving mechanism over a profit-maximizing one.

1 Introduction

In recent years, it has become common wisdom that data is a dominant source of power. This power is perhaps most clearly illustrated in markets where an incumbent with access to consumer data competes with an entrant who does not have such data. As stated in a crisp manner by [25], common wisdom holds that the incumbent's key advantage is data superiority: "If you run a market-leading company, you should never be blindsided by an invader. Locked within your own records is a huge, largely untapped asset that no attacker can hope to match: what we call the incumbent's advantage." The situation is not uncommon: In our data-driven economy, competing firms often find themselves in asymmetric situations where one of them has superior or even exclusive access to relevant data.

Such data asymmetry has become a major issue for debate. For example, in a June 2021 press release, the EU declared that it has opened an antitrust investigation that will "examine whether Google is distorting competition by restricting access by third parties to user data for advertising purposes on websites and apps, while reserving such data for its own use," [13]. One of the issues in such debates is the question of data sharing: Should the incumbent share its data in order to increase market competition and consumer welfare? And can the incumbent profitably sell its data to the entrant?

Strategic decisions about the sale of data to competitors appear in the online economy frequently, although they are not always stated explicitly. For example, when an advertiser buys a sponsored-search campaign through an ad exchange, the advertiser obtains useful information about a segment of consumers as part of the ad exchange reports. The advertiser might then use this information when bidding directly for display ads on other platforms, including platforms on which the ad exchange is also a competitor. The data-holder (e.g., ad exchange) thus faces a strategic decision about which consumer segments to sell to a competitor, and at what prices. The data-buyer (e.g., advertiser), in turn, must decide whether to pay the price and obtain data about these consumer segments, or whether to enter into competition without the data on offer.

These strategic considerations raise numerous questions: What are the data-holder's costs and benefits from selling data to a data-buyer? What are the effects of data sale on consumer welfare? And, in a regulated market, can data sale be regulated in a way that leads to Pareto-improving transactions, benefitting consumers as well as firms?

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 318–330, doi:10.4204/EPTCS.379.26

In this paper we study these questions in a paradigmatic model of imperfect competition between two firms who have asymmetric access to data. We consider the classic Hotelling model of imperfect competition: There are two firms, each located at a different endpoint of a unit interval, with a unit mass of consumers distributed across this interval. We model data about a consumer as information about the consumer's location on the interval. In the classic model, neither firm has data about any consumer, and so firms engage in competition via uniform prices that each offers to all consumers. In our variation of this model, in contrast, one firm is a data-holder who knows the locations of all consumers, whereas the other firm is a data-buyer who has no such data. The data-holder can use its data advantage in order to personalize prices to consumers, and can thus sometimes undercut the data-buyer's uniform price.

In order to study the costs and benefits of data sale to a competitor, we suppose the firms engage in a data-sharing mechanism. Such a mechanism consists of a segment of consumers whose data the data-holder shares with the data-buyer, as well as a price the data-buyer pays the data-holder. After engaging in such a mechanism, the data-buyer will hold location data about all consumers in the shared segment, allowing that firm to also personalize prices to them.

Within this model, we first show that full data-sharing, in which the data-holder shares all its data with the data-buyer, is harmful to the firms. We then show that there exist other data-sharing mechanisms—in which only some consumers' data is shared—that increase both firms' profits. In fact, we identify the mechanism that maximizes total firm profits. This last mechanism, however, increases firm profits at the expense of consumers. We thus proceed to show that there exist Pareto-improving mechanisms, in which each consumers' welfare, as well as both firms' profits, increase. Finally, we consider the question of how a regulator can induce firms to utilize a Pareto-improving mechanism rather than a profit-maximizing one that may harm consumers. We show that consumer opt-in may serve as such an instrument: If consumers are given the opportunity to opt-in to having their data sold, and if the data-holder is only permitted to share data about consumers who have opted-in, then in equilibrium the firms will choose a Pareto-improving mechanism.

Our results are driven by two forces, which we identify as the direct effect and the indirect effect of data sharing. The direct effect is the following: if the data-holder shares data about a particular consumer, then the data-buyer can now offer that consumer a personalized rather than the uniform price. This affects both firms' equilibrium personalized prices to that consumer, and may thus impact profits and welfare. The indirect effect of data sharing, on the other hand, is the following: by sharing data about a segment of consumers, the data-holder changes the set of consumers to whom the data-buyer's uniform price applies (since additional consumers will now be offered personalized prices). And since the uniform price is determined in equilibrium in part by the locations of consumers to whom that price will apply, a change in the set of consumers may effect a change in the equilibrium uniform price, thereby affecting profits and welfare. Our results highlight how the interplay between the direct and indirect effects of data sharing lead to changes in firms' profits and consumers' welfare.

In addition to identifying the two effects of data sharing, our analysis generates several general insights. First, and perhaps surprisingly, selling data to a competitor can be strictly beneficial to both firms.[1] Second, data can be sold in a way that is Pareto improving. And finally, such Pareto-improving data-sale can be induced by consumer opt-in regulation.

We note that the idea of selling data to a competitor has been advocated in financial markets (see, for example, [2]). In that context, the possibility of data sale allows a decision maker to choose between taking investment risks or obtaining direct monetary rewards. The incredibly fast-growing data-economy, in which some firms hold massive amounts of data about consumers, raises calls to consider such data

[1] We note that this holds even if data is sold at no cost—see Proposition 2.

sale in a broader context: Can it lead to increased profits to both data-holders and data-buyers? And can it benefit all of society, including consumers whose data is exchanged?

Related literature This paper is part of a large and growing literature on data markets (see, e.g., the survey of [5]). Work in this area focuses on related but orthogonal questions, such as the effects of data-sale by a third-party data-provider and of information sharing between competitors. Our paper bridges these strands by considering data sale to a competitor. To the best of our knowledge, the work of [2] is the only other paper to study such a scenario, and ours is the first to focus on the effects of such data sale on firm profits as well as consumer welfare.

The literature on the sale of data by a data provider (e.g., [1, 6, 26, 29, 34] and others) studies how a third-party data-provider can maximize profits by selling data to a monopolist or to competing firms who use this data to price discriminate. Within this literature, one paper that is closely related to ours is that of [12]. [12] consider an information designer who provides consumer information to oligopolists, and characterize the different market outcomes that can be achieved by the designer. Our paper differs from this research in that we suppose data is not held by a third-party, but rather by one of the competing firms. This firm may sell data to its direct competitor, affecting both firms' respective market positions.

A different but related setting is that of [3], where the consumers are holders of information who may share it with one or both firms so as to intensify competition. The model and results of [3] bear some similarity to ours. For example, they also consider a Hotelling model, and show that consumers are better off whenever those sufficiently closer to one firm than another share their location with that firm, and those closer to the middle share their location with both firms. Despite the similarities, our paper studies an orthogonal question, as we assume one of the firms already has data about consumers, and focus on whether that firm will sell data to its competitor. In contrast to the model of [3], in which each consumer chooses which firm has access to that consumer's location, in our model the informed firm strategically chooses whether or not to share this information. Under consumer opt-in the role of consumers is in determining whether such sharing could potentially take place, but not in whether it actually takes place. Finally, while [3] show that, consumers are always better off when they share some of their information, we show that when firms choose what information to share this may no longer be the case.

Because our work considers the sale of data from one firm to another, it is related to the literature on information sharing. Although information sharing between firms has been studied in a variety of settings,[2] our paper is most-closely related to that of competitive price discrimination—see, for example, the surveys of [31] and [14]. One of the main insights from this literature is that when firms have more data about consumers, competition between them is more intense, leading to lower prices and profits. In our paper, in contrast, data is sold by one firm to another in such a way as to increase profits.

Two papers that specifically analyze the effects of data sharing within a Hotelling model are [24] and [7]. [24] study a model in which each of two firms may have data both about consumers' locations and about their transportation costs, and consider the eight permutations in which each firm may have either a dataset about locations, a dataset about transportation costs, both datasets, or neither datasets. They then analyze the market effects of firms sharing one or both of their (full) datasets with each other, and provide conditions under which sharing is beneficial to the firms. [7] studies a Hotelling model in which locations are two-dimensional, and firms hold all data about one dimension, both dimensions, or neither dimension. He analyzes the various scenarios in terms of firm profits and consumer welfare, with a particular emphasis on the comparison to the regimes of full privacy (neither firm has any data) and no privacy (both firms have full data). Interestingly, [7] shows that total firm profits are hump-shaped

[2] These include oligopolistic competition [11, 28], financial intermediation [27, 23, 15], supply chain management [21, 30], competition between data brokers [20, 22], and advertising [18].

in the amount of information they hold; for example, the scenario in which each firm holds data about a different dimension yields higher profits than both full privacy and no privacy. Our work differs from both of these papers in that we study the sale of partial data from one firm to another, with an emphasis on mutually increasing profits.

In terms of modeling, our paper is most closely related to [26] and [17]. [26] consider a one-dimensional Hotelling model in which consumers' locations may be known to one, both, or neither firm. Their concern is not the sale of data from one firm to another, but rather the optimal strategy of a data broker who sells the data to the firms. They also consider the effects of a consumer-side technology that allows consumers the ability to protect their privacy. [17] also study a Hotelling model, but suppose that both firms have some data about consumers. Their main focus is on various forms of mutual data sharing between the firms.

2 The Model

We focus on a standard Hotelling model, in which a unit mass of consumers is spread over the unit interval according to an atomless distribution F with continuous, strictly positive density f that has full support. There are two firms: firm A is located at $\theta_A = 0$, and firm B is located at $\theta_B = 1$. Each consumer chooses at most one firm from which to purchase a good. Consumers derive value v from the good, but pay two costs: the price, and a linear transportation cost that scales with the distance between the consumer and the firm providing the good. Thus, a consumer located at θ who buys from firm i at price p_i obtains utility $v - p_i - t\,|\theta - \theta_i|$, where t is the marginal transportation cost. We assume throughout that the market is covered—namely, that $v > 2t$—so that all consumers purchase a good even when there is a monopolist firm. Finally, we also assume for simplicity that firms' marginal costs are 0, and so their profit from the sale of a good is equal to the price. These are all standard assumptions in Hotelling games.

The standard setup consists of a two-stage game: First, firms simultaneously set prices; second, consumers choose a firm and make a purchase. In the simple case where the distribution F of consumers is uniform the game has a unique subgame perfect equilibrium: firms' prices are $p_A = p_B = t$, consumers in $[0, 0.5)$ buy from A, and consumers in $(0.5, 1]$ buy from B (see, e.g., [4]).[3]

In this paper we will consider a variant of the standard model by supposing that firms may have additional information about some of the consumers. In particular, we will suppose that, for each consumer, one or both firms know the location of that consumer on the unit interval. For such consumers, firms will be able to offer a *personalized price*—a special offer specifically tailored to that consumer. If a firm does not know a consumer's location, however, then it cannot distinguish between that consumer and all other consumers whose location it does not know. All such consumers are offered the same *uniform price*.

In our model, firm B is the data-holder and firm A is the data-buyer. Thus, initially, we assume that firm B knows the locations of all consumers, whereas firm A does not know any consumer's location. Given this informational environment, a data-sharing mechanism $M = (M_B, r)$ between firms specifies a subset $M_B \subseteq [0, 1]$ and a number $r \in \mathbb{R}$, with the interpretation that firm B shares with firm A the locations of consumers in M_B, and firm A transfers to firm B a payment r. Two simple examples of data-sharing mechanisms are one that involves *no sharing*, $M = (\emptyset, r)$, and one that involves *full sharing*, $M = ([0, 1], r)$. Alternatively, firm B may share data about a subset of consumers. For example, under mechanism $([x, y], r)$, if consumer $\theta \in [x, y]$ arrives, both firms will know that consumer's location. On the other hand, if consumer $\theta \in [0, 1] \setminus [x, y]$ arrives, firm B will know that consumer's location, and firm A will only be able to deduce that the consumer is not located within $[x, y]$.

[3] The equilibrium is unique up to the choice of the indifferent consumer located at $\theta = 0.5$.

In our analysis, we consider the following order of events:

1. Firms engage in a data-sharing mechanism $M = (M_B, r)$.
2. Firm A announces uniform price p_A.[4]
3. A consumer arrives, and all firms who know the consumer's location θ simultaneously offer that consumer a personalized price, $p_A(\theta)$ and $p_B(\theta)$.
4. The consumer chooses a firm from which to buy, and payoffs are realized.

Note that firms share data, and firm A announces its uniform prices, before consumers arrive. After a consumer arrives to the market, the firms who know the consumer's specific location simultaneously offer personalized prices. If firm A offers a consumer a personalized price, this offer subsumes the firm's original uniform price. Thus, the uniform price p_A will apply only to those consumers who will not subsequently be offered a personalized price by firm A.

Importantly, when firms set personalized prices, they know the uniform price set by firm A in the previous stage. This is the standard timing considered in the literature (see, e.g., [33, 10, 9, 26, 8]).[5]

For any fixed mechanism M, we will consider the pure subgame perfect equilibria of the game that starts with data-sharing mechanism M. Such equilibria always exists, and consist of a uniform price for firm A followed by personalized prices for both firms. Once the uniform price is fixed, the equilibrium personalized prices for each consumer θ are uniquely fixed. We will be interested in designing mechanisms M that lead to equilibria with high firm-profits and high consumer-welfare.

One important desideratum of data-sharing mechanisms (with corresponding equilibria) is that they be *individually rational (IR)*: That the expected utility of each firm with data sharing be at least as high as without data sharing. A data-sharing mechanism should be IR if we expect firms to participate.

Our main focus will be on mechanisms that are not only IR, but also *Pareto-improving*: when sharing takes place, (i) the expected utility of each firm and *every* consumer be at least as high as without data sharing, and that (ii) either firm A's profits, firm B's profits, or total consumer welfare be strictly higher.

We note that many of our results make no assumptions about the distribution of consumers. In such a general setting there may be multiple equilibria, even with no data-sharing, each with different uniform prices. Hence, we will often describe mechanisms as being IR or Pareto-improving *relative to* a particular no-sharing equilibrium.

3 No Data-Sharing

We begin by analyzing equilibria under no data-sharing. To this end, define $\mu(p_A) = \frac{1}{2} - \frac{p_A}{2t}$. If firm A charges uniform price p_A, then the consumer located at $\mu(p_A)$ is indifferent between purchasing from firm A at that price and purchasing from firm B at price 0. All consumers located to the left of $\mu(p_A)$ will thus strictly prefer purchasing from firm A at price p_A than from firm B at any nonnegative price. In contrast, for every consumer located to the right of $\mu(p_A)$ there exists a nonnegative price of firm B such that that consumer will prefer to purchase from B than from A. This is formalized in the following proposition:

[4] Note that firm B knows all consumers' locations, and so personalizes prices to each. It therefore need not post a uniform price.

[5] An alternative model that we do not analyze is one in which firms set uniform and personalized prices simultaneously, for each consumer. [26] show that, in this case, a (pure) equilibrium may fail to exist.

Proposition 1 *Let $P_A = \arg\max_p p \cdot F(\mu(p_A))$. Without data sharing, the set of equilibria consist of any uniform price $p_A \in P_A$ for firm A and corresponding personalized prices $p_B(\theta) = \max\{0, p_A + t(2\theta - 1)\}$ for firm B. In the equilibrium with uniform price $p_A \in P_A$, consumers in $[0, \mu(p_A))$ purchase from firm A, whereas consumers in $[\mu(p_A), 1]$ purchase from B. The equilibrium with $p_A = \max\{P_A\}$ is strictly dominant for the firms.*

The proof of Proposition 1, and all other propositions, appear in the full version of this paper [19].

Throughout the paper we will illustrate our results with the simple case in which consumers are uniformly distributed on $[0,1]$. We note that this is the standard setup in Hotelling games.

Example 1 *When consumers are uniformly distributed on $[0,1]$, the set $P_A = \{t/2\}$. In the unique equilibrium, then, consumers between 0 and $\mu(t/2) = 1/4$ purchase from A at uniform price $p_A = t/2$, whereas the rest purchase from B at personalized prices $p_B(\theta) = \max\{t(2\theta - 1/2), 0\}$. Total firm profits are $\pi_A = t/8$ and*

$$\pi_B = \int_{1/4}^{1} t(2\theta - 1/2)d\theta = \frac{9t}{16},$$

whereas consumer welfare is

$$CW = \int_0^1 \max\{v - \theta t - p_A, v - t(1-\theta) - p_B(\theta)\}d\theta = \int_0^1 (v - t/2 - \theta t)d\theta = v - t.$$

4 The Direct Effect and Full Data-Sharing

In this section we begin our analysis of how data-sharing impacts profits and welfare. Data sharing has a direct effect and an indirect effect. The direct effect is that if firm A obtains information about a consumer's locations via the sharing mechanism, it can now offer that consumer a personalized price. This affects firm B's equilibrium personalized price to that consumer, and hence also profits and welfare. The indirect effect of data sharing is that it may change the set of consumers to whom firm A's uniform price applies, since additional consumers will now be offered personalized prices. And since the uniform price is determined in equilibrium in part by the locations of consumers to whom that price will apply, a change in the set of consumers may effect a change in the equilibrium uniform price. In this section we explore the direct effect, and then in Section 5 we explore the indirect effect.

Suppose that, absent data-sharing, firm A's uniform price is p_A. If firm B shares the location θ of some consumer with A, then the firms compete in personalized prices over that consumer, yielding equilibrium prices $p_A(\theta) = \max\{t(1-2\theta), 0\}$ and $p_B(\theta) = \max\{t(2\theta - 1), 0\}$. The direct effect of firm B sharing the location of a consumer is summarized in Lemma 1:

Lemma 1 *Consider mechanism $M = (\{\theta\}, 0)$ relative to no sharing, and suppose that consumer θ shows up.*

1. *If $\theta \in (1/2, 1]$, consumer θ still buys from B, but now at price $t(2\theta - 1)$. This is a net loss of p_A to firm B and a net gain of p_A to the consumer.*
2. *If $\theta \in [\mu(p_A), 1/2)$, consumer θ switches to purchasing from A, at price $t(1 - 2\theta)$. This is a loss of $p_A + t(2\theta - 1)$ to firm B, a gain of $t(1 - 2\theta)$ to firm A, and a gain of $p_A - t(1 - 2\theta) \geq 0$ to the consumer. Also, the gain to A is greater than the loss to B if and only if*

$$\theta < \frac{1}{2}\left(\mu(p_A) + \frac{1}{2}\right),$$

the midpoint of the interval of θ-s in the case under consideration.

3. *If $\theta \in [0, \mu(p_A))$, consumer θ still buys from A, now at personalized price $p_A(\theta) = t(1-2\theta) > p_A$.*

Given these direct effects, we now consider full data-sharing, namely, $M = ([0,1], r)$ for some r. Under this mechanism, both firms know the location of every consumer, and so both engage in personalized pricing. Firm A's uniform price thus applies to no consumer, and so only the direct effect has any bite. By Lemma 1, relative to the no-sharing mechanism with price p_A, consumers $\theta \in [\mu(p_A), 1]$ are better off, whereas consumers $\theta \in [0, \mu(p_A))$ are worse off, under full data-sharing. For the firms, naturally firm B is better off with no sharing and firm A with full sharing. The effect on total profits, however, depends on the distribution F. For the case of uniformly distributed consumers, full data-sharing harms firms:

Example 2 *When consumers are uniformly distributed, [32] show that profits are $\pi_A = \pi_B = t/4$ (see also [33]). Note that total profits $\pi_A + \pi_B$ are higher under no sharing ($t/8 + 9t/16 = 11t/16$, by Example 1) than under full sharing ($t/4 + t/4 = t/2$). This implies that no mechanism $([0,1], r)$ is IR, regardless of r.*

Although full data-sharing decreases total firm profits when consumers are uniformly distributed, there exist distributions of consumers under which full data-sharing increases profits—for example, this is the case when a $(1-\varepsilon)$-fraction of consumers are uniformly distributed on the sub-interval $[0, 1/4]$, and the remaining ε on $(1/4, 1]$, for some small enough $\varepsilon > 0$.[6] However, even then full sharing does not lead to *maximal* profits. We now turn to mechanisms that do.

5 The Indirect Effect and Firm-Optimal Data-Sharing

In this section we describe firm-optimal mechanisms, which exploit the *indirect* effect of data sharing. By Lemma 1, firm B's profit from a consumer $\theta \in (1/2, 1]$ is $p_A + t(2\theta - 1)$. If A's uniform price were to increase, this would likewise increase B's profit from consumer θ. Now, recall that, when there is no sharing, firm A sets its uniform price by choosing $p_A \in P_A = \arg\max_p p \cdot F(\mu(p))$. If B were to share data about consumers in some interval $[\underline{\theta}, \mu(p_A)]$, however, then A would offer consumers on this interval a personalized price. The uniform price would no longer apply to them, but would instead apply only to consumers $[0, \underline{\theta}) \cup (\mu(p_A), 1]$. Firm A may then benefit from increasing (decreasing) the uniform price above (below) p_A, at the same time increasing (decreasing) the profits of firm B from consumers $\theta \in (1/2, 1]$. This is the indirect effect of data sharing.

Firm B can exploit both the indirect and direct effects of data sharing by sharing data both about consumers in $[\underline{\theta}, \mu(p_A)]$ and about consumers in $(\mu(p_A), 1/2]$. Note, however, that sharing data about consumers in $(1/2, 1]$ is never beneficial, since it only results in a net loss to firms and has no indirect effect (by Lemma 1, above).

In general, the firm-optimal mechanism may depend on the distribution of consumers and other primitives of the model. In Proposition 2, however, we show that when v (the consumers' value for the good) is sufficiently high, then there is an essentially unique mechanism, with a corresponding equilibrium, that yield the firms maximal joint profits. The mechanism makes extreme use of the indirect effect of data sharing: Firm B shares data about consumers $[0, 1/2]$, implying that firm A's uniform price no longer applies to these consumers, and hence that this price can be almost arbitrarily high. A's uniform price does apply to consumers in $(1/2, 1]$, for whom it serves as an outside option. However, because these

[6] Such a consumer distribution does not satisfy our continuity assumption. However, the same result holds also when the kink at $1/4$ is smoothed out.

consumers will always purchase from B in equilibrium (by Lemma 1, above), the high outside option allows that firm to extract these consumers' entire surplus.

Proposition 2 *Fix $v > \frac{5t}{2(1-F(1/2))}$. Mechanism $M = ([0,1/2],0)$ with equilibrium uniform price $p_A = v - t/2$ maximizes joint firm profits and is IR relative to any no-sharing equilibrium. Every other firm-optimal mechanism is of the form $M' = ([0,1/2],r)$.*

Example 3 *When consumers are uniformly distributed, the mechanism described in Proposition 2 is actually firm-optimal for all $v > 2t$, as we now show. This mechanism leads to profits $\pi_A = t/4$ and*

$$\pi_B = \int_{1/2}^{1} (v - t(1-\theta))d\theta = \frac{v}{2} - \frac{t}{8} > \frac{7t}{8},$$

where the inequality follows since $v > 2t$. Thus, total profits are at least $9t/8$. In contrast, consider any mechanism $M' = (M_B, r)$ in which firm A's uniform price applies to a consumer in $[0,1/2]$, and fix some uniform price $p'_A \leq t$. Consumers $\theta \in (1/2,1]$ buy from B, leading to total profits at most $3t/4$ from these consumers. Consumers $\theta \in [0,1/2]$ either buy from A at uniform price p'_A or at personalized price $t(1-2\theta)$, or from B at personalized price $t(2\theta-1)$ or $p'_A + t(2\theta-1)$ (depending on whether $\theta \in M_B$). Total profits are maximized when $p'_A = t$, consumers $\theta \in [0,1/4]$ buy at A's personalized price, and the rest buy from B at price $t + t(2\theta-1)$. Profits to A from $[0,1/4]$ and to B from $(1/4,1]$ are each equal to $3t/16$. Total profits from M' are thus bounded above by $3t/4 + 2(3t/16) = 9t/8$, which is equal to the lower bound on profits from $([0,1/2],0)$.

Remark 1 Proposition 2 provides a sufficient condition under which mechanism $M = ([0,1/2],0)$ is firm-optimal for *some* equilibrium (namely, the one with uniform price $p_A = v - t/2$). However, under this mechanism there are other equilibria, which involve lower uniform prices, and that yield lower firm profits. In Proposition 6 in the full version of this paper [19] we describe a different mechanism with $r = 0$ that, while not firm-optimal, yields both firms strictly higher profits than under no sharing in *every* equilibrium.

6 Pareto-Improving Data-Sharing

In Section 5 above we show that firm B can sell data in a way that maximizes joint firm profits, and hence allows that firm to charge a high price for the data. Such sharing, however, comes at the expense of consumers. In particular, under the equilibrium of Proposition 2, firms extract the entire surplus of consumers located in $[1/2,1]$. In this section we show that there exist other data-sharing mechanisms that increase firm profits relative to no sharing, while at the same time also increasing consumers' utilities.

Recall that $P_A = \arg\max_p p \cdot F(\mu(p_A))$, and that, by Proposition 1, the set of equilibria under no data-sharing consist of uniform prices $p_A \in P_A$ by firm A and respective personalized prices $p_B(\theta) = \max\{0, p_A + t(2\theta-1)\}$ by firm B. Denote by $E(p_A)$ the no-sharing equilibrium with uniform price p_A. In the following proposition we show that for each such no-sharing equilibrium there exists a Pareto-improving mechanism.

Proposition 3 *For every $p_A \in P_A$ there exists r such that mechanism $M = \left(\left[\mu(p_A), \frac{1}{4} + \frac{\mu(p_A)}{2}\right], r\right)$ with uniform price p_A is IR and weakly beneficial to every consumer, relative to $E(p_A)$. Furthermore, M yields higher total firm profits than any other mechanism that is weakly beneficial to every consumer relative to $E(p_A)$.*

The mechanism M described in Proposition 3 does not decrease the utility of any consumer. Moreover, by bullet 2 of Lemma 1, that mechanism *strictly increases* the utilities of a subset of consumers—namely, those located in $\left(\mu(p_A), \frac{1}{4} + \frac{\mu(p_A)}{2}\right]$.

The main idea underlying the construction for Proposition 3 is that firm B shares data about every consumer θ that satisfies two conditions: (i) with no sharing, consumer θ prefers to pay B's personalized price than A's uniform price; (ii) sharing consumer θ's location leads to a net increase in firm profits. Note that these consumers are all closer to A than to B, so that the welfare and profit gain is obtained due to an increase in efficiency. Finally, the construction is such that A's uniform price under M remains the same as with no sharing, which guarantees that consumers close to A do not pay a higher price than under no sharing, but also that firms maximize their joint profits subject to this constraint.

7 Consumer Opt-In

Proposition 3 above shows that there exist mechanisms that are strictly Pareto-improving, increasing firm profits as well as consumer welfare. However, these mechanisms are not optimal for firms—Proposition 2 identifies a different mechanism as maximizing firm profits, a mechanism that does so at the expense of consumers. How can a policymaker induce firms to share data in a Pareto-improving manner, rather than in a profit-maximizing manner? In this section we identify one way in which a policymaker can do this: by asking each consumer whether or not they agree to have their data shared, and then permitting firms to share data only about consumers who have agreed.

In order to analyze such consumer opt-in regulation, we first extend the model to include a preliminary opt-in stage. After setting up the model, we present two results. The first, Proposition 4 in Section 7.2, states that, under consumer opt-in, there is an equilibrium of the extended model wherein firms choose the Pareto-improving mechanism of Proposition 3. Now, although consumer opt-in can lead to the choice of the Pareto-improving mechanism, there are other equilibria that do not. However, in our second result here—Proposition 5 in Section 7.3—we show that the equilibrium of Section 7.2, where the Pareto-improving mechanism is chosen, is, in a sense, optimal for the consumers.

7.1 The Extended Model

We begin by extending the model of Section 2 with a preliminary stage, in which each consumer simultaneously chooses whether or not to opt in to having location data shared. Denote the set of consumers who opted in as C. Only then do firms engage in a data-sharing mechanism $M = (M_B, r)$; however, firms are restricted to choosing a mechanism for which $M_B \subseteq C$. Such mechanisms are *feasible for C*.

We assume that firms bargain over the choice of mechanism efficiently—that is, they choose a mechanism M that maximizes total firm profits, subject to the opt-in constraint. One way to implement such efficient bargaining is when one of the firms makes the other a take-it-or-leave-it offer by suggesting a mechanism (M_B, r) that is feasible for C. Depending on which firm makes the offer, the chosen price transfer r will vary to favor the offering firm. Either way, however, firms will choose to offer a mechanism that maximizes joint firm-profits. This assumption is stated formally in Definition 1 below as part of the solution concept. In addition, as in the previous sections, we assume that, absent data-sharing, firms play the no-sharing equilibrium $E(p_A)$ for some $p_A \in P_A$.

In this extended model there is an additional, technical complication. We are assuming that firms choose a mechanism that is feasible for some C. However, since C is generated by the set of consumers who choose to opt in, it may not be a measurable set. Thus, the firms' optimization problem may not be well-defined at every C. One way to get around this problem is to consider the Nash equilibria of this extensive-form game (rather than the subgame perfect equilibria). However, this is somewhat unsatisfying, as such equilibria may be sustained by strange off-equilibrium behavior—namely, the presence of empty threats. Instead, we will use an equilibrium notion that is weaker than subgame perfect equilibrium but nonetheless suffices to eliminate empty threats. The general definition is due to [16]; here we give a specialized version that applies to our specific game.

For the definition, let L denote the set of subsets of $[0,1]$, and let $\mathscr{M}(C)$ denote the set of all mechanisms feasible for C.

Definition 1 *A set $C^* \in L$ and functions $m : L \to \mathscr{M}(L)$ and $p : L \to \mathbb{R}_+$ form a* threat-free Nash equilibrium (TFNE) *if*

1. *For every C, mechanism $m(C)$ is feasible for C, and $p(C)$ is an equilibrium uniform price for firm A under $m(C)$.*
2. *For every $\theta \in C^*$, consumer θ is weakly better off under $m(C^*)$ than under $m(C^* \setminus \{\theta\})$ (with respective uniform prices $p(C^*)$ and $p(C^* \setminus \{\theta\})$).*
3. *For every $\theta \in [0,1] \setminus C^*$, consumer θ is weakly better under $m(C^*)$ than under $m(C^* \cup \{\theta\})$ (with respective uniform prices $p(C^*)$ and $p(C^* \cup \{\theta\})$).*
4. *For every $\theta \in [0,1]$ and $C \in \{C^* \setminus \{\theta\}, C^* \cup \{\theta\}\}$, mechanism $m(C)$ with uniform price $p(C)$ is IR and jointly firm-optimal relative to all mechanisms that are feasible for C (with corresponding uniform prices).*

For comparison, in a Nash equilibrium bullet 4 would be replaced by requiring IR and joint firm-optimality only for C^*. In a subgame perfect equilibrium, in contrast, bullet 4 would require these for all sets C. A TFNE is a compromise between the two, requiring IR and joint firm-optimality for C^* and for all sets C that differ from C^* by a single consumer's unilateral deviation.

7.2 Pareto-Improving Equilibrium

Given the extended model above, we can now state our proposition on the benefit of consumer opt-in.

Proposition 4 *For every $p_A \in P_A$ there exists a TFNE (C^*, m, p) of the extended model in which $m(C^*)$ is the Pareto-improving mechanism $M = \left(\left[\mu(p_A), \frac{1}{4} + \frac{\mu(p_A)}{2}\right], r\right)$, for some r.*

Proposition 4 shows that, when consumers can choose whether or not to opt in to having their data shared, and firms are allowed to only share the data of consumers who have opted in, then the equilibrium mechanism is Pareto improving. There are other equilibria that lead to a Pareto-improving mechanism. In fact, as long as consumers $\theta \in [0, \mu(p_A))$ do *not* opt in to having their data shared, the mechanism that maximizes firms' profits will be Pareto improving.

However, there are also other equilibria in which the chosen mechanism is not Pareto improving. Consider the following strategies: Consumers $[0,1/2]$ opt in to having their data shared, and firms choose the firm-optimal mechanism $M = ([0,1/2], 0)$ from Proposition 2. If some consumer $\theta \in [0,1/2]$ does not opt in, then firms use the mechanism $M_\theta = ([0,1/2] \setminus \{\theta\}, 0)$. This mechanism is identical to M, except that consumer θ faces firm A's uniform price $p_A = v - t/2$ rather than the personalized price $p_A(\theta) = t(1-2\theta)$. This is no better for consumer θ, and so these strategies form an equilibrium. Why, then, would consumers choose to collectively opt in as in Proposition 4?

7.3 Consumer-Optimal Equilibrium

We now show that the equilibrium of Proposition 4 is focal for the consumers. In particular, we show that it maximizes consumer welfare, relative to all other equilibria that leave no consumer worse off.

Fix some $p_A \in P_A$, and observe that there always exists a TFNE of the extended game in which no consumer opts in, and that this leads to consumer utilities as derived from equilibrium $E(p_A)$ in mechanism $M_\emptyset = (\emptyset, 0)$. Next, let us consider other opt-in choices for consumers. For a set C and mechanism M feasible for C, say that M is *Pareto-improving for the consumers* if the resulting utility of every consumer is weakly higher than under $M_\emptyset$. We now show that consumers' utilities in the equilibrium of Proposition 4 are optimal:

Proposition 5 *Fix $C \subseteq [0,1]$ and a mechanism M with uniform price q_A that is feasible for C and that is Pareto-improving for the consumers. If M yields strictly higher total utility to the consumers than M^*, then M will not be chosen by the firms in any TFNE in which consumers C opt in.*

That is, if we assume consumers make their opt-in decisions in a way that leads to a weak improvement for each, then they can do no better than the opt-in strategy of Proposition 4.

8 Conclusion

In this paper we analyzed the benefits to a data-holder of selling consumer data to a data-buyer in a Hotelling model of imperfect competition. We identified the two effects of data sharing, and showed that the interplay of these effects can lead to Pareto-improving mechanisms that benefit consumers as well as firms. Finally, we showed that consumer opt-in can induce firms to choose such a Pareto-improving mechanism.

Acknowledgements

The work of Tennenholtz was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant number 740435).

References

[1] Anat R Admati & Paul Pfleiderer (1986): *A monopolistic market for information*. Journal of Economic Theory 39(2), pp. 400–438, doi:10.1016/0022-0531(86)90052-9.

[2] Anat R Admati & Paul Pfleiderer (1988): *Selling and trading on information in financial markets*. The American Economic Review 78(2), pp. 96–103.

[3] S Nageeb Ali, Greg Lewis & Shoshana Vasserman (forthcoming): *Voluntary disclosure and personalized pricing*. Review of Economic Studies, doi:10.1093/restud/rdac033.

[4] Paul Belleflamme & Martin Peitz (2015): *Industrial organization: markets and strategies*. Cambridge University Press, doi:10.1017/CBO9781107707139.

[5] Dirk Bergemann & Alessandro Bonatti (2019): *Markets for information: An introduction*. Annual Review of Economics 11, pp. 85–107, doi:10.2307/1912702.

[6] Dirk Bergemann, Alessandro Bonatti & Alex Smolin (2018): *The design and price of information*. American Economic Review 108(1), pp. 1–48, doi:10.1016/j.jet.2011.06.003.

[7] Francesco Clavorà Braulin (2023): *The effects of personal information on competition: Consumer privacy and partial price discrimination*. International Journal of Industrial Organization 87, p. 102923, doi:10.1016/j.ijindorg.2023.102923.

[8] Zhijun Chen, Chongwoo Choe & Noriaki Matsushima (2020): *Competitive personalized pricing*. Management Science 66(9), pp. 4003–4023, doi:10.1287/mksc.1100.0607.

[9] Chongwoo Choe, Stephen King & Noriaki Matsushima (2018): *Pricing with cookies: Behavior-based price discrimination and spatial competition*. Management Science 64(12), pp. 5669–5687, doi:10.1287/mksc.1100.0607.

[10] Vidyanand Choudhary, Anindya Ghose, Tridas Mukhopadhyay & Uday Rajan (2005): *Personalized pricing and quality differentiation*. Management Science 51(7), pp. 1120–1130, doi:10.1509/jmkr.37.3.292.18777.

[11] Richard N Clarke (1983): *Collusion and the incentives for information sharing*. The Bell Journal of Economics, pp. 383–394, doi:10.2307/3003640.

[12] Matthew Elliott, Andrea Galeotti, Andrew Koh & Wenhao Li (2021): *Market segmentation through information*. Available at SSRN 3432315, doi:10.2139/ssrn.3432315.

[13] European Commission (2021): *Antitrust: Commission opens investigation into possible anticompetitive conduct by Google in the online advertising technology sector*. Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3143.

[14] Drew Fudenberg & J Miguel Villas-Boas (2012): *Price discrimination in the digital economy*. The Oxford handbook of the digital economy, pp. 254–272.

[15] Thomas Gehrig & Rune Stenbacka (2007): *Information sharing and lending market competition with switching costs and poaching*. European Economic Review 51(1), pp. 77–99, doi:10.1016/j.euroecorev.2006.01.009.

[16] Ronen Gradwohl, Noam Livne & Alon Rosen (2013): *Sequential rationality in cryptographic protocols*. ACM Transactions on Economics and Computation (TEAC) 1(1), pp. 1–38, doi:10.1007/978-3-642-00457-5_3.

[17] Ronen Gradwohl & Moshe Tennenholtz (2022): *Pareto-Improving Data-Sharing*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, doi:10.1145/3298981.

[18] Ronen Gradwohl & Moshe Tennenholtz (2023): *Coopetition against an Amazon*. Journal of Artificial Intelligence Research 76, pp. 1077–1116, doi:10.1613/jair.1.14074.

[19] Ronen Gradwohl & Moshe Tennenholtz (2023): *Selling Data to a Competitor*. SSRN 4343686, doi:10.2139/ssrn.4343686. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4343686.

[20] Yiquan Gu, Leonardo Madio & Carlo Reggiani (2022): *Data brokers co-opetition*. Oxford Economic Papers 74(3), pp. 820–839, doi:10.1093/oep/gpab042.

[21] Albert Y Ha & Shilu Tong (2008): *Contracting and information sharing under supply chain competition*. Management Science 54(4), pp. 701–715, doi:10.1111/j.1937-5956.2002.tb00476.x.

[22] Shota Ichihashi (2021): *Competing data intermediaries*. The RAND Journal of Economics 52(3), pp. 515–537, doi:10.1111/1756-2171.12382.

[23] Tullio Jappelli & Marco Pagano (2002): *Information sharing, lending and defaults: Cross-country evidence*. Journal of Banking & Finance 26(10), pp. 2017–2045, doi:10.1016/S0378-4266(01)00185-6.

[24] Nicola Jentzsch, Geza Sapi & Irina Suleymanova (2013): *Targeted pricing and customer data sharing among rivals*. International Journal of Industrial Organization 31(2), pp. 131–144, doi:10.1016/j.ijindorg.2012.11.004.

[25] Ian Macmillan & Larry Selden (2008): *The incumbent's advantage*. Harvard Business Review 86(10), pp. 111–121.

[26] Rodrigo Montes, Wilfried Sand-Zantman & Tommaso Valletti (2019): *The value of personal information in online markets with endogenous privacy*. Management Science 65(3), pp. 1342–1362, doi:10.1016/0022-0531(84)90162-5.

[27] Marco Pagano & Tullio Jappelli (1993): *Information sharing in credit markets*. The Journal of Finance 48(5), pp. 1693–1718, doi:10.2307/2329064.

[28] Michael Raith (1996): *A general model of information sharing in oligopoly*. Journal of economic theory 71(1), pp. 260–288, doi:10.1006/jeth.1996.0117.

[29] Carlos Segura-Rodriguez (2021): *Selling Data*. PIER Working Paper No. 19-006. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3385500.

[30] Noam Shamir & Hyoduk Shin (2016): *Public forecast information sharing in a market with competing supply chains*. Management Science 62(10), pp. 2994–3022, doi:10.1287/mnsc.1040.0226.

[31] Lars A Stole (2007): *Price discrimination and competition*. Handbook of industrial organization 3, pp. 2221–2299, doi:10.1016/S1573-448X(06)03034-2.

[32] Curtis Taylor & Liad Wagman (2014): *Consumer privacy in oligopolistic markets: Winners, losers, and welfare*. International Journal of Industrial Organization 34, pp. 80–84, doi:10.1016/j.ijindorg.2014.02.010.

[33] Jacques-Francois Thisse & Xavier Vives (1988): *On the strategic choice of spatial price policy*. The American Economic Review, pp. 122–137.

[34] Kai Hao Yang (2022): *Selling consumer data for profit: Optimal market-segmentation design and its consequences*. American Economic Review 112(4), pp. 1364–93, doi:10.1016/0304-4068(87)90007-3.

An Acceptance Semantics for Stable Modal Knowledge
Extended Abstract

Peter Hawke
Philosophy Department, Lingnan University, Hong Kong
peterhawke@ln.edu.hk

We observe some puzzling linguistic data concerning ordinary knowledge ascriptions that embed an epistemic (im)possibility claim. We conclude that it is untenable to jointly endorse both classical logic and a pair of intuitively attractive theses: the thesis that knowledge ascriptions are always veridical and a 'negative transparency' thesis that reduces knowledge of a simple negated 'might' claim to an epistemic claim without modal content. We motivate a strategy for answering the trade-off: preserve veridicality and (generalized) negative transparency, while abandoning the general validity of contraposition. We survey and criticize various approaches for incorporating veridicality into *domain semantics*, a paradigmatic 'information-sensitive' framework for capturing negative transparency and, more generally, the non-classical behavior of sentences with epistemic modals. We then present a novel information-sensitive semantics that successfully executes our favored strategy: *stable acceptance semantics*.

1 *Introduction*

In this paper, we are concerned with the semantics and logic of ordinary knowledge ascriptions that embed an epistemic (im)possibility claim.

(1) Ann knows that it might be raining.

(2) Ann knows that it can't be raining.

It is natural to interpret the modals here as having an *epistemic* flavor. Intuitively, (1) communicates (perhaps *inter alia*) that Ann's knowledge leaves it open that it is raining; (2) communicates (perhaps *inter alia*) that Ann's knowledge rules out that it is raining. In support, notice how jarring the following sound:

(3) # Ann knows that it might be raining and Ann knows that it isn't raining.

(4) # Ann knows that it can't be raining and for all Ann knows, it is raining.

Note that (1) and (2) also provide evidence of the systematic *shiftiness* of ordinary epistemic modals. Compare a bare might claim:

(5) It might be raining.

In this case, the modal is most naturally taken to communicate that the knowledge of the *speaker* (who need not be Ann) leaves it open that that is raining. As evidence, note the incoherence of the following so-called (and much discussed) *epistemic contradiction* (cf. [22],[23]).

(6) # It might be raining and it isn't raining.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 331–343, doi:10.4204/EPTCS.379.27

The first aim of the present paper is to highlight some unusual and subtle logical features that attitude ascriptions like (1) and (2) plausibly display (§2 and §3), in particular in interaction with bare modal claims like (5). The second aim is to propose a novel formal semantics that successfully predicts these features (§5), in contrast to a salient rival theory (§4). The resulting theory is of linguistic, technical, and philosophical interest. On the linguistic side, we combine novel and known linguistic data to motivate a new entry in the tradition of 'information-sensitive' semantics for ordinary epistemic modals (cf. [21], [22], [23], [17], [12], [13], [1]), extending a standard 'state-based' account with a novel semantics for knowledge ascriptions. On the technical side, our system displays intriguing and striking non-classical logical behavior, motivating a fuller technical study of the underlying epistemic logic and its interactions with modals (cf. [6], [19],[26]). On the philosophical side, our semantics may be viewed as a new development in the expressivist tradition for epistemic vocabulary (cf. [24]) that treats assertion conditions as primary in semantics (cf. [20]).

2 *Linguistic Evidence for Transparency and Veridicality*

We work with formal language $\mathscr{L}$, intended to formalize the relevant fragment of declarative English. We use φ and ψ for arbitrary formulas. Intuitively, read $K_a\varphi$ as 'Agent a knows that φ' (with $a \in \{1, 2, \ldots, n\}$) and read $\diamond\varphi$ as 'It might be that φ'. We take atoms p and q to be declaratives without logical vocabulary (we include $\diamond$ in the logical vocabulary). We use $\vdash$ to denote entailment and $\equiv$ for logical equivalence, relative to our intended reading of $\mathscr{L}$. With this in mind, there are reasons to the think that the following principles are sound, and should be recovered by a formal semantics that aims to honor our intended reading of $\mathscr{L}$.

Negative Transparency (NTrans): $K_a\neg\diamond p \equiv K_a\neg p$
K-veridicality (Ver): $K_a\varphi \vdash \varphi$

As evidence, note that the following bare assertions (easily multiplied) have an air of incoherence.

(7) # Ann knows that Bob can't be here but, for all she knows, he is. (cf. (3))

(8) # Ann knows that Bob isn't here but, for all she knows, he might be.

(9) # Bob can't be here, but Ann mistakenly knows that he might be.

Compare (9) to the benign 'Bob can't be here, but Ann mistakenly believes that he might be'. **NTrans** predicts that (7) and (8) are contradictory; **Ver** predicts that (9) is contradictory.

As further evidence, note the difficulty in distinguishing the information communicated by the following in conversation:

(10) # For all Ann knows, Bob is here.

(11) # For all Ann knows, Bob might be here.

(10) and (11) seem to say the same thing: nothing that Ann knows rules out that Bob is here. Assuming that 'for all Ann knows, φ' is formalizable as '$\neg K\neg\varphi$', **NTrans** predicts this equivalence, as it entails (with minimal further assumptions) that $\neg K\neg p$ is equivalent to $\neg K\neg\diamond p$.

Observations of the above sort are not without precedent. **Ver** is orthodox (though it is notable, as (9) seems to demonstrate, that **Ver** is undisturbed by modal content). **NTrans** is related to Łukasiewicz' principle (i.e., $\neg p \vdash \neg\diamond p$), which is in turn related to the much-discussed incoherence of 'epistemic contradictions' (i.e, claims of the form $\neg p \wedge \diamond p$ or $p \wedge \diamond\neg p$) [4, 23].

Combining **NTrans** and **Ver** with classical logic has untoward effects. To see this, first note a seemingly benign consequence of **NTrans** and **Ver**.

Fact 1. ***NTrans+Ver*** *entails* ***Epistemic Łukasiewicz (ELuk)****:* $K_a\neg p \vdash \neg\Diamond p$

Proof. $K_a\neg p \underset{\texttt{NTrans}}{\vdash} K_a\neg\Diamond p \underset{\texttt{Ver}}{\vdash} \neg\Diamond p$ □

There is *prima facie* evidence that **ELuk** is an apt principle on our intended reading of $\mathscr{L}$. Consider:

(12) Ann knows that it isn't raining. So, it can't be raining.

(13) Ann has conclusively established that it isn't raining. So, it must not be raining.

(14) # Bob knows that it isn't snowing, but it might be.

(15) # Bob has conclusively established that it isn't snowing, but it might be.

(12) seems like unobjectionable ordinary reasoning (to bolster this, the effect seems heightened when considering the closely related reasoning in (13)). (14) has an air of incoherence (as does the closely related (15)). **ELuk** explains both. But combining **ELuk** with unfettered classical logic has puzzling results. Consider:

Double Negation (DN): $\neg\neg\varphi \equiv \varphi$
Contraposition (Con): $\varphi \vdash \psi$ implies $\neg\psi \vdash \neg\varphi$

Fact 2. ***ELuk+Con+DN*** *entails* ***Uniformity I****:* $\Diamond p \vdash \neg K_a\neg p$

Proof. $\Diamond p \underset{\texttt{DN}}{\vdash} \neg\neg\Diamond p \underset{\texttt{ELuk+Con}}{\vdash} \neg K_a\neg p.$ □

Fact 3. ***Uniformity I+Ver*** *entails* ***Uniformity II****:* $K_a\Diamond p \vdash \neg K_b\neg p$

Proof. $K_a\Diamond p \underset{\texttt{Ver}}{\vdash} \Diamond p \underset{\texttt{Uni}}{\vdash} \neg K_b\neg p.$ □

Uniformity I and **II** seem invalid, egregiously implying that if an agent is aware of but rightly uncertain about p, *every* agent is uncertain about p. To see this, note that **Uniformity I** (with minimal assumptions) entails: $\Diamond p \wedge \Diamond\neg p \vdash \neg K_a\neg p \wedge \neg K_a p$. But 'it might be raining and might not be raining' predominantly serves to express the *speaker's* ignorance about the rain, while 'Jones doesn't know that it is raining and doesn't know that it isn't raining' expresses that *Jones* is ignorant: it is generally agreed that $\Diamond p$ either has a solipsistic reading as its default, or something close (e.g.., expression of the information state of a select group of agents that includes the speaker). Similarly, note that **Uniformity II** (with minimal assumptions) entails: $K_1\Diamond p \wedge K_1\Diamond\neg p \vdash \neg K_2\neg p \wedge \neg K_2 p$. But 'Smith knows it might be raining and might not be raining' predominantly serves to express *Smith's* ignorance about the rain, while 'Jones doesn't know that it is raining and doesn't know that it isn't raining' predominantly serves to express that *Jones* is ignorant.

To bolster this assessment, consider a banal context. Suppose that your dinner partner has a severe allergy to shellfish. You ask your waiter, Smith, 'Does the daily soup contain shellfish?'. Smith replies:

(16) It might. The kitchen usually puts shellfish in the soup, but not always. I'll check with Chef Jones. She always knows exactly what's in the soup.

Upon hearing (16), and waiting for Smith to return, one would normally happily accept/say all of:

(17) The soup might have shellfish (that's why Smith is checking with the kitchen).

(18) Smith knows that the soup might have shellfish.

(19) Unlike Smith, Jones knows whether the soup has shellfish.

It would be odd to conclude from (17) and (18), per **Uniformity**, that Jones doesn't know that the soup doesn't have shellfish. For then an uncontentious application of disjunctive syllogism, using (19), would yield (even before Smith returns): chef Jones knows that the soup has shellfish. Surely one shouldn't conclude *this* given only (16).

The general pattern here is emulated by other epistemic vocabulary. Let's use $\triangledown\varphi$ for 'it is likely that φ'. Then $K_a\neg p \vdash \neg\triangledown p$ (and $K_a\neg p \vdash K_a\neg\triangledown p$) is similarly well-supported by *prima facie* linguistic evidence, while the contrapositive $\triangledown p \vdash \neg K_a\neg p$ does *not* seem true. Compare:

(20) Ann knows that it isn't raining. So, it isn't likely to be raining.

(21) Ann knows that it isn't raining. So, Ann knows that it isn't likely to be raining.

(22) It is likely to rain tomorrow, but only our local metereologist Jones knows for sure.

(20) and (21) strike me as good, if redundant, reasoning (easily generalized), while (22) seems perfectly intelligible.

One style of response to all this tries to exploit the context-sensitivity of epistemic 'might' to preserve restricted versions of **Ver** and **NTrans** without abandoning classical logic. In particular, the strategy would be to say that **NTrans**, **ELuk**, and **Uniformity I** hold only when the '$\diamond$' deployed in $K_a\neg\diamond p$ and $\neg\diamond p$ is indexed to the information available to agent a (i.e., the same agent referred to in $K_a\neg p$). This is best expressed by enriching the syntax for $\mathscr{L}$, to record the agent each instance of $\diamond$ is indexed to:

Restricted NTrans: $K_a\neg\diamond_a p \equiv K_a\neg p$
Restricted ELuk: $K_a\neg p \vdash \neg\diamond_a p$

It may then be claimed that any ill results (e.g. unrestricted **Uniformity**) leading from **Con** and **DN** are a mere illusion brought on by subtle shifts in context. This strategy should not be dismissed out of hand. Nevertheless, its execution will not be trivial. Among other complications, it sits uneasily with the data collected above (for example, our intuitive assessment of claims (12)-(15), in support of **ELuk**, does *not* seem to hinge on taking 'might'/'can't'/'must' to be indexed to Ann/Bob's information specifically) and risks introducing such loose criteria for contextual shifts that the relevant explanations become bereft of content.

To bolster the alternative strategy of dropping classical logic (at least when epistemic modals are in play), note that independent motivation for rejecting **Con** has been tabled. For example, one might think that the empirical case for Łukasiewicz' principle is compelling (cf. [3]) and argue on this basis that **Con** must be false (given that $\diamond p \vdash p$ is obviously false). Alternatively, a proposed counterexample to modus tollens from [25], utilizing 'likely', is easily modified to bear against **Con**. Suppose an urn contains 100 marbles, big and small. Of the big, 10 are blue and 30 are red. Of the small, 50 are blue and 10 are red. A marble, *m*, is randomly selected and placed under a cup. Given only this information, (23) sounds like good reasoning, but (24) does not:

(23) Suppose that *m* is big. It follows that *m* is likely to be red.

(24) *m* isn't likely to be red. # Thus, *m* isn't big.

To see why the second inference in (24) seems incorrect, note that we *already* know that the marble isn't likely to be red, yet accepting that it isn't big is rash.

The current paper thus pursues the strategy of giving an independently motivated formal semantics that delivers **Ver** and **NTrans**, while invalidating **Uniformity (I)** and invalidating **Con**.

We add one last wrinkle to our list of logical desiderata: it seems that **NTrans** can be generalized (in ways that bear on our discussion). Consider:

Generalized Negative Transparency (GeNT): $K_a\neg(p \wedge \Diamond q) \equiv K_a\neg(p \wedge q)$
$K_a(p \vee \neg\Diamond q) \equiv K_a(p \vee \neg q)$

In both cases, **NTrans** is a special case (respectively, $p = \top$ and $p = \bot$). For convenience, I assume the above claims are equivalent (they could be deployed individually in our coming argumentation, however). Note that the linguistic evidence in support of **GeNT** seems no worse than that for **NTrans** (though, unsurprisingly, parsing the relevant sentences requires slightly more effort). Consider:

(25) # Ann knows that it isn't both raining and a good day for a picnic, but for all she knows it's both raining and might be a good day for a picnic.

(26) # Ann knows that either it isn't raining or must not be a good day for a picnic, but for all she knows it's both raining and a good day for a picnic.

(27) Ann knows that it isn't both raining and a good day for a picnic. So, Ann knows that either it isn't raining or it must not be a good day for a picnic.

(25) and (26) sound incoherent; (27) sounds like good reasoning. **GeNT** explains all this.

3 *Strategy*

Altogether, our target in the current paper is this:

Goal: Provide an independently motivated formal semantics that validates **Ver** and **GeNT** (with **NTrans** as a special case), and invalidates **Uniformity I**.

We proceed as follows. In §4, we consider the *domain semantics* of [24] and [17], a standard 'information-sensitive' semantics for 'might' claims (designed to account, in particular, for non-classical behavior induced by epistemic contradictions). Equipping domain semantics with an account of attitude ascriptions presented by [24] (following [14] and [6]) delivers **NTrans**. A natural starting point is thus to ask if **Ver** and **GeNT** can be realized in this setting without fuss. However, *ad hoc* maneuvers aside, this system forces a choice between **NTrans** and **Ver**. What's more, even *with* said *ad hoc* maneuvers, the system fails to deliver **GeNT**.

§5 thus proposes a novel alternative theory, showcasing a related but distinct tradition of information-sensitive semantics: we propose a formal *acceptance semantics* (in the ballpark of [21],[20], [12],[13], [5], [1]) that delivers **Ver** and **GeNT** as desired. Our treatment of $\Diamond p$ is essentially standard for such a framework; the more novel aspect is our account of $K\varphi$, and its interaction with $\Diamond p$. The guiding idea is that knowledge ascription reflects the *stability* of knowledge under *available refinements* of veridical information. A notion of inter-subjective 'available information' sets the bound on available refinements. A variation of a classic example (cf. [11, pg. 148]) provides initial motivation (cf. the Schmolmes case in [10, sect.1]):

> **Salvaging Operation.** Imagine a salvage crew searching for a ship that sank a long time ago. The mate of the salvage ship works from an old log, but overlooks some pertinent entries in the log, and concludes that the wreck may be in a certain bay. He confidently says 'the hulk might be in these waters'. But, as it turns out later, careful examination of the log shows that the boat must have gone down at least thirty miles further south.

One hesitates to say 'the mate *knew* that the ship might be in the bay' (better to say 'he merely believed it might be'), given that his rational acceptance of 'it might be in the bay' did not survive the incorporation of readily available information.

Our semantics may thus be taken (i) as an abstract version of the *defeasibility theory of knowledge* (cf. [16], [2]) and (ii) as a novel implementation of the insight from [11] that the *available information* bears on whether a speaker is entitled to an epistemic possibility claim, going beyond the actual knowledge of the speaker or hearers.

4 *Domain Semantics*

Domain semantics invites a natural account of knowledge ascription that exhibits **NTrans**. This contrasts with the influential *descriptivist/factualist* school on epistemic modals, according to which 'it might be that p' is taken as synonymous with, roughly, 'p is not ruled out by what is mutually known, or easily known, by a relevant group of agents'. Negative transparency seems untenable on the descriptivist account: that Smith knows that the train isn't late does not entail that Smith knows anything about what the mutual knowledge of a certain group rules out (even if the group includes only Smith: she might well be uncertain what she knows).

An *information model* $\mathscr{I} = \langle W, \mathtt{I} \rangle$ is a pair, with W the set of all possible worlds and $\mathtt{I}$ an assignment of an information state $\mathtt{I}(p)$ to each atomic sentence of $\mathscr{L}$. We take an information state – generically denoted $\mathbf{i}$ – to just be an *intension*, i.e., a subset of W. State $\mathbf{i}$ is *veridical at* w when $w \in \mathbf{i}$. We evaluate sentences in $\mathscr{L}$ as true (1) or false (0) relative to a possible world w and an information state $\mathbf{i}$: the valuation function $[\cdot]^{w,\mathbf{i}}$ is as follows.

Definition 1 (**Domain Semantics**). *Given an information model* $\mathscr{I}$*:*

$[p]^{w,i} = 1$ *iff* $w \in \mathtt{I}(p)$
$[\neg\varphi]^{w,i} = 1$ *iff* $[\varphi]^{w,i} = 0$
$[\varphi \wedge \psi]^{w,i} = 1$ *iff* $[\varphi]^{w,i} = 1$ *and* $[\psi]^{w,i} = 1$
$[\diamond\varphi]^{w,i} = 1$ *iff* $\exists u \in \boldsymbol{i}$*:* $[\varphi]^{u,i} = 1$

The following notion (following [23]) will be important for our account of attitude ascriptions:

Definition 2 (**Acceptance**). $\boldsymbol{i} \Vdash \varphi$ *iff* $\forall w \in \boldsymbol{i}$*:* $[\varphi]^{w,i} = 1$

If $\mathbf{i} \Vdash \varphi$, we say information $\mathbf{i}$ *accepts* or *supports* sentence φ, modeling the idea that having exactly the information $\mathbf{i}$ is sufficient for establishing φ, rendering φ correctly assertable (putting aside Gricean considerations, anyway). To get a feel for $\Vdash$, note that the following sensible properties are readily verified (though note that, given domain semantics, they do not generalize; cf. §5, [13]):

$\mathbf{i} \Vdash p$ iff $\forall w \in \mathbf{i}$: $w \in \mathtt{I}(p)$
$\mathbf{i} \Vdash \neg p$ iff $\forall w \in \mathbf{i}$: $w \notin \mathtt{I}(p)$
$\mathbf{i} \Vdash p \wedge q$ iff $\mathbf{i} \Vdash p$ and $\mathbf{i} \Vdash q$
$\mathbf{i} \Vdash p \vee q$ iff $\exists \mathbf{i}_1, \mathbf{i}_2$ s.t. $\mathbf{i} = \mathbf{i}_1 \cup \mathbf{i}_2$ and $\mathbf{i}_1 \Vdash p$ and $\mathbf{i}_2 \Vdash q$
$\mathbf{i} \Vdash \diamond p$ iff $\exists w \in \mathbf{i}$: $\{w\} \Vdash p$
$\mathbf{i} \Vdash \neg \diamond p$ iff $\forall w \in \mathbf{i}$: $\{w\} \Vdash \neg p$

As for logical consequence, two notions of entailment are prominent in this framework. First, a truth-preservation relation $\models$ is straightforwardly defined: $\varphi \models \psi$ holds exactly when $[\varphi]^{w,\mathbf{i}} = 1$ implies $[\psi]^{w,\mathbf{i}} = 1$ for every w and $\mathbf{i}$ in every model $\mathscr{I}$. Second, an acceptance-preservation relation $\Vdash$ is straightforwardly defined: $\varphi \Vdash \psi$ holds exactly when $\mathbf{i} \Vdash \varphi$ implies $\mathbf{i} \Vdash \psi$ for every $\mathbf{i}$ in every model $\mathscr{I}$. Both consequence relations serve as useful tools for explaining ordinary intuitions about entailment and contradiction. For example, the domain semanticist utilizes $\Vdash$, not $\models$, to explain the incoherence of epistemic contradictions of the form $p \wedge \diamond\neg p$: while $p \wedge \diamond\neg p$ is consistent with respect to $\models$, there is no $\mathbf{i}$ such that $\mathbf{i} \Vdash p \wedge \Diamond\neg p$.

To introduce attitude ascriptions, we transfer an account of belief ascription from [24] to knowledge ascription. Call this the *classical approach*. A *classical model* $\mathscr{C}$ supplements an information model with function $\mathbf{k}$, mapping a world to a non-empty intension $\mathbf{k}^w$. The idea is that $\mathbf{k}^w$ models Smith's epistemic state at w as a set of *epistemic alternatives* (the total informational content of Smith's knowledge). As an agent's knowledge can never rule out the actual world, we stipulate:

C1. $\forall w \in W$: $w \in \mathbf{k}^w$

Definition 3 (**Classicism**). *Given classical* $\mathscr{C}$*, we extend domain semantics with:*
$[K\varphi]^{w,\mathbf{i}} = 1$ *iff* $\mathbf{k}^w \Vdash \varphi$

However, relative to the strategy of §3, classicism is only a partial success.

Fact 4. *For classicists,* ***NTrans*** *holds.*

Proof. $[K\neg\Diamond p]^{w,\mathbf{i}}$ iff $\mathbf{k}^w \Vdash \neg\Diamond p$ iff $\forall u \in \mathbf{k}^w$: $\{u\} \Vdash \neg p$ iff $\forall u \in \mathbf{k}^w$: $u \notin \mathtt{I}(p)$ iff $\mathbf{k}^w \Vdash \neg p$ iff $[K\neg p]^{w,\mathbf{i}}$ □

Fact 5. *For classicists,* ***Ver*** *fails.*

Proof. Counter-model: consider $\mathscr{C}$ where (i) $W = \{w_1, w_2\}$, (ii) $\mathtt{I}(p) = \{w_2\}$, (iii) $\mathbf{k}^{w_1} = W$. Let $\mathbf{i} = \{w_1\}$. So, by (ii) and (iii), $[K\Diamond p]^{w_1,\mathbf{i}} = 1$, as there is a p-world in $\mathbf{k}^{w_1}$. But $[\Diamond p]^{w_1,\mathbf{i}} = 0$, as there is no p-world in $\mathbf{i}$. □

Of course, a small modification to the semantics secures **Ver**:

$[K\varphi]^{w,\mathbf{i}} = 1$ iff: $\mathbf{k}^w \Vdash \varphi$ and $[\varphi]^{w,\mathbf{i}} = 1$.

However, the modified proposal abandons **NTrans**. For a counter-model, take $\mathscr{C}$ where, for some $@ \in W$, every world in $\mathbf{k}^@$ (including @ itself) is a $\neg p$-world (assuring $\mathbf{k}^@ \Vdash \neg p \wedge \neg\Diamond p$ and $[\neg p]^{@,\mathbf{i}} = 1$), but there is a p-world in $\mathbf{i}$ (so $[\neg\Diamond p]^{@,\mathbf{i}} = 0$). So, given $\mathscr{C}$, $[K\neg p]^{@,\mathbf{i}} = 1$ and $[K\neg\Diamond p]^{@,\mathbf{i}} = 0$.

However, it is readily checked that the modified proposal yields: $\mathbf{i} \Vdash K\neg\Diamond p$ iff $\mathbf{i} \Vdash K\neg p$. So, **NTrans** emerges at the level of acceptance, in tandem with **Ver**. Nevertheless, two problems remain. First, the modified proposal is, as it stands, markedly *ad hoc*: adding the clause $[\varphi]^{w,\mathbf{i}} = 1$ to the truth condition for $K\varphi$ raises interpretive questions about the nature of $\mathbf{k}^w$ and serves *purely* to assure factivity in the case of modalized formulas (it is readily checked that **Ver** holds for $\Diamond$-free formulas in the original account of $K\varphi$). Second, even more pointedly, the modified proposal does not yield **GeNT**: in particular, there exists $\mathscr{C}$ and $\mathbf{i}$ where $\mathbf{i} \Vdash K\neg(p \wedge q)$ but $\mathbf{i} \nVdash K\neg(p \wedge \Diamond q)$. To see this, let $\mathbf{i}$ contain only worlds w_1 and w_2, with p only true at w_1, and q only true at w_2. Thus, $\mathbf{i} \Vdash \neg(p \wedge q)$ but $\mathbf{i} \nVdash \neg(p \wedge \Diamond q)$ (as $[p \wedge \Diamond q]^{w_1,\mathbf{i}} = 1$). If we further set $\mathbf{k}^w$ to be $\mathbf{i}$ for every $w \in \mathbf{i}$, we get: $\mathbf{i} \Vdash K\neg(p \wedge q)$ but $\mathbf{i} \nVdash K\neg(p \wedge \Diamond q)$.

5 *Stable Acceptance Semantics*

We now present an information-sensitive semantic theory that achieves the goal of §3. The leading idea behind this theory is that Smith's knowledge at w is *stable* under refinement of her veridical information at w - or at least refinements that are 'available' at w, in a sense to be clarified.

Our system may be seen as a novel implementation of a well-known (alleged) insight that the truth/aptness of an epistemic possibility claim is sensitive to *objective factors* that go beyond the actual knowledge of the speaker or other relevant agents: in particular, it is sensitive to information that has not been acquired but is (in some sense) *available* to the relevant agents. Consider two cases from [11].

> Imagine a salvage crew searching for a ship that sank a long time ago. The mate of the salvage ship works from an old log, makes a mistake in his calculations, and concludes that the wreck may be in a certain bay. It is possible, he says, that the hulk is in these waters. No one knows anything to the contrary. But in fact, as it turns out later, it simply was not possible for the vessel to be in that bay; more careful examination of the log shows that the boat must have gone down at least thirty miles further south. The mate said something false when he said, "It is possible that we shall find the treasure here", but the falsehood did not arise from what anyone actually knew at the time. [11, pg. 148]

As for the second case:

> Consider a person who buys a lottery ticket. At the time he buys his ticket we shall say it is possible he will win, though probably he will not. As expected, he loses. But retrospectively it would be absurd to report that it only seemed possible that the man would win. It was perfectly possible that he would win. To see this clearly, consider a slightly different case, in which the lottery is not above board; it is rigged so that only the proprietors can win. Thus, however it may have seemed to the gullible customer, it really was not possible that he would win. It only seemed so. "Seemed possible" and "was possible" both have work cut out for them. [11, pg. 148]

This suggests a proposal along the following lines: that whether an epistemic possibility claim is aptly assertible depends, in context, not only on the information that is already possessed, but that is available via "practicable investigation" (as Hacking puts it), or depends (as [7] puts it) on the "relevant way[s] by which members of the relevant community can come to know", or tracks (as [18, pg. 402] puts it) a distinction between what the speaker or other relevant agents "easily might know" versus "couldn't easily know or have known". We needn't commit to any particular elaboration here (cf.[10, sect.1]).

Exactly what to make of the above cases is debatable, as [17, Sects. 10.2.2, 10.4.2] points out. For our purposes, we need only observe the following. First, one hesitates to say that the mate *knew* that they might find the treasure in the bay: as his claim could not be maintained were accessible further evidence collected, it does not rise to knowledge. Second, it seems reasonable to say that we *knew*, at the time, that the person with the fair lottery ticket might win (but probably would not). Our beliefs seemed sufficiently sensitive to the available information: given the intrinsic limits on predicting a lottery, the possibility of his winning could not be ruled out even with all accessible evidence on the table.

Two strategies are available to theorists for explaining these observations. First, one could incorporate objective factors as a constraint on *epistemic possibility claims*. As [17, Sect. 10.2.2] notes, this has the cost that it becomes hard to see how the casual 'might' claims we make in ordinary life are ever warranted. Alternatively, one could incorporate objective factors as a constraint on *knowledge ascriptions* (with an eye to delivering plausible interactions with epistemic modals). As the conditions for asserting a knowledge claim are plausibly relatively demanding, the analogue of the previous objection has less force in this case. Our own theory exploits this second approach, citing the precedent and independent motivation provided by the tradition of *defeasibility* theories of knowledge, in the spirit of [16] (we leave more detailed comparisons for elsewhere).

In contrast to domain semantics, we offer a bilateral *acceptance semantics*: instead of evaluating sentences at world-information pairs and deriving acceptance conditions, sentences are evaluated at just an information state. Hence, acceptance conditions (and, simultaneously, rejection conditions) are *directly* provided. For some independent advantages of working with an acceptance semantics, see [21], [20], [5] and [1]; for independent drawbacks to domain semantics, see [13].

A *bounded model* $\mathcal{M}$ supplements an information model with functions $\mathbf{k}$ and $\mathbf{i}$, each mapping a world to an information state (a non-empty intension), respectively denoted $\mathbf{k}^w$ and $\mathbf{i}^w$. We call $\mathbf{i}^w$ the *worldly information at* w, while $\mathbf{k}^w$ again models the set of *epistemic alternatives*: the possible worlds compatible with the agent's total knowledge state (for simplicity we proceed with a single agent, writing K instead of K_1). We say that intension $\mathbf{j}$ *refines* intension $\mathbf{i}$ when $\mathbf{j} \subseteq \mathbf{i}$. We say that $\mathbf{i}$ is *internally coherent* when $\mathbf{i}$ is non-empty and, for every $w \in \mathbf{i}$, $\mathbf{i}^w$ refines $\mathbf{i}$. Intuitively, an internally coherent information state $\mathbf{i}$ is coherent in the following sense: if $\mathbf{i}$ leaves it open that the best available information (the 'worldly information') cannot rule out a certain possibility, then $\mathbf{i}$ does not itself rule out that possibility. We say that $\mathbf{i}$ is *accessible* at w exactly when $\mathbf{i}$ is both internally coherent and veridical at w, i.e., $w \in \mathbf{i}$. We stipulate, for all $w \in W$, that $\mathbf{k}^w$ and $\mathbf{i}^w$ are both accessible at w.

Lemma 1. *If* $\boldsymbol{i}$ *is internally coherent then* $\boldsymbol{i} = \bigcup_{w\in \boldsymbol{i}} \boldsymbol{i}^w$.

Proof. As $\mathbf{i}^w$ refines $\mathbf{i}$ for all $w \in \mathbf{i}$, we have $\bigcup_{w\in\mathbf{i}} \mathbf{i}^w \subseteq \mathbf{i}$. Suppose that $w \in \mathbf{i}$. As $\mathbf{i}^w$ is accessible at w, $w \in \mathbf{i}^w$. So, $\mathbf{i} \subseteq \bigcup_{w\in\mathbf{i}} \mathbf{i}^w$. □

Definition 4 (**Accessible Refinement**). *Given information state* $\boldsymbol{i}$*, let* $Acc(\boldsymbol{i})$ *be the set of information states* $\boldsymbol{j}$ *where (i)* $\boldsymbol{j}$ *refines* $\boldsymbol{i}$ *and (ii)* $\boldsymbol{j}$ *is accessible at* w *for some* $w \in \boldsymbol{i}$*. We call the members of* $Acc(\boldsymbol{i})$ *the accessible refinements of* $\boldsymbol{i}$*.*

Note that every $\mathbf{j} \in Acc(\mathbf{i})$ has the property: there exists $w \in \mathbf{i}$ such that $\mathbf{i}^w \subseteq \mathbf{j} \subseteq \mathbf{i}$. Thus, the accessible refinements of $\mathbf{i}$ are bounded by the candidates left open by $\mathbf{i}$ for what the worldly information might be.

Definition 5 (**Stable Acceptance Semantics**). *Given bounded* $\mathcal{M}$*, intension* $\boldsymbol{i}$*:*

$\boldsymbol{i} \Vdash p$	*iff*	$\forall w \in \boldsymbol{i}$: $w \in \mathtt{I}(p)$
$\boldsymbol{i} \dashV p$	*iff*	$\forall w \in \boldsymbol{i}$: $w \notin \mathtt{I}(p)$
$\boldsymbol{i} \Vdash \neg\varphi$	*iff*	$\boldsymbol{i} \dashV \varphi$
$\boldsymbol{i} \dashV \neg\varphi$	*iff*	$\boldsymbol{i} \Vdash \varphi$
$\boldsymbol{i} \Vdash \varphi \wedge \psi$	*iff*	$\boldsymbol{i} \Vdash \varphi$ *and* $\boldsymbol{i} \Vdash \psi$
$\boldsymbol{i} \dashV \varphi \wedge \psi$	*iff*	$\exists \boldsymbol{i}_1, \boldsymbol{i}_2$ *s.t.* $\boldsymbol{i} = \boldsymbol{i}_1 \cup \boldsymbol{i}_2$ *and* $\boldsymbol{i}_1 \dashV \varphi$ *and* $\boldsymbol{i}_2 \dashV \psi$
$\boldsymbol{i} \Vdash \Diamond\varphi$	*iff*	$\exists w \in \boldsymbol{i}$: $\{w\} \Vdash \varphi$
$\boldsymbol{i} \dashV \Diamond\varphi$	*iff*	$\forall w \in \boldsymbol{i}$: $\{w\} \dashV \varphi$
$\boldsymbol{i} \Vdash K\varphi$	*iff*	$\forall w \in \boldsymbol{i}$, $\forall \boldsymbol{j} \in Acc(\boldsymbol{k}^w)$: $\boldsymbol{j} \Vdash \varphi$
$\boldsymbol{i} \dashV K\varphi$	*iff*	$\forall w \in \boldsymbol{i}$, $\exists \boldsymbol{j} \in Acc(\boldsymbol{k}^w)$: $\boldsymbol{j} \nVdash \varphi$

Read $\mathbf{i} \Vdash \varphi$ as '$\mathbf{i}$ accepts φ' or '$\mathbf{i}$ supports φ', and $\mathbf{i} \dashV \varphi$ as '$\mathbf{i}$ rejects φ' or '$\mathbf{i}$ refutes φ'. The most unusual entry (cf. [21], [13], [1] and §4) is that for $K\varphi$: according to our semantics, 'Smith knows that φ' can be accepted exactly when it is established that every accessible refinement of Smith's knowledge state supports φ; 'Smith knows that φ' can be rejected exactly when it is established that an accessible refinement of Smith's knowledge state doesn't support φ.

A couple of technical lemmas will prove useful.

Lemma 2. *If* $\boldsymbol{i} \Vdash \varphi$ *and* $\boldsymbol{j} \Vdash \varphi$ *then* $\boldsymbol{i} \cup \boldsymbol{j} \Vdash \varphi$*. Likewise, if* $\boldsymbol{i} \dashV \varphi$ *and* $\boldsymbol{j} \dashV \varphi$ *then* $\boldsymbol{i} \cup \boldsymbol{j} \dashV \varphi$*.*

Proof. This can be established by a routine induction on φ, with respect to the following stronger property: (i) if $\mathbf{a} \Vdash \varphi$ and $\mathbf{b} \Vdash \varphi$ for all $\mathbf{a} \subseteq \mathbf{i}$ and $\mathbf{b} \subseteq \mathbf{j}$ then $\mathbf{a} \cup \mathbf{b} \Vdash \varphi$ for all $\mathbf{a} \subseteq \mathbf{i}$ and all $\mathbf{b} \subseteq \mathbf{j}$ and (ii) if $\mathbf{a} \dashV \varphi$ and $\mathbf{b} \dashV \varphi$ for all $\mathbf{a} \subseteq \mathbf{i}$ and all $\mathbf{b} \subseteq \mathbf{j}$ then $\mathbf{a} \cup \mathbf{b} \dashV \varphi$ for all $\mathbf{a} \subseteq \mathbf{i}$ and all $\mathbf{b} \subseteq \mathbf{j}$. □

Lemma 3. *If* $\boldsymbol{i}$ *is internally coherent, the following are equivalent:*

A. $\forall \boldsymbol{j} \in Acc(\boldsymbol{i})$: $\boldsymbol{j} \Vdash \varphi$

B. $\forall u \in \boldsymbol{i}$*:* $\boldsymbol{i}^u \Vdash \varphi$

Proof. As $\mathbf{i}^u \in Acc(\mathbf{i})$ for all $u \in \mathbf{i}$, the direction from A to B is trivial. For the other direction, consider $\mathbf{j} \in Acc(\mathbf{i})$ and use a routine induction on the structure of φ to show that if $\mathbf{i}^u \Vdash \varphi$ holds for all $u \in \mathbf{i}$, then $\mathbf{j} \Vdash \varphi$ holds, and if $\mathbf{i}^u \dashv\!\mid \varphi$ holds for all $u \in \mathbf{i}$, then $\mathbf{j} \dashv\!\mid \varphi$ holds, with Lemma 1 and Lemma 2 being put to crucial use (the latter for the case of $\varphi \wedge \psi$). □

It follows that our entries for $K\varphi$ have the following convenient reformulation, which we deploy in coming proofs:

$\mathbf{i} \Vdash K\varphi$ iff $\forall w \in \mathbf{i}, \forall u \in \mathbf{k}^w$: $\mathbf{i}^u \Vdash \varphi$
$\mathbf{i} \dashv\!\mid K\varphi$ iff $\forall w \in \mathbf{i}, \exists u \in \mathbf{k}^w$: $\mathbf{i}^u \nVdash \varphi$

Thus, according to our semantics, ‘Smith knows that φ’ can be accepted exactly when it is established that Smith’s knowledge state establishes that the worldly information establishes φ; ‘Smith knows that φ’ can be rejected exactly when it is established that Smith’s knowledge state leaves it open that the worldly information doesn’t establish φ.

This system invites the following notion of logical consequence:

Definition 6 (**Coherent Consequence**). $\varphi \VDash \psi$ *iff, for every bounded model* $\mathscr{M}$*, if* $\boldsymbol{i}$ *is internally coherent and* $\boldsymbol{i} \Vdash \varphi$*, then* $\boldsymbol{i} \Vdash \psi$*.*

Definition 7 (**Assertoric Equivalence**). *Sentences* φ *and* ψ *are* assertorically equivalent *if*

$$\boldsymbol{i} \Vdash \varphi \textit{ iff } \boldsymbol{i} \Vdash \psi$$

for every information state $\boldsymbol{i}$ *in every bounded model* $\mathscr{M}$*.*

For example, $p \wedge q$ and $q \wedge p$ are assertorically equivalent.

Definition 8. *A sentence* φ *is* $\diamond$-restricted *if the only occurrences of* $\diamond$ *are in the scope of a K operator.*

For example, $\neg(p \wedge q)$ and $K \diamond p$ are $\diamond$-restricted; $\diamond p$ and $\neg \diamond (p \vee q)$ aren’t.

To efficiently demonstrate the key properties of our system, we require some preliminary results, which are of independent technical interest.

Lemma 4. *If* φ *is* $\diamond$*-restricted then:*

(1) $\boldsymbol{i} \Vdash \varphi$ *iff* $\forall w \in \boldsymbol{i}$*:* $\{w\} \Vdash \varphi$

(2) $\boldsymbol{i} \dashv\!\mid \varphi$ *iff* $\forall w \in \boldsymbol{i}$*:* $\{w\} \dashv\!\mid \varphi$

Proof. A routine induction. □

Lemma 5. *If* φ *is* $\diamond$*-restricted then:* $\boldsymbol{i} \dashv\!\mid \diamond\varphi$ *iff* $\boldsymbol{i} \dashv\!\mid \varphi$

Proof. Suppose that $\mathbf{i} \dashv\!\mid \diamond\varphi$. Thus, $\forall w \in \mathbf{i}$: $\{w\} \dashv\!\mid \varphi$. Thus, by Lemma 4, $\mathbf{i} \dashv\!\mid \varphi$. The reasoning can be reversed. □

Theorem 1 (Normal Form). *For every sentence* φ*, there exists* $n \geqslant 0$ *and* $\diamond$*-restricted sentences* α_0*,* α_1*, ...,* α_n *such that for any internally coherent* $\boldsymbol{i}$*:*

$$\boldsymbol{i} \Vdash \varphi \textit{ iff } \boldsymbol{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \cdots \wedge \diamond\alpha_n$$

Proof. See the appendix. □

Now for the key results.

Fact 6. ***Generalized Negative Transparency*** *holds:* $K\neg(p \wedge \Diamond q) \mathrel{=\!\mid\mid\!\!\models} K\neg(p \wedge q)$.

Proof. Suppose that $\mathbf{i} \Vdash K\neg(p \wedge \Diamond q)$. So, $\forall w \in \mathbf{i}$, $\forall u \in \mathbf{k}^w$: $\mathbf{1}^u \Vdash \neg p$ and $\mathbf{2}^u \Vdash \neg \Diamond q$, where $\mathbf{1}^u \cup \mathbf{2}^u = \mathbf{i}^u$. By Lemma 5: $\forall w \in \mathbf{i}$, $\forall u \in \mathbf{k}^w$: $\mathbf{2}^u \Vdash \neg q$. So, $\mathbf{i} \Vdash K\neg(p \wedge q)$. The reasoning can be reversed. □

Fact 7. ***K-Veridicality*** *holds:* $K\varphi \mid\!\models \varphi$.

Proof. Assume that $\mathbf{i}$ is internally coherent and $\mathbf{i} \Vdash K\varphi$. So, $\forall w \in \mathbf{i}$, $\forall u \in \mathbf{k}^w$: $\mathbf{i}^u \Vdash \varphi$. By Theorem 1, there exists $n \geqslant 1$ and $\Diamond$-restricted sentences $\alpha_0, \alpha_1, \dots, \alpha_n$ such that, $\forall w \in \mathbf{i}$, $\forall u \in \mathbf{k}^w$: $\mathbf{i}^u \Vdash \alpha_0 \wedge \Diamond\alpha_1 \wedge \cdots \wedge \Diamond\alpha_n$.

We show that $\mathbf{i} \Vdash \alpha_0$. Let $w \in \mathbf{i}$. Now, as $w \in \mathbf{k}^w$ and $\mathbf{i}^u \Vdash \alpha_0$ for any $u \in \mathbf{k}^w$, we have $\mathbf{i}^w \Vdash \alpha_0$. So, by Lemma 4, we have $\forall u \in \mathbf{i}^w$: $\{u\} \Vdash \alpha_0$. Thus, as $w \in \mathbf{i}^w$, we have $\{w\} \Vdash \alpha_0$. Generalizing: $\forall w \in \mathbf{i}$: $\{w\} \Vdash \alpha_0$. So, by Lemma 4, $\mathbf{i} \Vdash \alpha_0$.

We show that $\mathbf{i} \Vdash \Diamond\alpha_k$ for $1 \leqslant k \leqslant n$. Let $w \in \mathbf{i}$. Now, for any $u \in \mathbf{k}^w$, there exists $v \in \mathbf{i}^u$ such that $\{v\} \Vdash \alpha_k$, as $\mathbf{i}^u \Vdash \Diamond\alpha_k$. As $w \in \mathbf{k}^w$, it follows that there exists $v \in \mathbf{i}^w$ such that $\{v\} \Vdash \alpha_k$. Thus, as $\mathbf{i}$ is internally coherent, $\exists v \in \mathbf{i}$ such that $\{v\} \Vdash \alpha_k$. So, $\mathbf{i} \Vdash \Diamond\alpha_k$.

Altogether: $\mathbf{i} \Vdash \alpha_0 \wedge \Diamond\alpha_1 \wedge \cdots \wedge \Diamond\alpha_n$. So, by Theorem 1, $\mathbf{i} \Vdash \varphi$. □

It is instructive to linger on the broad explanation as to why $\Diamond p$ is a coherent consequence of $K \Diamond p$. Suppose that $\mathbf{i}$ is internally coherent and supports $K \Diamond p$. Thus, $\mathbf{i}$ establishes that Smith's knowledge state establishes that the worldly information establishes $\Diamond p$. Thus, the candidates for the worldly information – those $\mathbf{i}$ cannot rule out – all contain a p-world. As $\mathbf{i}$ is internally coherent, $\mathbf{i}$ cannot itself rule out these worlds. So, $\mathbf{i}$ accepts $\Diamond p$.

Finally:

Fact 8. ***Uniformity*** *fails:* $\Diamond p \mid\!\not\models \neg K\neg p$.

Proof. Consider any bounded model $\mathscr{M}$ where: (i) $w_1 \in \mathtt{I}(p)$ and $w_2 \notin \mathtt{I}(p)$; (ii) $\mathbf{i}^{w_1} = \mathbf{k}^{w_1} = \{w_1\}$ and $\mathbf{i}^{w_2} = \mathbf{k}^{w_2} = \{w_2\}$. Set $\mathbf{i} = \{w_1, w_2\}$. Note that $\mathbf{i}$ is internally coherent.

By (i), $\{w_1\} \Vdash p$. So, $\exists w \in \mathbf{i}$: $\{w\} \Vdash p$. So, $\mathbf{i} \Vdash \Diamond p$.

By (i), $\{w_2\} \dashv\!\mid p$. Thus, by Lemma 4 and (ii), $\mathbf{i}^{w_2} \dashv\!\mid p$. Thus, by (ii), $\forall u \in \mathbf{k}^{w_2}$: $\mathbf{i}^u \dashv\!\mid p$. Thus, $\mathbf{i} \not\dashv\!\mid K\neg p$. Thus, $\exists w \in \mathbf{i}$ such that $\forall u \in \mathbf{k}^w$: $\mathbf{i}^u \dashv\!\mid p$. Thus, $\mathbf{i} \not\dashv\!\mid K\neg p$. Thus, $\mathbf{i} \nVdash \neg K\neg p$. □

References

[1] Maria Aloni (2022): *Logic and Conversation: The Case of Free Choice*. Semantics and Pragmatics 15(5), pp. 565–589, doi:10.3765/sp.15.5.

[2] Alexandru Baltag, Nick Bezhanishvili, Aybüke Özgün & Sonja Smets (forthcoming): *Justified Belief, Knowledge and the Topology of Evidence*. Synthese, doi:10.1007/s11229-022-03967-6.

[3] Justin Bledin (2014): *Logic Informed*. Mind 123(490), pp. 277–316, doi:10.1093/mind/fzu073.

[4] Justin Bledin & Tamar Lando (2018): *Closure and Epistemic Modals*. Philosophy and Phenomenological Research 97(1), pp. 3–22, doi:10.1111/phpr.12335.

[5] Ivano Ciardelli (2021): *Restriction without Quantification: Embedding and Probability for Indicative Conditionals*. Ergo 8, doi:10.3998/ergo.1158.

[6] Adam Dabrowski, Lawrence S. Moss & Rohit Parikh (1996): *Topological Reasoning and the Logic of Knowledge*. Annals of Pure and Applied Logic 78, pp. 73–110, doi:10.1016/0168-0072(95)00016-X.

[7] Keith DeRose (1991): *Epistemic Possibilities*. Philosophical Review 100(4), pp. 581–605, doi:10.2307/2185175.

[8] Cian Dorr & John Hawthorne (2013): *Embedding Epistemic Modals*. Mind 122(488), pp. 867–913, doi:10.1093/mind/fzt091.

[9] Andy Egan & Brian Weatherson, editors (2011): *Epistemic Modality*. Oxford University Press, doi:10.1093/acprof:oso/9780199591596.001.0001.

[10] Kai von Fintel & Anthony S Gillies (2011): *'Might' Made Right*. In Andy Egan & Brian Weatherson, editors: Epistemic Modality, Oxford University Press, Oxford, pp. 108–130, doi:10.1093/acprof:oso/9780199591596.003.0004.

[11] Ian Hacking (1967): *Possibility*. Philosophical Review 76(2), pp. 143–168, doi:10.2307/2183640.

[12] Peter Hawke & Shane Steinert-Threlkeld (2018): *Informational dynamics of epistemic possibility modals*. Synthese 195(10), pp. 4309–4342, doi:10.1007/s11229-016-1216-8.

[13] Peter Hawke & Shane Steinert-Threlkeld (2021): *Semantic Expressivism for Epistemic Modals*. Linguistics and Philosophy 44, pp. 475–511, doi:10.1007/s10988-020-09295-7.

[14] Jaakko Hintikka (1962): *Knowledge and Belief*. Cornell University Press, Ithaca.

[15] Nathan Klinedinst & Daniel Rothschild (2012): *Connectives without truth tables*. Natural Language Semantics 20(2), pp. 137–175, doi:10.1007/s11050-011-9079-5.

[16] Keith Lehrer & Thomas Paxson (1969): *Knowledge: Undefeated Justified True Belief*. Journal of Philosophy 66, pp. 225–37, doi:10.2307/2024435.

[17] John MacFarlane (2014): *Assessment Sensitivity*. Oxford University Press, doi:10.1093/acprof:oso/9780199682751.001.0001.

[18] G. E. Moore (1962): *Commonplace Book, 1919-53*. Allen and Unwin.

[19] Vít Punčochář (2015): *Weak Negation in Inquisitive Semantics*. Journal of Logic, Language and Information 24(3), pp. 323–355, doi:10.1007/s10849-015-9219-2.

[20] Mark Schroeder (2008): *Expression for Expressivists*. Philosophy and Phenomenological Research 76(1), pp. 86–116, doi:10.1111/j.1933-1592.2007.00116.x.

[21] Frank Veltman (1985): *Logics for Conditionals*. Ph.D. thesis, Universiteit van Amsterdam.

[22] Frank Veltman (1996): *Defaults in Update Semantics*. Journal of Philosophical Logic 25(3), pp. 221–261, doi:10.1007/BF00248150.

[23] Seth Yalcin (2007): *Epistemic Modals*. Mind 116(464), pp. 983–1026, doi:10.1093/mind/fzm983.

[24] Seth Yalcin (2011): *Nonfactualism About Epistemic Modality*. In Andy Egan & Brian Weatherson, editors: Epistemic Modality, Oxford University Press, Oxford, pp. 295–332, doi:10.1093/acprof:oso/9780199591596.003.0011.

[25] Seth Yalcin (2012): *A Counterexample to Modus Tollens*. Journal of Philosophical Logic 41(6), pp. 1001–1024, doi:10.1007/s10992-012-9228-4.

[26] Fan Yang & Jouko Väänänen (2017): *Propositional Team Logics*. Annals of Pure and Applied Logic 168(7), pp. 1406–1441, doi:10.1016/j.apal.2017.01.007.

A Appendix: Normal Form for Acceptance Semantics

Theorem 1. For every sentence φ, there exists $n \geqslant 0$ and $\diamond$-restricted sentences $\alpha_0, \alpha_1, \dots, \alpha_n$ such that for any internally coherent $\mathbf{i}$:

$$\mathbf{i} \Vdash \varphi \text{ iff } \mathbf{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \cdots \wedge \diamond\alpha_n$$

Proof. We proceed by induction on sentence structure, with respect to the following stronger property: there exists $m, n \geqslant 0$ and $\diamond$-restricted sentences $\alpha_0, \alpha_1, \dots, \alpha_m$ and $\beta_0, \beta_1, \dots, \beta_n$ such that, for any internally coherent $\mathbf{i}$:

$\mathbf{i} \Vdash \varphi$ iff $\mathbf{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m$

$\mathbf{i} \dashv\Vert \varphi$ iff $\mathbf{i} \Vdash \beta_0 \wedge \diamond\beta_1 \wedge \ldots \wedge \diamond\beta_n$

The case for atom p is trivial, as this sentence is itself $\diamond$-restricted: set $m = n = 0$, $\alpha_0 = p$ and $\beta_0 = \neg p$.

The case for knowledge ascription $K\varphi$ is trivial, as this sentence is itself $\diamond$-restricted: set $m = n = 0$, $\alpha_0 = K\varphi$ and $\beta_0 = \neg K\varphi$.

For the induction hypothesis IH, assume, for arbitrary φ and ψ, that there exists $m,n,x,y \geqslant 0$ and $\diamond$-restricted sentences

$$\alpha_0, \alpha_1, \ldots, \alpha_m, \beta_0, \beta_1, \ldots, \beta_n, \delta_0, \delta_1, \ldots, \delta_x, \varepsilon_0, \varepsilon_1, \ldots, \varepsilon_y$$

such that, for any internally coherent $\mathbf{i}$:

$\mathbf{i} \Vdash \varphi$ iff $\mathbf{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m$

$\mathbf{i} \dashv\Vert \varphi$ iff $\mathbf{i} \Vdash \beta_0 \wedge \diamond\beta_1 \wedge \ldots \wedge \diamond\beta_n$

$\mathbf{i} \Vdash \psi$ iff $\mathbf{i} \Vdash \delta_0 \wedge \diamond\delta_1 \wedge \ldots \wedge \diamond\delta_x$

$\mathbf{i} \dashv\Vert \psi$ iff $\mathbf{i} \Vdash \varepsilon_0 \wedge \diamond\varepsilon_1 \wedge \ldots \wedge \diamond\varepsilon_y$

Using the IH, we can prove the following.

$$\begin{array}{lll}
\mathbf{i} \Vdash \neg\varphi & \text{iff} & \mathbf{i} \dashv\Vert \varphi \\
 & \text{iff} & \mathbf{i} \Vdash \beta_0 \wedge \diamond\beta_1 \wedge \ldots \wedge \diamond\beta_n \\
\mathbf{i} \dashv\Vert \neg\varphi & \text{iff} & \mathbf{i} \Vdash \varphi \\
 & \text{iff} & \mathbf{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m \\
\mathbf{i} \Vdash \varphi \wedge \psi & \text{iff} & \mathbf{i} \Vdash \varphi \text{ and } \mathbf{i} \Vdash \psi \\
 & \text{iff} & \mathbf{i} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m \text{ and } \mathbf{i} \Vdash \delta_0 \wedge \diamond\delta_1 \wedge \ldots \wedge \diamond\delta_x \\
 & \text{iff} & \mathbf{i} \Vdash (\alpha_0 \wedge \delta_0) \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m \wedge \diamond\delta_1 \wedge \ldots \wedge \diamond\delta_x \\
\mathbf{i} \dashv\Vert \varphi \wedge \psi & \text{iff} & \exists \mathbf{i}_1, \mathbf{i}_2\text{: } \mathbf{i} = \mathbf{i}_1 \cup \mathbf{i}_2 \text{ and } \mathbf{i}_1 \dashv\Vert \varphi \text{ and } \mathbf{i}_2 \dashv\Vert \psi \\
 & \text{iff} & \exists \mathbf{i}_1, \mathbf{i}_2\text{: } \mathbf{i} = \mathbf{i}_1 \cup \mathbf{i}_2 \text{ and } \mathbf{i}_1 \Vdash \beta_0 \wedge \diamond\beta_1 \wedge \ldots \wedge \diamond\beta_n \\
 & & \text{and } \mathbf{i}_2 \Vdash \varepsilon_0 \wedge \diamond\varepsilon_1 \wedge \ldots \wedge \diamond\varepsilon_y \\
 & \text{iff} & \mathbf{i} \Vdash (\beta_0 \vee \varepsilon_0) \wedge \diamond(\beta_0 \wedge \beta_1) \wedge \ldots \wedge \diamond(\beta_0 \wedge \beta_m) \\
 & & \quad \wedge \diamond(\varepsilon_0 \wedge \varepsilon_1) \wedge \ldots \wedge \diamond(\varepsilon_0 \wedge \varepsilon_x) \\
\mathbf{i} \Vdash \diamond\varphi & \text{iff} & \exists w \in \mathbf{i}\text{: } \{w\} \Vdash \varphi \\
 & \text{iff} & \exists w \in \mathbf{i}\text{: } \{w\} \Vdash \alpha_0 \wedge \diamond\alpha_1 \wedge \ldots \wedge \diamond\alpha_m \\
 & \text{iff} & \exists w \in \mathbf{i}\text{: } \{w\} \Vdash \alpha_0 \wedge \alpha_1 \wedge \ldots \wedge \alpha_m \\
 & \text{iff} & \mathbf{i} \Vdash \diamond(\alpha_0 \wedge \alpha_1 \wedge \ldots \wedge \alpha_m) \\
 & \text{iff} & \mathbf{i} \Vdash (p \vee \neg p) \wedge \diamond(\alpha_0 \wedge \alpha_1 \wedge \ldots \wedge \alpha_m) \\
\mathbf{i} \dashv\Vert \diamond\varphi & \text{iff} & \forall w \in \mathbf{i}\text{: } \{w\} \dashv\Vert \varphi \\
 & \text{iff} & \forall w \in \mathbf{i}\text{: } \{w\} \Vdash \beta_0 \wedge \diamond\beta_1 \wedge \ldots \wedge \diamond\beta_n \\
 & \text{iff} & \forall w \in \mathbf{i}\text{: } \{w\} \Vdash \beta_0 \wedge \beta_1 \wedge \ldots \wedge \beta_n \\
 & \text{iff} & \mathbf{i} \Vdash \beta_0 \wedge \beta_1 \wedge \ldots \wedge \beta_n
\end{array}$$

□

Incentive Engineering for Concurrent Games

David Hyland
University of Oxford
Oxford, United Kingdom
david.hyland@cs.ox.ac.uk

Julian Gutierrez
Monash University
Melbourne, Australia
julian.gutierrez@monash.edu

Michael Wooldridge
University of Oxford
Oxford, United Kingdom
mjw@cs.ox.ac.uk

We consider the problem of incentivising desirable behaviours in multi-agent systems by way of taxation schemes. Our study employs the concurrent games model: in this model, each agent is primarily motivated to seek the satisfaction of a goal, expressed as a Linear Temporal Logic (LTL) formula; secondarily, agents seek to minimise costs, where costs are imposed based on the actions taken by agents in different states of the game. In this setting, we consider an external principal who can influence agents' preferences by imposing taxes (additional costs) on the actions chosen by agents in different states. The principal imposes taxation schemes to motivate agents to choose a course of action that will lead to the satisfaction of their goal, also expressed as an LTL formula. However, taxation schemes are limited in their ability to influence agents' preferences: an agent will always prefer to satisfy its goal rather than otherwise, no matter what the costs. The fundamental question that we study is whether the principal can impose a taxation scheme such that, in the resulting game, the principal's goal is satisfied in at least one or all runs of the game that could arise by agents choosing to follow game-theoretic equilibrium strategies. We consider two different types of taxation schemes: in a *static* scheme, the same tax is imposed on a state-action profile pair in all circumstances, while in a *dynamic* scheme, the principal can choose to vary taxes depending on the circumstances. We investigate the main game-theoretic properties of this model as well as the computational complexity of the relevant decision problems.

1 Introduction

Rational verification is the problem of establishing which temporal logic properties will be satisfied by a multi-agent system, under the assumption that agents in the system choose strategies that form a game-theoretic equilibrium [12, 41, 17]. Thus, rational verification enables us to verify which desirable and undesirable behaviours could arise in a system through individually rational choices. This article, however, expands beyond verification and studies methods for incentivising outcomes with favourable properties while mitigating undesirable consequences. One prominent example is the implementation of Pigovian taxes, which effectively discourage agents from engaging in activities that generate negative externalities. These taxes have been extensively explored in various domains, including sustainability and AI for social good, with applications such as reducing carbon emissions, road congestion, and river pollution [24, 22, 32].

We take as our starting point the work of [40], who considered the possibility of influencing one-shot Boolean games by introducing taxation schemes, which impose additional costs onto a game at the level of individual actions. In the model of preferences considered in [40], agents are primarily motivated to achieve a goal expressed as a (propositional) logical formula, and only secondarily motivated to minimise costs. This logical component limits the possibility to influence agent preferences: an agent can never be motivated by a taxation scheme away from achieving its goal. In related work, Wooldridge et al. defined the following implementation problem: given a game G and an objective Υ, expressed as a propositional logic formula, does there exists a taxation scheme τ that could be imposed upon G such that, in the resulting game G^{τ}, the objective Υ will be satisfied in at least one Nash equilibrium [40].

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 344–358, doi:10.4204/EPTCS.379.28

We develop these ideas by applying models of finite-state automata to introduce and motivate the use of history-dependent incentives in the context of *concurrent games* [2]. In a concurrent game, play continues for an infinite number of rounds, where at each round, each agent simultaneously chooses an action to perform. Preferences in such a multiplayer game are defined by associating with each agent i a Linear Temporal Logic (LTL) goal γ_i, which agent i desires to see satisfied. In this work, we also assume that actions incur costs, and that agents seek to minimise their limit-average costs.

Since, in contrast to the model of [40], play in our games continues for an infinite number of rounds, we find there are two natural variations of taxation schemes for concurrent games. In a *static* taxation scheme, we impose a fixed cost on state-action profiles so that the same state-action profile will always incur the same tax, no matter when it is performed. In a *dynamic* taxation scheme, the same state-action profile may incur different taxes in different circumstances: it is history-dependent. We first show that dynamic taxation schemes are strictly more powerful than static taxation schemes, making them a more appropriate model of incentives in the context of concurrent games, characterise the conditions under which an LTL objective Υ can be implemented in a game using dynamic taxation schemes, and begin to investigate the computational complexity of the corresponding decision problems.

2 Preliminaries

Where S is a set, we denote the powerset of S by $\mathbf{2}^S$. We use various propositional languages to express properties of the systems we consider. In these languages, we will let Φ be a finite and non-empty vocabulary of Boolean variables, with typical elements $p, q, \ldots$. Where a is a finite word and b is also a word (either finite or infinite), we denote the word obtained by concatenating a and b by ab. Where a is a finite word, we denote by a^ω the infinite repetition of a. Finally, we use $\mathbb{R}^+_n$ for the set of n-tuples of non-negative real numbers.

Concurrent Game Arenas: We work with concurrent game structures, which in this work we will refer to as *arenas* (to distinguish them from the game structures that we introduce later in this section) [2]. Formally a *concurrent game arena* is given by a structure

$$\mathscr{A} = (\mathscr{S}, \mathscr{N}, Ac_1, \ldots, Ac_n, \mathscr{T}, \mathscr{C}, \mathscr{L}, s_0),$$

where: $\mathscr{S}$ is a finite and non-empty set of *arena states*; $\mathscr{N} = \{1, \ldots, n\}$ is the set of *agents* – for any $i \in \mathscr{N}$, we let $-i = \mathscr{N} \setminus \{i\}$ denote the set of *all agents excluding i*; for each $i \in \mathscr{N}$, Ac_i is the finite and non-empty set of unique actions available to agent i – we let $Ac = \bigcup_{i \in \mathscr{N}} Ac_i$ denote the set of all *actions* available to all players in the game and $\vec{Ac} = Ac_1 \times \cdots \times Ac_n$ denote the set of all *action profiles*; $\mathscr{T} : \mathscr{S} \times Ac_1 \times \cdots \times Ac_n \to \mathscr{S}$ is the *state transformer function* which prescribes how the state of the arena is updated for each possible action profile – we refer to a pair $(s, \vec{\alpha})$, consisting of a state $s \in \mathscr{S}$ and an action profile $\vec{\alpha} \in \vec{Ac}$ as a *state-action profile*; $\mathscr{C} : \mathscr{S} \times Ac_1 \times \cdots \times Ac_n \to \mathbb{R}^n_+$ is the *cost function* – given a state-action profile $(s, \vec{\alpha})$ and an agent $i \in \mathscr{N}$, we write $\mathscr{C}_i(s, \vec{\alpha})$ for the i-th component of $\mathscr{C}(s, \vec{\alpha})$, which corresponds to the cost that agent i incurs when $\vec{\alpha}$ is executed at s; $\mathscr{L} : \mathscr{S} \to \mathbf{2}^\Phi$ is a *labelling function* that specifies which propositional variables are true in each state $s \in \mathscr{S}$; and $s_0 \in \mathscr{S}$ is the *initial state* of the arena. In what follows, it is useful to define for every agent $i \in \mathscr{N}$ the value c^*_i to be the maximum cost that i could incur through the execution of a state-action profile: $c^*_i = \max\{\mathscr{C}_i(s, \vec{\alpha}) \mid s \in \mathscr{S}, \vec{\alpha} \in \vec{Ac}\}$.

Runs: Games are played in an arena as follows. The arena begins in its initial state s_0, and each agent $i \in \mathscr{N}$ selects an action $\alpha_i \in Ac_i$ to perform; the actions so selected define an action profile, $\vec{\alpha} \in Ac_1 \times$

$\cdots \times Ac_n$. The arena then transitions to a new state $s_1 = \mathcal{T}(s_0, \alpha_1, \ldots, \alpha_n)$. Each agent then selects another action $\alpha_i' \in Ac_i$, and the arena again transitions to a new state $s_2 = \mathcal{T}(s_1, \vec{\alpha}')$. In this way, we trace out an infinite interleaved sequence of states and action profiles, referred to as a *run*, $\rho : s_0 \xrightarrow{\vec{\alpha}_0} s_1 \xrightarrow{\vec{\alpha}_1} s_2 \xrightarrow{\vec{\alpha}_2} \cdots$.

Where ρ is a run and $k \in \mathbb{N}$, we write $s(\rho, k)$ to denote the state indexed by k in ρ, so $s(\rho, 0)$ is the first state in ρ, $s(\rho, 1)$ is the second, and so on. In the same way, we denote the k-th action profile played in a run ρ by $\vec{\alpha}(\rho, k-1)$ and to single out an individual agent i's k-th action, we write $\alpha_i(\rho, k-1)$.

Above, we defined the cost function $\mathcal{C}$ with respect to individual state-action pairs. In what follows, we find it useful to lift the cost function from individual state-action pairs to sequences of state-action pairs and runs. Since runs are infinite, simply taking the sum of costs is not appropriate: instead, we consider the cost of a run to be the *average* cost incurred by an agent i over the run; more precisely, we define the average cost incurred by agent i over the first t steps of the run ρ as $\mathcal{C}_i(\rho, 0:t) = \frac{1}{t+1} \sum_{j=0}^{t} \mathcal{C}_i(\rho, j)$ for $t \geq 1$, whereby $\mathcal{C}_i(\rho, j)$ we mean $\mathcal{C}_i(s(\rho, j), \vec{\alpha}(\rho, j))$. Then, we define the cost incurred by an agent i over the run ρ, denoted $\mathcal{C}_i(\rho)$, as the *inferior limit of means*: $\mathcal{C}_i(\rho) = \liminf_{t \to \infty} \mathcal{C}_i(\rho, 0: \ t)$. It can be shown that the value $\mathcal{C}_i(\rho)$ always converges because the sequence of averages $\mathcal{C}_i(\rho, 0:t)$ is Cauchy.

Linear Temporal Logic: We use the language of Linear Temporal Logic (LTL) to express properties of runs [33, 11]. Formally, the syntax of LTL is defined wrt. a set Φ of Boolean variables by the following grammar:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathbf{X}\varphi \mid \varphi \, \mathbf{U} \, \varphi \tag{1}$$

where $p \in \Phi$. Other usual logic connectives ("$\bot$", " & ", "$\to$", "$\leftrightarrow$") are defined in terms of $\neg$ and $\vee$ in the conventional way. Given a set of variables Φ, let $LTL(\Phi)$ be the set of LTL formulae over Φ; where the variable set Φ is clear from the context, we simply write LTL. We interpret formulae of LTL with respect to pairs (ρ, t), where ρ is a run, and $t \in \mathbb{N}$ is a temporal index into ρ. Any given LTL formula may be true at none or multiple time points on a run; for example, a formula $\mathbf{X}q$ will be true at a time point $t \in \mathbb{N}$ on a run ρ if q is true on a run ρ at time $t+1$. We will write $(\rho, t) \models \varphi$ to mean that $\varphi \in LTL$ is true at time $t \in \mathbb{N}$ on run ρ. The rules defining when formulae are true (i.e., the semantics of LTL) are defined as follows:

$$\begin{array}{lll}
(\rho,t) \models \top & & \\
(\rho,t) \models p & \text{iff} & p \in \mathcal{L}(s(\rho,t)) \quad (\text{where } p \in \Phi) \\
(\rho,t) \models \neg\varphi & \text{iff} & \text{it is not the case that } (\rho,t) \models \varphi \\
(\rho,t) \models \varphi \vee \psi & \text{iff} & (\rho,t) \models \varphi \text{ or } (\rho,t) \models \psi \\
(\rho,t) \models \mathbf{X}\varphi & \text{iff} & (\rho,t+1) \models \varphi \\
(\rho,t) \models \varphi \, \mathbf{U} \, \psi & \text{iff} & \text{for some } t' \geq t : \ (\rho,t') \models \psi \text{ and} \\
 & & \text{for all } t \leq t'' < t' : (\rho,t'') \models \varphi
\end{array}$$

We write $\rho \models \varphi$ as a shorthand for $(\rho, 0) \models \varphi$, in which case we say that ρ *satisfies* φ. A formula φ is *satisfiable* if there is some run satisfying φ. Checking satisfiability for LTL formulae is known to be PSPACE-complete [38], while the synthesis problem for LTL is 2EXPTIME-complete [34]. In addition to the LTL tense operators $\mathbf{X}$ ("in the next state…") and $\mathbf{U}$ ("…until…"), we make use of the two derived operators $\mathbf{F}$ ("eventually…") and $\mathbf{G}$ ("always…"), which are defined as follows [11]: $\mathbf{F}\varphi = \top \, \mathbf{U} \, \varphi$ and $\mathbf{G}\varphi = \neg\mathbf{F}\neg\varphi$.

Strategies: We model strategies for agents as finite-state machines with output. Formally, strategy σ_i for agent $i \in \mathcal{N}$ is given by a structure $\sigma_i = (Q_i, next_i, do_i, q_i^0)$, where Q_i is a finite set of machine states, $next_i : Q_i \times Ac_1 \times \cdots \times Ac_n \to Q_i$ is the machine's state transformer function, $do_i : Q_i \to Ac_i$ is the machine's action selection function, and $q_i^0 \in Q_i$ is the machine's initial state. A collection of strategies, one for each agent $i \in \mathcal{N}$, is a *strategy profile*: $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n)$. A strategy profile $\vec{\sigma}$ enacted in an arena $\mathcal{A}$ will generate a unique run, which we denote by $\rho(\vec{\sigma}, \mathcal{A})$; the formal definition is standard, and we will omit it here [17]. Where $\mathcal{A}$ is clear from the context, we will simply write $\rho(\vec{\sigma})$. For each agent $i \in \mathcal{N}$, we write Σ_i for the set of all possible strategies for the agent and $\Sigma = \Sigma_1 \times \cdots \times \Sigma_n$ for the set of all possible strategy profiles for all players.

For a set of distinct agents $A \subseteq \mathcal{N}$, we write $\Sigma_A = \prod_{i \in A} \Sigma_i$ for the set of partial strategy profiles available to the group A and $\Sigma_{-A} = \prod_{j \in \mathcal{N} \setminus A} \Sigma_j$ for the set of partial strategy profiles available to the set of all agents excluding those in A. Where $\vec{\sigma} = (\sigma_1, \ldots, \sigma_i, \ldots, \sigma_n)$ is a strategy profile and σ_i' is a strategy for agent i, we denote the strategy profile obtained by replacing the i-th component of $\vec{\sigma}$ with σ_i' by $(\vec{\sigma}_{-i}, \sigma_i')$. Similarly, given a strategy profile $\vec{\sigma}$ and a set of agents $A \subseteq \mathcal{N}$, we write $\vec{\sigma}_A = (\sigma_i)_{i \in A}$ to denote a partial strategy profile for the agents in A and if $\vec{\sigma}_A' \in \Sigma_A$ is another partial strategy profile for A, we write $(\vec{\sigma}_{-A}, \vec{\sigma}_A')$ for the strategy profile obtained by replacing $\vec{\sigma}_A$ in $\vec{\sigma}$ with $\vec{\sigma}_A'$.

Games, Utilities, and Preferences: We obtain a *concurrent game* from an arena $\mathcal{A}$ by associating with each agent i a goal γ_i, represented as an LTL formula. Formally, a concurrent game $\mathcal{G}$ is given by a structure

$$\mathcal{G} = (\mathcal{S}, \mathcal{N}, Ac_1, \ldots, Ac_n, \mathcal{T}, \mathcal{C}, \mathcal{L}, s_0, \gamma_1, \ldots, \gamma_n),$$

where $(\mathcal{S}, \mathcal{N}, Ac_1, \ldots, Ac_n, \mathcal{T}, \mathcal{C}, \mathcal{L}, s_0)$ is a concurrent game arena, and γ_i is the LTL goal of agent i, for each $i \in \mathcal{N}$. Runs in a concurrent game $\mathcal{G}$ are defined over the game's arena $\mathcal{A}$, and hence we use the notations $\rho(\vec{\sigma}, \mathcal{G})$ and $\rho(\vec{\sigma}, \mathcal{A})$ interchangeably. When the game or arena is clear from the context, we omit the $\mathcal{G}$ and simply write $\rho(\vec{\sigma})$. Given a strategy profile $\vec{\sigma}$, the generated run $\rho(\vec{\sigma})$ will satisfy the goals of some agents and not satisfy the goals of others, that is, there will be a set $W(\vec{\sigma}) = \{i \in \mathcal{N} : \rho(\vec{\sigma}) \models \gamma_i\}$ of *winners* and a set $L(\vec{\sigma}) = \mathcal{N} \setminus W(\vec{\sigma})$ of *losers*.

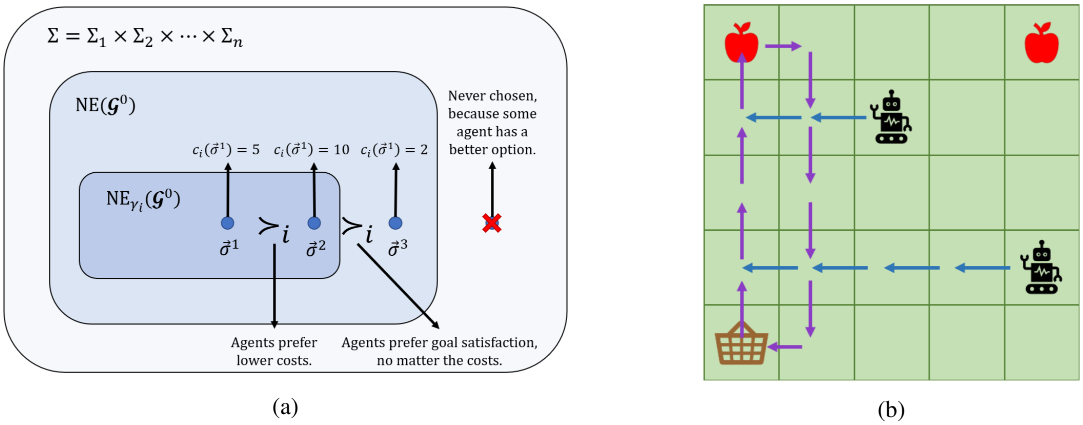
We are now ready to define preferences for agents. Our basic idea is that, as in [40], agents' preferences are structured: they first desire to accomplish their goal, and secondarily desire to minimise their costs. To capture this idea, it is convenient to define preferences via utility functions u_i over runs, where i's utility for a run ρ is

$$u_i(\rho) = \begin{cases} 1 + c_i^* - \mathcal{C}_i(\rho) & \text{if } \rho \models \gamma_i \\ -\mathcal{C}_i(\rho) & \text{otherwise.} \end{cases}$$

Defined in this way, if an agent i gets their goal achieved, their utility will lie in the range $[1, c_i^* + 1]$ (depending on the cost she incurs), whereas if she does not achieve their goal, then their utility will lie within $[-c_i^*, 0]$. Preference relations $\succeq_i$ over runs are then defined in the obvious way: $\rho_1 \succeq_i \rho_2$ if and only if $u_i(\rho_1) \geq u_i(\rho_2)$, with indifference relations $\sim_i$ and strict preference relations $\succ_i$ defined as usual.

Nash equilibrium: A strategy profile $\vec{\sigma}$ is a (pure strategy) Nash equilibrium if there is no agent i and strategy σ_i' such that $\rho(\vec{\sigma}_{-i}, \sigma_i') \succ_i \rho(\vec{\sigma})$. If such a strategy σ_i' exists for a given agent i, we say that σ_i' is a *beneficial deviation* for i from $\vec{\sigma}$. Given a game $\mathcal{G}$, let $\textsc{ne}(\mathcal{G})$ denote its set of Nash equilibria. In general, Nash equilibria in this model of concurrent games may require agents to play infinite memory strategies [8], but we do not consider these in this study [1]. Where φ is an LTL formula, we find it useful to define $\textsc{ne}_\varphi(\mathcal{G})$ to be the set of Nash equilibrium strategy profiles that result in φ being satisfied: $\textsc{ne}_\varphi(\mathcal{G}) = \{\vec{\sigma} \in \textsc{ne}(\mathcal{G}) \mid \rho(\vec{\sigma}) \models \varphi\}$. It is sometimes useful to consider a concurrent game that is modified so that no costs are incurred in it. We call such a game a *cost-free game*. Where $\mathcal{G}$ is a game, let $\mathcal{G}^{\mathbf{0}}$ denote

[1] Even in the purely quantitative setting where all agents' goals are $\top$, it is still possible that some Nash equilibria require infinite memory [16].



$\Sigma = \Sigma_1 \times \Sigma_2 \times \cdots \times \Sigma_n$
$\mathrm{NE}(\mathcal{G}^0)$
$\mathrm{NE}_{\gamma_i}(\mathcal{G}^0)$
$c_i(\vec{\sigma}^1) = 5$
$c_i(\vec{\sigma}^1) = 10$
$c_i(\vec{\sigma}^1) = 2$
$\vec{\sigma}^1$
$\succ_i$
$\vec{\sigma}^2$
$\succ_i$
$\vec{\sigma}^3$
Never chosen, because some agent has a better option.
Agents prefer lower costs.
Agents prefer goal satisfaction, no matter the costs.
(a)
(b)

Figure 1: (a) Illustration of the lexicographic quantitative and qualitative preferences of agents. (b) A concurrent game where two robots are situated in a grid world and are programmed to 1) never crash into another robot and 2) to secondarily minimise their limit-average costs. The arrows indicate how a run may be decomposed into a non-repeating and an infinitely-repeating component.

the game that is the same as $\mathcal{G}$ except that the cost function $\mathscr{C}^{\mathbf{0}}$ of $\mathcal{G}^{\mathbf{0}}$ is such that $\mathscr{C}_i^{\mathbf{0}}(s,\vec{\alpha}) = 0$ for all $i \in \mathcal{N}$, $s \in \mathcal{S}$, and $\vec{\alpha} \in \vec{Ac}$. Given this, the following is readily established (cf., [17]):

Theorem 1. *Given a game $\mathcal{G}$, the problem of checking whether* $\mathrm{NE}(\mathcal{G}^{\mathbf{0}}) \neq \emptyset$ *is* 2EXPTIME*-complete.*

The notion of Nash equilibrium is closely related to the concept of beneficial deviations. Given how preferences are defined in this study, it will be useful to introduce terminology that captures the potential deviations that agents may have [19]. Firstly, given a game $\mathcal{G}$, we say that a strategy profile $\vec{\sigma}^1 \in \Sigma$ is *distinguishable* from another strategy profile $\vec{\sigma}^2 \in \Sigma$ if $\rho(\vec{\sigma}^1,\mathcal{G}) \neq \rho(\vec{\sigma}^2,\mathcal{G})$. Then, for an agent i, a strategy profile $\vec{\sigma}$, and an alternative strategy $\sigma_i' \neq \sigma_i$, we say that σ_i' is an *initial deviation* for agent i from strategy profile $\vec{\sigma}$, written $\vec{\sigma} \rightarrow_i (\vec{\sigma}_{-i},\sigma_i')$, if we have $i \in W(\vec{\sigma}) \Rightarrow i \in W(\vec{\sigma}_{-i},\sigma_i')$ and strategy profile $\vec{\sigma}$ is distinguishable from $(\vec{\sigma}_{-i},\sigma_i')$.

3 Taxation Schemes

We now introduce a model of incentives for concurrent games. For incentives to work, they clearly must appeal to an agent's preferences $\succeq_i$. As we saw above, incentives for our games are defined with respect to both goals and costs: an agent's primary desire is to see their goal achieved – the desire to minimise costs is strictly secondary to this. We will assume that we cannot change agents' goals: they are assumed to be fixed and immutable. It follows that any incentives we offer an agent to alter their behaviour must appeal to the costs incurred by that agent. Our basic model of incentives assumes that we can alter the cost structure of a game by imposing *taxes*, which depend on the collective actions that agents choose in different states. Taxes may increase an agent's costs, influencing their preferences and rational choices.

Formally, we model static taxation schemes as functions $\tau : \mathcal{S} \times \vec{Ac} \rightarrow \mathbb{R}_+^n$. A static taxation scheme τ imposed on a game $\mathcal{G} = (\mathcal{S},\mathcal{N},Ac_1,\ldots,Ac_n,\mathcal{T},\mathscr{C},\mathcal{L},s_0,\gamma_1,\ldots,\gamma_n)$ will result in a new game, which we denote by

$$\mathcal{G}^\tau = (\mathcal{S},\mathcal{N},Ac_1,\ldots,Ac_n,\mathcal{T},\mathscr{C}^\tau,\mathcal{L},s_0,\gamma_1,\ldots,\gamma_n),$$

which is the same as $\mathcal{G}$ except that the cost function $\mathscr{C}^\tau$ of $\mathcal{G}^\tau$ is defined as $\mathscr{C}^\tau(s,\vec{\alpha}) = \mathscr{C}(s,\vec{\alpha}) + \tau(s,\vec{\alpha})$. Similarly, we write $\mathscr{A}^\tau$ to denote the arena with modified cost function $\mathscr{C}^\tau$ associated with $\mathcal{G}^\tau$ and $u_i^\tau(\rho)$

to denote the utility function of agent i over run ρ with the modified cost function $\mathscr{C}^{\tau}$. Given $\mathscr{G}$ and a taxation scheme τ, we write $\rho_1 \succeq_i^{\tau} \rho_2$ iff $u_i^{\tau}(\rho_1) \geq u_i^{\tau}(\rho_2)$. The indifference relations $\sim_i^{\tau}$ and strict preference relations $\succ_i^{\tau}$ are defined analogously.

The model of static taxation schemes has the advantage of simplicity, but it is naturally limited in the range of behaviours it can incentivise—particularly with respect to behaviours Υ expressed as LTL formulae. To overcome this limitation, we therefore introduce a *dynamic* model of taxation schemes. This model essentially allows a designer to impose taxation schemes that can choose to tax the same action in different amounts, depending on the history of the run to date. A very natural model for dynamic taxation schemes is to describe them using a finite state machine with output—the same approach that we used to model strategies for individual agents. Formally, a *dynamic taxation scheme T* is defined by a tuple $T = (Q_T, next_T, do_T, q_T^0)$ where Q_T is a finite set of taxation machine states, $next_T : Q_T \times Ac_1 \times \cdots \times Ac_n \to Q_T$ is the transition function of the machine, $q_T^0 \in Q_T$ is the initial state, and $do_T : Q_T \to (\mathscr{S} \times \vec{Ac} \to \mathbb{R}_+^n)$ is the output function of the machine. With this, let $\mathscr{T}$ be the set of all dynamic taxation schemes for a game $\mathscr{G}$. As a run unfolds, we think of the taxation machine being executed alongside the strategies. At each time step, the machine outputs a static taxation scheme, which is applied at that time step only, with $do_T(q_T^0)$ being the initial taxation scheme imposed.

When we impose dynamic taxation schemes, we no longer have a simple transformation $\mathscr{G}^{\tau}$ on games as we did with static taxation schemes τ. Instead, we define the effect of a taxation scheme with respect to a run ρ. Formally, given a run ρ of a game $\mathscr{G}$, a dynamic taxation scheme T induces an infinite sequence of static taxation schemes, which we denote by $t(\rho, T)$. We can think of this sequence as a function $t(\rho, T) : \mathbb{N} \to (\mathscr{S} \times \vec{Ac} \to \mathbb{R}_+^n)$. We denote the cost of the run ρ in the presence of a dynamic taxation scheme T by $\mathscr{C}^T(\rho)$:

$$\mathscr{C}^T(\rho) = \liminf_{u \to \infty} \frac{1}{u} \sum_{v=0}^{u} \mathscr{C}(\rho, v) + \underbrace{t(\rho, T)(v)(s(\rho, v), \vec{\alpha}(\rho, v))}_{(*)}$$

The expression $(*)$ denotes the vector of taxes incurred by the agents as a consequence of performing the action profile which they chose at time step v on the run ρ. The cost $\mathscr{C}_i^T(\rho)$ to agent i of the run ρ under T is then given by the i-th component of $\mathscr{C}^T(\rho)$.

Example 1. *Two robots are situated in a grid world (Figure 1b), where atomic propositions represent events where a robot picks up an apple (label a_{ij} represents agent i picking up apple j), has delivered an apple to the basket (label b_i represents agent i delivering an apple to the basket), or where the robots have crashed into each other (label c). Additionally, suppose that both robots are programmed with LTL goals $\gamma_1 = \gamma_2 = \mathbf{G}\neg c$. In this way, the robots are not pre-programmed to perform specific tasks, and it is therefore the duty of the principal to design taxes that motivate the robots to perform a desired function, e.g., pick apples and deliver them to the basket quickly. Because the game is initially costless, there is an infinite number of Nash equilibria that could arise from this scenario and it is by no means obvious that the robots will choose one in which they perform the desired function. Hence, the principal may attempt to design a taxation scheme to eliminate those that do not achieve their objective, thus motivating the robots to collect apples and deliver them to the basket. Clearly, using dynamic taxation schemes affords the principal more control over how the robots should accomplish this than static taxation schemes.*

4 Nash Implementation

We consider the scenario in which a principal, who is external to the game, has a particular goal that they wish to see satisfied within the game; in a general economic setting, the goal might be intended to capture some principle of social welfare, for example. In our setting, the goal is specified as an LTL formula Υ, and will typically represent a desirable system/global behaviour. The principal has the power to influence the game by choosing a taxation scheme and imposing it upon the game. Then, given a game $\mathcal{G}$ and a goal Υ, our primary question is whether it is possible to design a taxation scheme T such that, assuming the agents, individually and independently, act rationally (by choosing strategies $\vec{\sigma}$ that collectively form a Nash equilibrium in the modified game), the goal Υ will be satisfied in the run $\rho(\vec{\sigma})$ that results from executing the strategies $\vec{\sigma}$. In this section, we will explore two ways of interpreting this problem.

E-Nash Implementation: A goal Υ is *E-Nash implemented* by a taxation scheme T in $\mathcal{G}$ if there is a Nash equilibrium strategy profile $\vec{\sigma}$ of the game $\mathcal{G}^T$ such that $\rho(\vec{\sigma}) \models \Upsilon$. The notion of E-Nash implementation is thus analogous to the E-Nash concept in rational verification [14, 15]. Observe that, if the answer to this question is "yes" then this implies that the game $\mathcal{G}^T$ has at least one Nash equilibrium. Let us define the set $\text{ENI}(\mathcal{G},\Upsilon)$ to be the set of taxation schemes T that E-Nash implements Υ in $\mathcal{G}$:

$$\text{ENI}(\mathcal{G},\Upsilon) = \{T \in \mathcal{T} \mid \text{NE}_{\Upsilon}(\mathcal{G}^T) \neq \emptyset\} .$$

The obvious decision problem is then as follows:

E-NASH IMPLEMENTATION:
Given: Game $\mathcal{G}$, LTL goal Υ.
Question: Is it the case that $\text{ENI}(\mathcal{G},\Upsilon) \neq \emptyset$?

This decision problem proves to be closely related to the E-NASH problem [14, 15], and the following result establishes its complexity:

Theorem 2. E-NASH IMPLEMENTATION *is* 2EXPTIME*-complete, even when $\mathcal{T}$ is restricted to static taxation schemes.*

Proof. For membership, we can check whether Υ is satisfied on any Nash equilibrium of the cost-free concurrent game $\mathcal{G}^{\mathbf{0}}$ obtained from $\mathcal{G}$ by effectively removing its cost function using a static taxation scheme which makes all costs uniform for all agents. This then becomes the E-NASH problem, known to be 2EXPTIME-complete. The answer will be "yes" iff Υ is satisfied on some Nash equilibrium of $\mathcal{G}^{\mathbf{0}}$; and if the answer is "yes", then observing that $\text{NE}(\mathcal{G}^T) \subseteq \text{NE}(\mathcal{G}^{\mathbf{0}})$ for all taxation schemes $T \in \mathcal{T}$ [40], the given LTL goal Υ can be E-Nash implemented in $\mathcal{G}$. For hardness, we can reduce the problem of checking whether a cost-free concurrent game G has a Nash equilibrium (Theorem 1). Simply ask whether $\Upsilon = \top$ can be E-Nash implemented in $\mathcal{G}^{\mathbf{0}}$.

For the second part of the result, observe that the reduction above only involves removing the costs from the game and checking the answer to E-NASH, which can be done using a simple static taxation scheme. Hardness follows in a similar manner. □

A-Nash Implementation: The universal counterpart of E-Nash implementation is *A-Nash Implementation*. We say that Υ is *A-Nash implemented* by T in $\mathcal{G}$ if we have both 1) Υ is E-Nash implemented by T in game $\mathcal{G}$; and 2) $\text{NE}(\mathcal{G}^T) = \text{NE}_{\Upsilon}(\mathcal{G}^T)$. We thus define $\text{ANI}(\mathcal{G},\Upsilon)$ as follows:

$$\text{ANI}(\mathcal{G},\Upsilon) = \{T \in \mathcal{T} \mid \text{NE}(\mathcal{G}^T) = \text{NE}_{\Upsilon}(G^T) \neq \emptyset\}$$

The decision problem is then:

A-NASH IMPLEMENTATION:
Given: Game $\mathcal{G}$, LTL goal Υ.
Question: Is it the case that $\text{ANI}(\mathcal{G},\Upsilon) \neq \emptyset$?

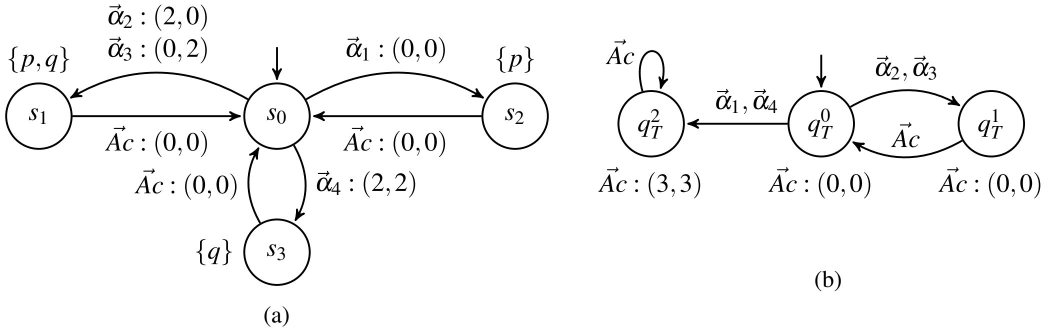


Figure 2: (a): A two-agent concurrent game $\mathcal{G}$ with action sets $Ac_1 = \{a,b\}$ and $Ac_2 = \{c,d\}$ and goals $\gamma_1 = \gamma_2 = \mathbf{GF}p$, where we let $\vec{\alpha}_1 = (a,c), \vec{\alpha}_2 = (a,d), \vec{\alpha}_3 = (b,c), \vec{\alpha}_4 = (b,d)$. Cost vectors associated with sets denote that all action profiles within the set are assigned those costs. (b): A dynamic taxation scheme that could be imposed on the agents in the game from (a). Labels below the states represent a static taxation scheme that applies a uniform tax for all agents and all action profiles.

The following result shows that, unlike the case of E-Nash implementation, dynamic taxation schemes are *strictly more powerful* than static taxation schemes for A-Nash implementation. It can be verified that the game in Figure 2a, the taxation scheme in Figure 2b, and the principal's goal being $\Upsilon = G(p \leftrightarrow q)$ are witnesses to this result (see Appendix for the full proof):

Proposition 1. *There exists a game $\mathcal{G}$ and an LTL goal Υ such that* $\textsc{ani}(\mathcal{G},\Upsilon) \neq \emptyset$*, but not if $\mathcal{T}$ is restricted to static taxation schemes.*

Before proceeding with the A-NASH IMPLEMENTATION problem, we will need to introduce some additional terminology and concepts, beginning first with deviation graphs, paths, and cycles. A *deviation graph* is a directed graph $\Gamma = (\mathcal{V}, E)$, where $\mathcal{V} \subseteq \Sigma$ is a set of nodes which represent strategy profiles in Σ and $E \subseteq \{(\vec{\sigma}, \vec{\sigma}') \in \mathcal{V} \times \mathcal{V} \mid \vec{\sigma} \rightarrow_i \vec{\sigma}' \text{ for some } i \in \mathcal{N}\}$ is a set of directed edges between strategy profiles that represent initial deviations. Additionally, we say that a dynamic taxation scheme T *induces* a deviation graph $\Gamma = (\mathcal{V}, E)$ if for every $(\vec{\sigma}, \vec{\sigma}') \in \mathcal{V} \times \mathcal{V}$, it holds that $\vec{\sigma}' \succ_i^T \vec{\sigma}$ for some $i \in \mathcal{N}$ if and only if $(\vec{\sigma}, \vec{\sigma}') \in E$. In other words, if the edges in a deviation graph precisely capture all of the beneficial deviations between its nodes under T, then the deviation graph is said to be induced by T.[2] Then, a *deviation path* is simply any path $P = (\vec{\sigma}^1, \dots, \vec{\sigma}^m)$ within a deviation graph Γ where $(\vec{\sigma}^j, \vec{\sigma}^{j+1}) \in E$ for all $j \in \{1, \dots, m-1\}$.

Because the principal is only able to observe the actions taken by the agents and not their strategies directly, any taxation scheme that changes the cost of some strategy profile $\vec{\sigma}$ will also change the cost of all strategy profiles that are indistinguishable from $\vec{\sigma}$ by the same amount. This naturally suggests that we modify the concept of a deviation path to take indistinguishability into account. To this end, we say that a sequence of runs $P_o = (\rho^1, \rho^2, \dots, \rho^m)$ is an *observed deviation path* in a deviation graph $\Gamma = (\mathcal{V}, E)$ if there exists an *underlying tuple* $(\vec{\sigma}^1, \vec{\sigma}^2, \dots, \vec{\sigma}^m)$ such that for all $j \in \{1, \dots, m\}$, it holds that 1) $\rho^j = \rho(\vec{\sigma}^j)$, and 2) if $j < m$, then $(\vec{\sigma}^j, \vec{\sigma}^{j+1'}) \in E$ for some $\vec{\sigma}^{j+1'}$ such that $\rho(\vec{\sigma}^{j+1'}) = \rho(\vec{\sigma}^{j+1})$. Then, a *deviation cycle* is a deviation path $(\vec{\sigma}^1, \dots, \vec{\sigma}^m)$ where $\rho(\vec{\sigma}^1) = \rho(\vec{\sigma}^m)$. A deviation path $P = (\vec{\sigma}^1, \vec{\sigma}^2, \dots, \vec{\sigma}^m)$ is said to *involve* an agent i if $\vec{\sigma}^j \rightarrow_i \vec{\sigma}^{j+1}$ for some $j \in \{1, \dots, m-1\}$ and similarly, an observed deviation path P_o in a deviation graph involves agent i if the analogous property holds for all of its underlying sets. Given a game $\mathcal{G}$ and a set of strategy profiles X, a taxation scheme T *eliminates*

[2] This definition implies that a taxation scheme may induce many possible deviation graphs in general, depending on the nodes selected to be part of the graph.

X if $NE(\mathscr{G}^T) \cap X = \emptyset$. Finally, a set of strategy profiles X is said to be *eliminable* if there exists a taxation scheme that eliminates it. With this, we can characterise the conditions under which a finite set of strategy profiles is eliminable:

Proposition 2. *Let $\mathscr{G}$ be a game and $X \subset \Sigma$ be a finite set of strategy profiles in $\mathscr{G}$. Then, X is eliminable if and only if there exists a finite deviation graph $\Gamma = (\mathscr{V}, E)$ that satisfies the following properties: 1) For every $\vec{\sigma} \in X$, there is some $\vec{\sigma}' \in \mathscr{V}$ such that $(\vec{\sigma}, \vec{\sigma}') \in E$; and 2) Every deviation cycle in Γ involves at least two agents.*

Proof Sketch. The forward direction follows by observing that if all deviation graphs fail to satisfy at least one of the two properties, then every deviation graph will either fail to eliminate some $\vec{\sigma} \in X$ if induced, or will not be inducible by any dynamic taxation scheme. The backward direction can be established by constructing a dynamic taxation scheme T^Γ that induces a deviation graph Γ satisfying the two properties. Using these properties, it follows that T^Γ eliminates X. □

To conclude our study of dynamic taxation schemes, we present a characterisation of the A-Nash implementation problem.[3]

Theorem 3. *Let $\mathscr{G}$ be a game and Υ be an LTL formula. Then $\text{ANI}(\mathscr{G}, \Upsilon) \neq \emptyset$ if and only if the following conditions hold:*

1. $\text{ENI}(\mathscr{G}, \Upsilon) \neq \emptyset$;
2. $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$ *is eliminable.*

Proof. For the forward direction, it follows from the definition of the problem that if $\text{ENI}(\mathscr{G}, \Upsilon) = \emptyset$, then $\text{ANI}(\mathscr{G}, \Upsilon) = \emptyset$. Moreover, it is also clear that if $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$ is not eliminable, then it is impossible to design a (dynamic) taxation scheme such that only good equilibria remain in the game and hence, $\text{ANI}(\mathscr{G}, \Upsilon) = \emptyset$.

For the backward direction, suppose that the two conditions hold and let T be a taxation scheme that only affects the limiting-average costs incurred by agents under strategy profiles in $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$, and eliminates this set. Such a taxation scheme is guaranteed to exist by the assumption that condition (2) holds and because it is known that no good equilibrium is indistinguishable from a bad one. Now consider a static taxation scheme τ such that $c_i(s, \vec{\alpha}) + \tau_i(s, \vec{\alpha}) = \hat{c}$ for all $i \in \mathscr{N}$, $(s, \vec{\alpha}) \in \mathscr{S} \times \vec{Ac}$, and some $\hat{c} \geq \max_{i \in \mathscr{N}} c_i^*$. Combining τ with T gives us a taxation scheme T^* such that for each state $q \in Q_{T^*} = Q_T$ and $(s, \vec{\alpha}) \in \mathscr{S} \times \vec{Ac}$, we have $do_{T^*}(q)(s, \vec{\alpha}) = do_T(q)(s, \vec{\alpha}) + \tau(s, \vec{\alpha})$. Now, because T eliminates $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$, and $\text{NE}(\mathscr{G}^\tau) = \text{NE}(\mathscr{G}^{\mathbf{0}})$, it follows that T^* eliminates $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$. Finally, note that because the satisfaction of an LTL formula on a given run is solely dependent on the run's trace, it follows that all good equilibria, i.e., strategy profiles in $\text{NE}_\Upsilon(\mathscr{G}^{\mathbf{0}})$, are distinguishable from all bad equilibria, so we have $\text{NE}_\Upsilon(\mathscr{G}^{\mathbf{0}}) \cap \text{NE}(\mathscr{G}^{T^*}) \neq \emptyset$. □

It is straightforward to see that A-NASH IMPLEMENTATION is 2EXPTIME-hard via a simple reduction from the problem of checking whether a Nash equilibrium exists in a concurrent game – simply ask if the formula $\top$ can be A-Nash implemented in $\mathscr{G}^{\mathbf{0}}$. However, it is an open question whether a matching upper bound exists and we conjecture that it does not. This problem is difficult primarily for two reasons. Firstly, it is well documented that Nash equilibria may require infinite memory in games with

[3] Note that, in general, Proposition 2 cannot be directly applied to Theorem 3, because it is assumed that the set to be eliminated is finite, whereas $\text{NE}_{\neg\Upsilon}(\mathscr{G}^{\mathbf{0}})$ is generally infinite. However, this can be reconciled if some restriction is placed on the agents' strategies so that Σ is finite, which is the case in many game-theoretic situations of interest, e.g., in games with memoryless, or even bounded memory, strategies – both used to model bounded rationality.

lexicographic ω-regular and mean-payoff objectives [8], and the complexity of deciding whether a Nash equilibrium even exists in games with our model of preferences has yet to be settled [15]. Secondly, Theorem 3 and Proposition 2 suggest that unless the strategy space is restricted to a finite set, a taxation scheme that A-Nash implements a formula may require reasoning over an infinite deviation graph, and hence require infinite memory. Nevertheless, our characterisation under such restrictions provides the first step towards understanding this problem in the more general setting.

5 Related Work and Conclusions

This work was motivated by [40], and based on that work, presents four main contributions: the introduction of *static and dynamic* taxation schemes as an extension to concurrent games expanding the model in (one-shot) Boolean games [40, 18, 19]; a study of the *complexity* of some of the most relevant computational decision problems building on previous work in rational verification [14, 17, 15]; evidence (formal proof) of the strict *advantage of dynamic taxation schemes* over static ones, which illustrates the role of not just observability but also *memory* to a principal's ability to (dis)incentivise certain outcomes [13, 20]; and a full characterisation of the *eliminability of sets of strategy profiles* under dynamic taxation schemes and the A-Nash implementation problem.

The incentive design problem has been studied in many different settings, and [35] group existing approaches broadly into those from the economics, control theory, and machine learning communities. However, more recent works in this area adopt multi-disciplinary methods such as automated mechanism design [30, 27, 37, 3], which typically focus on the problem of constructing incentive-compatible mechanisms to optimise a particular objective such as social welfare. Other approaches in this area reduce mechanism design to a program synthesis problem [29] or a satisfiability problem for quantitative strategy logic formulae [25, 28]. The notion of dynamic incentives has also been investigated in (multi-agent) learning settings [7, 26, 36, 42, 10]. These works focus solely on adaptively modifying the rewards for quantitative reward-maximising agents. In contrast, our model of agent utilities more naturally captures fundamental constraints on the principal's ability to (dis)incentivise certain outcomes due to the lexicographic nature of agents' preferences [4].

Another area closely related to incentives is that of norm design [23]. Norms are often modelled as the encouragement or prohibition of actions that agents may choose to take by a regulatory agent. The most closely related works in this area are those of [21, 31, 1], who study the problem of synthesising dynamic norms in different classes of concurrent games to satisfy temporal logic specifications. Whereas norms in these frameworks have the ability to *disable* actions at runtime, our model confers only the power to *incentivise* behaviours upon the principal. Finally, other studies model norms with violation penalties, but differ from our work in how incentives, preferences, and strategies are modelled [6, 5, 9].

In summary, a principal's ability to align self-interested decision-makers' interests with higher-order goals presents an important research challenge for promoting cooperation in multi-agent systems. The present study highlights the challenges associated with incentive design in the presence of constraints on the kinds of behaviours that can be elicited, makes progress on the theoretical aspects of this endeavour through an analysis of taxation schemes, and suggests several avenues for further work. Promising directions include extensions of the game model to probabilistic/stochastic or learning settings, finding optimal complexity upper bounds for the A-Nash implementation problems, and consideration of different formal models of incentives. We expect that this and such further investigations will positively contribute to our ability to develop game-theoretically aware incentives in multi-agent systems.

References

[1] Natasha Alechina, Giuseppe De Giacomo, Brian Logan & Giuseppe Perelli (2022): *Automatic Synthesis of Dynamic Norms for Multi-Agent Systems*. In: *19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022: KR 2022*, doi:10.24963/kr.2022/2.

[2] Rajeev Alur, Thomas A. Henzinger & Orna Kupferman (2002): *Alternating-time temporal logic*. *J. ACM* 49(5), pp. 672–713, doi:10.1145/585265.585270.

[3] Jan Balaguer, Raphael Koster, Christopher Summerfield & Andrea Tacchetti (2022): *The Good Shepherd: An Oracle Agent for Mechanism Design*. *arXiv preprint arXiv:2202.10135*, doi:10.48550/arXiv.2202.10135.

[4] Michael Bräuning, Eyke Hüllermeier, Tobias Keller & Martin Glaum (2017): *Lexicographic preferences for predictive modeling of human decision making: A new machine learning method with an application in accounting*. *European Journal of Operational Research* 258(1), pp. 295–306, doi:10.1016/j.ejor.2016.08.055.

[5] Nils Bulling & Mehdi Dastani (2016): *Norm-based mechanism design*. *Artificial Intelligence* 239, pp. 97–142, doi:10.1016/j.artint.2016.07.001.

[6] Henrique Lopes Cardoso & Eugénio Oliveira (2009): *Adaptive Deterrence Sanctions in a Normative Framework*. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2, pp. 36–43, doi:10.1109/WI-IAT.2009.123.

[7] Roberto Centeno & Holger Billhardt (2011): *Using incentive mechanisms for an adaptive regulation of open multi-agent systems*. In: *Twenty-Second International Joint Conference on Artificial Intelligence*, doi:10.5591/978-1-57735-516-8/IJCAI11-035.

[8] Krishnendu Chatterjee, Thomas A. Henzinger & Marcin Jurdzinski (2005): *Mean-Payoff Parity Games*. In: *20th IEEE Symposium on Logic in Computer Science (LICS 2005), 26-29 June 2005, Chicago, IL, USA, Proceedings*, IEEE Computer Society, pp. 178–187, doi:10.1109/LICS.2005.26.

[9] Davide Dell'Anna, Mehdi Dastani & Fabiano Dalpiaz (2020): *Runtime Revision of Sanctions in Normative Multi-Agent Systems*. *Autonomous Agents and Multi-Agent Systems* 34(2), doi:10.1007/s10458-020-09465-8.

[10] Mahmoud Elbarbari, Florent Delgrange, Ivo Vervlimmeren, Kyriakos Efthymiadis, Bram Vanderborght & Ann Nowé (2022): *A framework for flexibly guiding learning agents*. *Neural Computing and Applications*, pp. 1–17, doi:10.1007/s00521-022-07396-x.

[11] E. Allen Emerson (1990): *Temporal and Modal Logic*. In Jan van Leeuwen, editor: *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, Elsevier and MIT Press, pp. 995–1072, doi:10.1016/b978-0-444-88074-1.50021-4.

[12] Dana Fisman, Orna Kupferman & Yoad Lustig (2010): *Rational synthesis*. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, Springer, pp. 190–204, doi:10.1007/978-3-642-12002-2_16.

[13] Sanford J. Grossman & Oliver D. Hart (1992): *An analysis of the principal-agent problem*. In: *Foundations of insurance economics*, Springer, pp. 302–340, doi:10.1007/978-94-015-7957-5_16.

[14] Julian Gutierrez, Paul Harrenstein & Michael J. Wooldridge (2017): *Reasoning about equilibria in game-like concurrent systems*. *Annals of Pure and Applied Logic* 169(2), pp. 373–403, doi:10.1016/j.apal.2016.10.009.

[15] Julian Gutierrez, Aniello Murano, Giuseppe Perelli, Sasha Rubin, Thomas Steeples & Michael J. Wooldridge (2021): *Equilibria for games with combined qualitative and quantitative objectives*. *Acta Informatica* 58(6), pp. 585–610, doi:10.1007/s00236-020-00385-4.

[16] Julian Gutierrez, Muhammad Najib, Giuseppe Perelli & Michael J. Wooldridge (2019): *Equilibrium Design for Concurrent Games*. In: *30th International Conference on Concurrency Theory*, doi:10.4230/LIPIcs.CONCUR.2019.22.

[17] Julian Gutierrez, Muhammad Najib, Giuseppe Perelli & Michael J. Wooldridge (2020): *Automated temporal equilibrium analysis: Verification and synthesis of multi-player games*. *Artificial Intelligence* 287, p. 103353, doi:10.1016/j.artint.2020.103353.

[18] Paul Harrenstein, Paolo Turrini & Michael J. Wooldridge (2014): *Hard and soft equilibria in boolean games*. In Ana L. C. Bazzan, Michael N. Huhns, Alessio Lomuscio & Paul Scerri, editors: *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, IFAAMAS/ACM, pp. 845–852, doi:10.5555/2615731.2615867. Available at http://dl.acm.org/citation.cfm?id=2615867.

[19] Paul Harrenstein, Paolo Turrini & Michael J. Wooldridge (2017): *Characterising the Manipulability of Boolean Games*. In Carles Sierra, editor: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, ijcai.org, pp. 1081–1087, doi:10.24963/ijcai.2017/150.

[20] Bengt Holmstrom (1982): *Moral hazard in teams*. *The Bell journal of economics*, pp. 324–340, doi:10.2307/3003457.

[21] Xiaowei Huang, Ji Ruan, Qingliang Chen & Kaile Su (2016): *Normative Multiagent Systems: A Dynamic Generalization*. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, AAAI Press, p. 1123–1129.

[22] Kenneth S. Lyon & Dug Man Lee (2001): *Pigouvian tax and the congestion externality: a benefit side approach*. *Economics Research Institute Study Paper* 10, p. 1.

[23] Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, Mohd Zaliman Mohd Yusoff & Aida Mustapha (2014): *A review of norms and normative multiagent systems*. *The Scientific World Journal* 2014, doi:10.1155/2014/684587.

[24] N. Gregory Mankiw (2009): *Smart taxes: An open invitation to join the pigou club*. *Eastern Economic Journal* 35(1), pp. 14–23, doi:10.1057/EEJ.2008.43.

[25] Bastien Maubert, Munyque Mittelmann, Aniello Murano & Laurent Perrussel (2021): *Strategic reasoning in automated mechanism design*. In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, 18, pp. 487–496, doi:10.24963/kr.2021/46.

[26] David Mguni, Joel Jennings, Emilio Sison, Sergio Valcarcel Macua, Sofia Ceppi & Enrique Munoz de Cote (2019): *Coordinating the Crowd: Inducing Desirable Equilibria in Non-Cooperative Systems*. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 386–394.

[27] David Mguni & Marcin Tomczak (2019): *Efficient reinforcement dynamic mechanism design*. In: *GAIW: Games, agents and incentives workshops, at AAMAS, Montreal, Canada.*

[28] Munyque Mittelmann, Bastien Maubert, Aniello Murano & Laurent Perrussel (2022): *Automated synthesis of mechanisms*. In: *31st International Joint Conference on Artificial Intelligence (IJCAI-22)*, International Joint Conferences on Artificial Intelligence Organization, pp. 426–432, doi:10.24963/ijcai.2022/61.

[29] Sai Kiran Narayanaswami, Swarat Chaudhuri, Moshe Vardi & Peter Stone (2022): *Automating Mechanism Design with Program Synthesis*. *Proc. of the Adaptive and Learning Agents Workshop (ALA 2022).*

[30] David C Parkes, Ruggiero Cavallo, Florin Constantin & Satinder Singh (2010): *Dynamic incentive mechanisms*. *Ai Magazine* 31(4), pp. 79–94, doi:10.1609/aimag.v31i4.2316.

[31] Giuseppe Perelli (2019): *Enforcing equilibria in multi-agent systems*. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 188–196, doi:10.5555/3306127.3331692.

[32] Arthur C. Pigou & Nahid Aslanbeigui (2017): *The Economics of Welfare*. Routledge, doi:10.4324/9781351304368.

[33] Amir Pnueli (1977): *The Temporal Logic of Programs*. In: *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*, IEEE Computer Society, pp. 46–57, doi:10.1109/SFCS.1977.32.

[34] Amir Pnueli & Roni Rosner (1989): *On the Synthesis of an Asynchronous Reactive Module*. In Giorgio Ausiello, Mariangiola Dezani-Ciancaglini & Simona Ronchi Della Rocca, editors: *Automata, Languages*

and Programming, 16th International Colloquium, ICALP89, Stresa, Italy, July 11-15, 1989, Proceedings, *Lecture Notes in Computer Science* 372, Springer, pp. 652–671, doi:10.1007/BFb0035790.

[35] Lillian J Ratliff, Roy Dong, Shreyas Sekar & Tanner Fiez (2019): *A perspective on incentive design: Challenges and opportunities*. *Annual Review of Control, Robotics, and Autonomous Systems* 2, pp. 305–338, doi:10.1146/ANNUREV-CONTROL-053018-023634.

[36] Lillian J Ratliff & Tanner Fiez (2020): *Adaptive incentive design*. *IEEE Transactions on Automatic Control* 66(8), pp. 3871–3878, doi:10.1109/tac.2020.3027503.

[37] Weiran Shen, Pingzhong Tang & Song Zuo (2019): *Automated Mechanism Design via Neural Networks*. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, p. 215–223, doi:10.5555/3306127.3331696.

[38] A. Prasad Sistla & Edmund M. Clarke (1985): *The Complexity of Propositional Linear Temporal Logics*. *J. ACM* 32(3), pp. 733–749, doi:10.1145/3828.3837.

[39] Michael Ummels & Dominik Wojtczak (2011): *The complexity of Nash equilibria in limit-average games*. In: *International Conference on Concurrency Theory*, Springer, pp. 482–496, doi:10.1007/978-3-642-23217-6_32.

[40] Michael J. Wooldridge, Ulle Endriss, Sarit Kraus & Jérôme Lang (2013): *Incentive engineering for Boolean games*. *Artif. Intell.* 195, pp. 418–439, doi:10.1016/j.artint.2012.11.003.

[41] Michael J. Wooldridge, Julian Gutierrez, Paul Harrenstein, Enrico Marchioni, Giuseppe Perelli & Alexis Toumi (2016): *Rational Verification: From Model Checking to Equilibrium Checking*. In Dale Schuurmans & Michael P. Wellman, editors: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, AAAI Press, pp. 4184–4191, doi:10.1016/J.ARTINT.2017.04.003. Available at http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12268.

[42] Jiachen Yang, Ethan Wang, Rakshit Trivedi, Tuo Zhao & Hongyuan Zha (2021): *Adaptive Incentive Design with Multi-Agent Meta-Gradient Reinforcement Learning*. *arXiv preprint arXiv:2112.10859*, doi:10.5555/3535850.3536010.

6 Supplementary Material

Proposition 1. *There exists a game $\mathcal{G}$ and an LTL goal Υ such that $\text{ANI}(\mathcal{G},\Upsilon) \neq \emptyset$, but not if $\mathcal{T}$ is restricted to static taxation schemes.*

Proof. Consider the concurrent game $\mathcal{G}$ in Figure 2a. Intuitively, both agents desire to always eventually visit either s_1 or s_2. Suppose that the principal's objective is $\Upsilon = G(p \leftrightarrow q)$, i.e., they would like the agents to never visit s_2 or s_3. Firstly, observe that there is no *static* taxation scheme which can A-Nash implement Υ, as any modification to the costs of the game will not eliminate any Nash equilibria where the agents visit s_2 or s_3 a finite number of times. This is due to the prefix-independence of costs in infinite games with limiting-average payoffs [39]. However, the dynamic taxation scheme depicted in Figure 2b A-Nash implements Υ. To see this, observe that for any strategy profile that visits s_2 or s_3 a finite number of times, there exists a deviation for some agent to ensure that s_2 and s_3 are never visited. Such a deviation will result in all agents $i \in \{1,2\}$ satisfying their goals γ_i and strictly reducing their average costs from at least $c_i^* + 1$ to some value strictly below this. This constitutes a beneficial deviation and hence, there is no Nash equilibrium under T that does not satisfy Υ. Moreover, any strategy profile $\vec{\sigma}$ that leads to the sequence of states $s(\rho(\vec{\sigma}),0:) = (s_0 s_1)^\omega$ is a Nash equilibrium of $\mathcal{G}^T$ and hence goal Υ is A-Nash implemented by T in this game. □

Proposition 2. *Let $\mathcal{G}$ be a game and $X \subset \Sigma$ be a finite set of strategy profiles in $\mathcal{G}$. Then, X is eliminable if and only if there exists a finite deviation graph $\Gamma = (\mathcal{V}, E)$ that satisfies the following properties: 1) For every $\vec{\sigma} \in X$, there is some $\vec{\sigma}' \in \mathcal{V}$ such that $(\vec{\sigma},\vec{\sigma}') \in E$; and 2) Every deviation cycle in Γ involves at least two agents.*

Proof. For the forward direction, suppose that there is no deviation graph Γ satisfying both properties (1) and (2) in the statement. Then, for all deviation graphs Γ, either for some $\vec{\sigma} \in X$, there is no $\vec{\sigma}' \in \mathcal{V}$ such that $(\vec{\sigma},\vec{\sigma}') \in \mathcal{V}$, or there is some deviation cycle in Γ involving only one agent. Now consider any deviation graph $\Gamma = (\mathcal{V}, E)$, where $\mathcal{V} = X \cup \{\vec{\sigma}' \mid \vec{\sigma} \rightarrow_i \vec{\sigma}' \text{ for some } \vec{\sigma} \in X \text{ and } i \in \mathcal{N}\}$. In the first case, it is clear that any taxation scheme that induces Γ does not eliminate $\{\vec{\sigma}\}$ and hence X. In the second case, no taxation scheme can induce the deviation graph Γ. To see why, suppose for contradiction that some taxation scheme T induces Γ and let i be the agent for which there is a deviation cycle $C = \{\vec{\sigma}^1, \dots, \vec{\sigma}^m\}$ in Γ involving only agent i. Then, we have $\vec{\sigma}^1 \succ_i^T \vec{\sigma}^2 \succ_i^T \dots \succ_i^T \vec{\sigma}^m$ and by transitivity of the preference relation $\succ_i^T$, we can conclude that $\vec{\sigma}^1 \succ_i^T \vec{\sigma}^m$. However, by definition of a deviation cycle, $\vec{\sigma}^1$ and $\vec{\sigma}^m$ are indistinguishable, so agent i will always receive the same utility under both $\vec{\sigma}^1$ and $\vec{\sigma}^m$, no matter what taxation scheme is imposed on them and hence, we have a contradiction. From this, we can conclude that every deviation graph that can be induced by a taxation scheme does not eliminate X and hence, X is not eliminable, proving this part of the statement.

For the backward direction, assume that there is a deviation graph Γ that satisfies both properties. Under this assumption, we will construct a dynamic taxation scheme T that eliminates X. To assign the appropriate costs to different strategy profiles, we will make use of the lengths of deviation paths within Γ. For every $i \in \mathcal{N}$, let ℓ_i denote the length of the longest observed deviation path in Γ that involves only agent i. Additionally, for all $\vec{\sigma} \in \mathcal{V}$, let $d_i(\rho(\vec{\sigma}))$ denote the length of the longest *observed* deviation path in Γ that starts from $\rho(\vec{\sigma})$ and involves only i. The difference between these two quantities will serve as the basis for how much taxation an agent i will incur for any given strategy profile in $\mathcal{V}$. Observe that because it is assumed that no deviation cycle involves only one agent, both quantities are well-defined and finite for all agents and strategy profiles. Then, for a deviation graph Γ and a run ρ, let $\text{IN}(\rho,\Gamma)$ be the set of agents $i \in \mathcal{N}$ for which there is some pair of strategy profiles $\vec{\sigma}, \vec{\sigma}' \in \mathcal{V}$ such that we have both

$(\vec{\sigma}, \vec{\sigma}') \in E_D$ and $\rho = \rho(\vec{\sigma}')$. In other words, $\textsc{in}(\rho, \Gamma)$ represents the set of agents who have an initial deviation from some other strategy profile in $\mathscr{V}$ to one that generates the run ρ. With this, we would like to construct a dynamic taxation scheme such that for any strategy profile $\vec{\sigma}$, the following criteria are satisfied:

- $C_i^T(\rho(\vec{\sigma})) \geq (\ell_i - d_i(\rho(\vec{\sigma}))) \cdot (c_i^* + 1)$ if $i \in \textsc{in}(\rho, \Gamma)$;
- $C_i^T(\rho(\vec{\sigma})) = C_i(\rho(\vec{\sigma}))$ otherwise.

Intuitively, the idea is to ensure that for every edge $(\vec{\sigma}, \vec{\sigma}') \in E$, the agent $i \in \mathscr{N}$ for whom $\vec{\sigma} \rightarrow_i \vec{\sigma}'$ gets taxed by a significantly higher amount for choosing $\vec{\sigma}$ compared to when they choose $\vec{\sigma}'$. To see why it is possible to construct such a taxation scheme, first observe that if $\rho \neq \rho'$ for any two runs ρ, ρ', then there is some dynamic taxation scheme that can distinguish between the two by simply tracing out the two runs up to the first point in which they differ and then branching accordingly. From this point onwards, the dynamic taxation scheme can then output static taxation schemes, which assign different limiting average costs to the agents according to the above criteria. Extending this approach to a taxation scheme that distinguishes between all unique runs generated by elements of $\mathscr{V}$, it follows that there is a dynamic taxation scheme T that satisfies the two criteria. Consequently, for all $(\vec{\sigma}, \vec{\sigma}') \in E$, it follows that $\vec{\sigma}' \succ_i^T \vec{\sigma}$ because $\rho(\vec{\sigma}) \neq \rho(\vec{\sigma}')$ by definition of the initial deviation relation $\rightarrow_i$. Moreover, because it is assumed that no deviation cycle involves only one agent, T gives rise to a strict total ordering $\succ_i^T$ on the elements of $\mathscr{V}$ for each $i \in \mathscr{N}$. Finally, by property (1), it holds that for every $\vec{\sigma} \in X$, some agent has a beneficial deviation from $\vec{\sigma}$ to another $\vec{\sigma}' \in \mathscr{V}$ under T and hence, T eliminates X. □

Metatickles and Death in Damascus

Saira Khan
University of California, Irvine
Irvine, California

The prescriptions of our two most prominent strands of decision theory, evidential and causal, differ in a general class of problems known as Newcomb problems. In these, evidential decision theory prescribes choosing a dominated act. Attempts have been made at reconciling the two theories by relying on additional requirements such as ratification ([13]) or "tickles" ([3]). It has been argued that such attempts have failed ([18]; [23]). More recently, Huttegger ([11]) has developed a version of deliberative decision theory that reconciles the prescriptions of the evidentialist and causalist. In this paper, I extend this framework to problems characterised by decision instability, and show that it cannot deliver a resolute answer under a plausible specification of the tickle. I prove that there exists a robust method of determining whether the specification of the tickle matters for all two-state, two-act problems whose payoff tables exhibit some basic mathematical relationships. One upshot is that we have a principled way of knowing *ex-ante* whether a reconciliation of evidential and causal decision theory is plausible for a wide range of decision problems under this framework. Another upshot is that the tickle approach needs further work to achieve full reconciliation.

1 Introduction

Decision theory offers a normative framework for determining rational choice. Its primary components are a set of beliefs (probabilities) over states of the world and a set of valuations (utilities) over the different outcomes of acts in these states of the world. Two prominent forms of decision theory are the causalist and the evidentialist approaches. Causal decision theory determines rational action by evaluating what an agent can expect to bring about by her action. Evidential decision theory determines rational action by evaluating what evidence an agent's action provides her with.

The theories prescribe different acts as rational under a class of problems known as Newcomb problems. It is frequently held that the causalist prescription is the correct one ([26]; [6]; [22]; [18]).[1] The characteristic feature of Newcomb problems is that there is a correlation between state and act such that the choosing of the act is understood to be good evidence for a state of the world. The result is that evidentialism prescribes choosing an act which is strictly worse in both states of the world. The evidentialist recognises that though the agent cannot causally bring about a different state of the world, they deny that causality is important for practical rationality ([1]). Rather, the rational act should be based on its "news value". That is, an agent ought to prefer a proposition to another just in case she would rather learn that proposition over the other. In light of criticism of this position, attempts have been made – notably by Jeffrey ([13]) and Eells ([3]) – to amend evidential decision theory to better accord with causalist prescriptions.

In this paper, I focus on a version of reconciliation developed by Huttegger ([11]) and show that it cannot reconcile evidential and causal decision theory without further, questionable assumptions. Huttegger uses an idea due to Eells called the "tickle" defence: that the evidentialist becomes increasingly confident that the state of the world is not causally dependent on her act as a result of knowledge of her

[1] Though some, such as [1], [9] and [10] support the evidentialist conclusion.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 359–378, doi:10.4204/EPTCS.379.29

beliefs and desires. However, Huttegger employs the deliberative apparatus developed by Skyrms ([23]) and thus overcomes some objections to the original Eellsian approach.[2] Section 2 of this paper expounds the technical differences between causal and evidential decision theory and briefly outlines two decision problems: the Newcomb problem and Death in Damascus. Section 3 discusses Eells' approach to resolving the difference between the evidentialist and causalist prescriptions and details Huttegger's proposed amendment using deliberative dynamics. Huttegger's approach delivers the (commonly considered) correct answer for the evidentialist in the Newcomb problem.

Section 4 considers the same framework applied to a class of problems characterised by decision instability. These are where, as soon as the agent leans toward performing one action, the other looks preferable. In more technical terms: there is no dominant act (no act which is preferred regardless of the state of the world) and every act is in principle causally unratifiable (after we have chosen the act we would prefer to have chosen otherwise). In particular, I consider a decision problem known as Death in Damascus ([6]). When the payoff table is symmetric, the received view is that both naïve evidentialism and naïve causalism (without any deliberative dynamics) remain silent on which is the correct act to perform. When it is asymmetric, the evidentialist is decisive whereas the causalist is trapped in a state of indecision. A more sophisticated (deliberative) causalist may settle upon choosing an act with probability slightly less than 0.5. In this paper, we see that Huttegger's framework, when applied to this problem, cannot straightforwardly reconcile the evidentialist prescription with the prescription of the causalist (both sophisticated and naïve).

In Section 5, I offer an original analysis of the deliberative framework to explicate why it is irresolute in the Death in Damascus problem, and prove some general facts about its irresoluteness given a plausible version of the dynamical process, which I call the *shortest-path independence dynamics*. I identify the existence of what I call the *plane of indifference* in all two-act, two-state decision problems which exhibit the basic mathematical structure of either Newcomb or Death in Damascus problems. The key insight is that the specification of the tickle matters only depending on the positioning of this plane of indifference. In particular, regardless of the precise operation of the tickle during deliberation – shortest-path or not – the positioning of the plane in the Newcomb problem renders it the case that deliberation will always lead us to the same conclusion. This is not so in Death in Damascus and reconciliation of evidential and causal decision theory here requires more questionable assumptions. Section 6 discusses the status of reconciliation and the importance of the proof of the indifference plane for future work in deliberative decision theory. Section 7 concludes an offers a view on the status of the Eellsian project.

2 The decision problems

The canonical form of evidential decision theory is attributable to Jeffrey ([13]). Under his framework, states of the world, acts and outcomes are all propositions of the same kind, forming a Boolean algebra. Probabilities and desirabilities may be applied to any of these propositions. Call the Boolean closure of the set of acts, states and outcomes, the *decision-relevant* propositions. The agent's conditional expected utility of an act is calculated from her probabilities and desirabilities for maximally specific decision-relevant propositions. Formally, the evidential decision theorist prescribes performing the act, A, that maximises the following conditional expected utility formula, where D denotes desirability, P denotes probability, and S, the state of the world.

[2] In particular, the assumption that the agent access to a proposition which fully describes her beliefs and desires. Under Huttegger's approach, this is not assumed but rather reached through a process of deliberation.

$$EU_{evid}(A) = \sum_i D(S_i \& A)P(S_i|A)$$

There are multiple versions of causal decision theory.[3] For simplicity, I present Lewis' ([18]) account. Like the traditional decision-theoretic framework of Savage ([21]), states, acts and outcomes are not propositions of the same Boolean algebra but are separate entities. Probabilities attach to states of the world, and desirabilities or utilities, to outcomes. Lewis builds on the Savage framework but introduces *dependency hypotheses* which determine the appropriate partition of the state space. A dependency hypothesis is defined as the maximally specific proposition about how outcomes do, and do not, causally depend on the agent's present acts. Formally, the causal decision theorist prescribes performing the act, A, that maximises the following expected utility formula relative to the partition given by the dependency hypothesis.[4]

$$EU_{caus}(A) = \sum_i D(S_i \& A)P(S_i)$$

I now present two decision problems. One which has caused particular worry for the evidentialist, and one which has caused worry for both theories, though it is more frequently levied against the causalist ([5]). The first, Newcomb's problem, can be described as follows ([19]). Tomas is in a room with two boxes, one of which is opaque and one of which is transparent. Under the transparent box lies $1,000. Under the opaque box, there is either nothing or $1,000,000 and Tomas does not know which. He is offered the option to take either only the opaque box, or both the transparent one and the opaque one. The catch is that there is a predictor who, if she predicts Tomas chooses only the opaque box puts $1,000,000 under it and, if she predicts he chooses both boxes, puts nothing under it. Tomas believes the predictor is reliable. The payoff table is illustrated Table 1.

	Box empty	Box not empty
Take opaque box	0	1,000,000
Take both boxes	1,000	1,001,000

Table 1: Newcomb's Problem

In this decision problem, the causalist recommends taking both boxes, as it can be seen that this act strictly dominates taking only the opaque box. That is, it has higher expected utility under both states of the world. The naïve evidentialist, however, recommends taking only the opaque box, as choosing only the opaque box is good evidence that the predictor put $1,000,000 there. In this decision problem, the evidentialist seems to prescribe the wrong answer and Tomas loses out on a guaranteed $1,000.

The Death in Damascus problem is as follows ([6]). Death works from an appointment book which specifies a time and a place. If and only if Tereza happens to be in the time and place when Death is there, she dies. Suppose Tereza is in Damascus and she accidentally bumps into Death. He tells her that

[3] Most notably, the subjunctive accounts of Stalnaker ([26]) and Gibbard and Harper ([6]), as well as the non-subjunctive accounts of Skyrms ([22]) and Lewis ([18]).

[4] The merit of the evidential approach is that it is partition invariant and it is much less sensitive to the formal specification of the decision problem. Indeed, it is a more general framework that can be reduced to Savage's decision theory under correct specification of the state space. In comparison, in many causal decision theories, the decision problem must be specified in such a way that each state-act pair is guaranteed to lead to a unique outcome; there is state-act independence; and the desirabilities of the outcomes are not influenced by the state-act pair which eventuated them. None of these restrictions are required in the evidential framework. See Eells ([3]) for discussion.

he is coming for her tomorrow. Her options are either to stay where she is or to flee to Aleppo. The catch is that Death is a reliable predictor of where she will be, so as soon as Tereza believes it is better for her to flee, this constitutes good evidence that Death's appointment for her is in Aleppo and it seems as though she should stay. Analogously, however, if she decides to stay, this constitutes good evidence that Death knows that she stays and so she would be better off fleeing. The problem is therefore one of decision instability. The moment Tereza becomes confident in one option, the other appears more attractive. Here, I consider an asymmetric problem where the cost of fleeing is 1 util. The payoff table is given in Table 2, where we assign 10 utils to Tereza's survival.[5]

	Death in Damascus	Death in Aleppo
Stay in Damascus	0	10
Flee to Aleppo	9	-1

Table 2: Asymmetric Death in Damascus Problem

In this decision problem, the naïve evidentialist believes that, as Tereza's act is good evidence of the state of the world no matter what she chooses, she ought to stay in Damascus, since she should not pay the extra 1 util to flee to Aleppo. The causalist, however, believes that staying is irrational as it will put the agent in a position from which fleeing looks superior. She is therefore in a state of decision instability. Gibbard and Harper ([6]) argue that this is the correct answer as neither choice is ratifiable. Other forms of causal decision theory, for example, the deliberative framework of Joyce ([15]) or Arntzenius ([2]), prescribe the mixed act of fleeing with probability 0.474.[6] In Skyrms' and Huttegger's deliberative dynamics, the agent only has access to pure acts and is therefore in a state of indecision when deliberation assigns an act probability of less than 1. In Joyce's framework, the mixed act is a choice for the agent should she have access to a random chance device she may use to pick her final, pure act. That is, a chance device which will determine that she flees with probability 0.474. One might ask whether the evidentialist should be reconciled with the naïve causalist or deliberative causalist. If we sought similar instability as the naïve causalist, it will be clear from the analysis which follows that this will not be achieved: in many cases the deliberative evidentialist is decided. So I ask whether reconciliation with the Joycean causalist is possible on Huttegger's model – whether evidential decision theory can prescribe the mixed act of fleeing with probability 0.474. First, we must explicate the framework.

[5] While only the asymmetric case is presented in this paper, for completeness, the symmetric case was also analysed. This exhibits multiple lines of equilibria on the faces of the dynamical cube and therefore constitutes greater instability on the boundary than the asymmetric case. However, some would deny that indecision in such a circumstance constitutes a flaw in the theory. See, for example, [7].

[6] This is derived using Joyce's ([15]) framework for Murder Lesion applied to Death in Damascus assuming conditional probabilities $P(S2|A2) = P(S1|A1) = 0.99$. Under this framework, one's unconditional probabilities are revised in light of the expected utility calculation of an act in conjunction with the probabilistic correlation between state and act. More precisely, let α be a real number, $P_{t+1}(S2) = P_t(S2|EU_t(A2) = \alpha) \neq P_t(S2)$ when $\alpha \neq 0$, so the probability of a state of the world is updated based on its probability conditional upon the expected utility of an act. Further, let x and y be real numbers, if $P_t(A2) < 1$ and $x > y$, then $P_t(A2|EU_t(A2) = x \;\&\; EU_t(\sim A2) = y) > P_t(A2)$, so the choice probability of an act is updated based on its expected utility. The iterative process of updating one's choice probability will continue in this fashion until $P_t(A2) = P_{t+1}(A2) = P_t(A2|EU_t(A2))$, so information about its expected utility does not change its choice probability. As in Skyrms' ([23]) deliberational framework, this occurs when the expected utility of the two acts are equal.

3 A brief history of the metatickle approach and Huttegger's dynamics

A prominent evidentialist attempt to prescribe the causalist action in the Newcomb problems is attributable to Eells ([3]; [4]). This has been referred to as the "tickle" or "metatickle" defence ([18]; [23]).[7] Eells argues that the mistake being made by the naïve evidentialist in the Newcomb problem is the inference from some underlying common cause of both state and act, to a dependence *of* the state *on* the act. Eells argues that the only way in which the underlying cause could affect an agent's act is through the agent's beliefs and desires since, under our decision theories, these are the entities that determine action.[8] This implies that if the agent had full knowledge of his beliefs and desires, knowledge of the presence or absence of the common cause would be irrelevant to his act.

The intuition is clear with a simple example. Consider a decision problem with the same structure as the Newcomb problem but is instead a decision about whether or not to smoke cigarettes. Suppose that there is a genetic cause, C, that results in both lung cancer and a proclivity to enjoy cigarettes but smoking does not itself result in lung cancer. It is correlated with lung cancer but there is no causal state-act dependence. Causal decision theory recognises this independence and thus prescribes smoking insofar as it is enjoyable to the agent. The naïve evidentialist prescribes abstaining as smoking is good evidence for the presence of the gene which determines lung cancer. The Eellsian evidential decision theorist, however, believes that the only way the common cause can affect the agent's acts is through his beliefs and desires. Let the proposition which describes his beliefs and desires be denoted T for metatickle. We have:

$$P(A|T\&C) = P(A|T\& \sim C)$$

If an agent has full knowledge of her beliefs and desires, $P(T) = 1$. So in the presence of the metatickle,

$$P(A|C) = P(A| \sim C)$$

By symmetry of probabilistic independence,

$$P(C|A) = P(C| \sim A)$$

Since the cause is not probabilistically dependent on the act in the presence of the metatickle, neither is the state of the world. This means

$$P(S|A\&T) = P(S|T)$$

Eells believed that the proposition T was a proposition available to an agent ([3]; [4]). Conditional upon T, state and act are independent, and if this is the case, evidential decision theory will make the correct prescription: to smoke. Knowledge of the beliefs and desires of the kind caused by the common cause screens off what was previously thought to be evidence about the state of the world: the act. Analogous reasoning will lead the Eellsian evidential decision theorist to two-box in Newcomb's problem;

[7] It is so named for the following thought experiment. Suppose the agent feels a tickle in his left pinkie just in case the predictor has put \$1,000,000 in the opaque box. Then, even though the presence of money depends probabilistically on the agent's act, the tickle is sufficient to screen off the relevance of that act to the state of the world – the tickle tells the agent all he needs to know. A tickle may not always be available but, according to Eells, a "metatickle" is. This is a proposition which describes the agent's beliefs and desires.

[8] Eells suggests the common cause could not affect an agent's act by changing his decision rule. In particular "the agent believes that the causal influence of the common cause is sufficiently insignificant as to be irrelevant to the eventual determination of which act is correct in light of his beliefs and desires... This is because he believes that the causal influence of whatever is causally responsible for his rationality – his training, genetic make-up, and so on – will be overwhelming" ([3, 147]).

the act of two-boxing is irrelevant to the \$1,000,000 being there or not, and one should therefore choose the strictly dominant act.[9] The reasoning behind the metatickle approach is diagrammed in Figure 1.

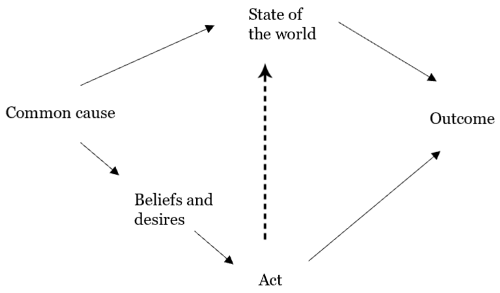


Figure 1: Diagrammatic depiction of Eells' metatickle defence where the causal connection between act and state is erroneously drawn on the basis of the common cause

For both Eells and Jeffrey ([12]), it is the agent's ability to anticipate her own choices that screens off the evidential import of her acts for states of the world.[10] However, unlike Eells, Jeffrey does not make reference to common causes. For Jeffrey, deliberation is what allows the sophisticated evidentialist to screen off the correlation between act and state which caused her to disagree with the causalist. He states "it is my credences and desirabilities at the end of deliberation that correspond to the preferences in the light of which I act, i.e., it is my final credence and desirability functions [...] not the initial ones [...] that underlie my choice" ([12, 486]). The idea is that the agent should not choose to maximise news value as she now sees it, but as she now expects herself to estimate it after having made the decision. This is known as "ratificationism". However, Huttegger believes that both Eells and Jeffrey did not adequately specify how the agent comes to fully know her beliefs and desires and achieve this screening off ([11]).[11] To fill this lacuna, he first turns to the deliberational dynamics of Skyrms ([23]).

Skyrms models a deliberational process where, as an agent deliberates about which act to choose, this is incorporated into her up-to-date probabilities and desirabilities. The agent has some information at the start of deliberation upon which she can assess expected utility but the deliberation process itself generates information that causes her to recalculate her expected utility. Suppose we assign probabilities to acts that represent the agent's belief that she will choose that particular act at the end of deliberation. Since states and acts are correlated, act probabilities provide evidence about states of the world which the agent can use to update her expected utility. Deliberation then pushes the agent in the direction of the act that has the higher expected utility in his current assessment. In particular, the direction of his choice probability of choosing both boxes, denoted $P(A2)$, is proportional to the difference in expected utility so that we have:

$$\frac{dP(A2)}{dt} \propto EU(A2) - EU(A1)$$

And

[9] See also Reichenbach's principle of screening off ([20]).

[10] Indeed, Skyrms ([24, 74]) refers to Jeffrey's idea as a "hypothetical version of the metatickle defense".

[11] See also [18]; [1]; [10]; [14]; [23]; [24] for criticisms of the metatickle approach.

$$\frac{dP(A2)}{dt} = \begin{cases} \text{positive if } EU(A2) > EU(A1) \\ \text{negative if } EU(A2) < EU(A1) \\ 0 \text{ if } EU(A2) = EU(A1) \end{cases}$$

We will refer to this as the "adaptive dynamics".[12]. It is assumed, in both Skyrms' and Huttegger's frameworks, that the adaptive dynamics operates continuously, though others, such as Eells [4] have developed discontinuous approaches. Since this paper is engaging with Huttegger's reconciliation project, I will assume a continuous adaptive dynamics. For Skyrms, the updating of one's choice probability continues until such a time as the agent reaches probability 1 of performing a certain act or the agent reaches a mixed equilibrium where there is no change in her choice probabilities ($\frac{dP(A2)}{dt} = 0$). The basic intuition capturing the metatickle is that, if Tomas leans toward only taking the one box, the probability of the $1,000,000 being there increases, and so he begins to believe that choosing both boxes is better. Let $S2$ denote the presence of the $1,000,000. Formally, as $P(A2)$ approaches 0 or 1, the conditional probabilities $P(S2|A1)$ and $P(S2|A2)$ approach 1 and 0, respectively. The value of $P(A2)$ where the expected utility of $A2$ and the expected utility of $A1$ are equal is where deliberation stops, and this is Tomas' final probability of two-boxing. On Skyrms' model this does not in fact end in a reconciliation of evidential and causal decision theory. Supposing Tomas is an evidentialist and begins on the fence, he ends deliberation most probably one-boxing, but also attaches some positive probability to two-boxing.

To this, Eells ([4]) introduces a model called "continual conditional expected utility maximization" which embraces Skyrms' idea that deliberation generates information upon which we should update our expected utilities but also introduces the notion that agents may face an urgency to act. Thus, depending on whether one wants to reach a decision quickly, one might eschew the states of indecision that Skyrms claims the evidentialist stuck in. Eells believes this reconciles the prescriptions of evidential and causal decision theory on Newcomb's problem, resulting in two-boxing. However, as Huttegger ([11]) rightly points out, this is a large deviation from traditional evidential decision theory. Whether an agent rushes to a decision or procrastinates are features of an agent not well captured by her preferences. Therefore, the proposed solution arguably fails.

Huttegger takes a different approach to reconciliation in light on Skyrms' findings. His amendment to Skyrms' model is a relaxation of the assumption that as $P(A2)$ approaches 0 or 1, the conditional probabilities $P(S2|A1)$ and $P(S2|A2)$ approach 1 and 0. That is, conditional probabilities of the states given acts are not functions of our choice probabilities. Indeed, in the original Eellsian account, there is nothing over and above one's informed beliefs and desires upon which the agent's decision is based; convergence towards one or the other act is not required for the appropriate screening off. Instead, conditional probabilities change by a separate "independence dynamics" as a function of time, or stages, in the deliberational process, moving closer to one another over the course of deliberation.[13] The independence

[12] Skyrms also refers to this informally as a dynamical rule which "seeks the good" ([25, 30]). He describes such rules as "qualitatively Bayesian" in the sense that the dynamical rule should reflect the agent's knowledge that she is an expected utility maximiser and the status of her present expected utilities as an expectation of her final utilities. Informally, such rules state that act probabilities should increase if the act has utility greater than the status quo, and that the probability of all acts with utilities greater than the status quo should increase. Frequently used dynamical rules that meet these conditions are the replicator dynamics or Nash dynamics, and the dynamics of Brown and von Neumann ([25]). Formally, $\frac{dP(A)}{dt} = \frac{cov(A)-P(A)\sum_j cov(A)_j}{k+\sum_j cov(A)_j}$ and $\frac{dP(A)}{dt} = cov(A)^2 - P(A)\sum_j cov(A)_j^2$ respectively, where the constant k represents how quickly the agent adjusts her act probabilities.

[13] One may argue against deliberation generating such information for the agent. However, for the purpose of my current analysis, I leave aside these issues. See [11] for a discussion.

dynamics is formally defined as follows.[14]

$$\frac{dP(S2|A1)}{dt} = \begin{cases} \text{positive if } P(S2|A1) > P(S2|A2) \\ \text{negative if } P(S2|A1) < P(S2|A2) \end{cases}$$

Likewise,

$$\frac{dP(S2|A2)}{dt} = \begin{cases} \text{positive if } P(S2|A2) > P(S2|A1) \\ \text{negative if } P(S2|A2) < P(S2|A1) \end{cases}$$

There are also no reappearances of correlations, so

$$\frac{d[P(S2|A2) - P(S2|A1)]}{dt} = 0 \text{ if } P(S2|A2) = P(S2|A1)$$

Under this dynamical process, evidential deliberation converges to two-boxing since the choice probability of two-boxing is governed by the adaptive dynamics when state and act are independent. It is precisely the introduction of the independence dynamics that brings us to this reconciliation. If the evidentialist does not believe her act is evidence for a state of the world, she in effect uses the same probabilities the causalist uses.

Furthermore, while in Skyrms' work, the end point of deliberation is where the choice probability of an act is 1 or $\frac{dP(A2)}{dt} = 0$, this is not the case under Huttegger's framework.[15] Rather, deliberation, in most cases, will continue until $\frac{dP(A2)}{dt} = 0$ *and* the agent reaches state-act independence. I say "in most cases" since Huttegger does not assume deliberation always leads to full state-act independence. This is because deliberation can sometimes fail provide all the information we need, for example, if the agent believes that the predictor in Newcomb's problem knows more about how he makes decisions than he knows about himself. If this is so, there are hidden factors influencing his choice which he cannot access via deliberation. Nonetheless, Huttegger states that situations where agents' acts are determined solely on the basis of their desires, beliefs and decision rule are the "most natural setting for decision theory" ([11, 22]). As such, I will be considering those cases in which the agent's deliberative process is sufficient to screen off state-act correlations.

In Huttegger's framework, the reason that the independence dynamics can continue after the adaptive dynamics concludes is because the operation of the independence dynamics is independent of the adaptive dynamics: it is not a function of the agent's choice probabilities. It is important to note that, on this interpretation, the relative strength of the independence and adaptive dynamics becomes relevant to where the agent ends deliberation. Huttegger's work finds that the exact specification of the operation of the independence dynamics relative to the adaptive dynamics does not matter for Eells' reconciliation project on Newcomb's problem. In this paper, I show that it does matter for other decision problems on which evidential and causal decision theory diverge.

[14] If $P(A1) = 0$, then $P(S2|A1)$ is not well defined. Huttegger states this obstacle can be overcome by requiring that dynamics of $P(S2|A1)$ is continuous with the dynamics for arbitrarily close states that have $P(A1) > 0$.

[15] In Skyrms ([23]), that the adaptive dynamics continues until the probability of an act equals 1, and does not exceed 1, is guaranteed by the fact that this is when deliberation ends. This is not the case under Huttegger's framework; deliberation does not end when the probability of an act reaches 1. Therefore, as stated here, it is possible that $P(A2)$ exceeds 1 since the rule that the change in choice probability is proportional to the difference in expected utility does not ensure that $P(A2)$ remains within the probability simplex. As such, we stipulate that the adaptive dynamic rules which are permissible under this general formulation are those which effectively slow as they reach the boundary, therefore remaining within the probability simplex over the course of deliberation.

I will not reconstruct Huttegger's work on Newcomb's problem here but rather apply his same framework to Death in Damascus. I begin by determining the dynamics on the boundaries and discuss the more complicated interior dynamics in Sections 5 and 6.

4 Death in Damascus for the deliberative evidentialist

In the language of metatickles, both Tereza's act of staying or fleeing and Death's appointment in Damascus or Aleppo are effects of a common cause; that is, the cognitive architecture of the agent upon which Death bases his appointment, sometimes referred to as the agent's "type" ([16]). Thus, conditional on the metatickle, T, which fully captures Tereza's beliefs and desires, states and acts are independent, and knowledge of the beliefs and desires of the kind caused by the common cause screens off the evidence that her choice provided for Death's location. Without making reference to common causes, but noting that deliberation can screen off state-act correlations, Huttegger introduces the independence dynamics, which, along with the adaptive dynamics describes the changes in an agent's choice probability over the course of her deliberation.

Under Huttegger's framework, $P(S2|A2)$ and $P(S2|A1)$ may vary independently so the deliberational space is represented in three dimensions; one being $P(S2|A1)$; the other $P(S2|A2)$; and the final being Tereza's probability of fleeing, $P(A2)$, all of which change during the deliberative process. The deliberational space is depicted in Figure 2. Note that the cube does not represent a phase diagram as the magnitude of the movement in any particular direction has not been specified. It should rather be thought of as a qualitative tool by which we may analyse where deliberation leads us.

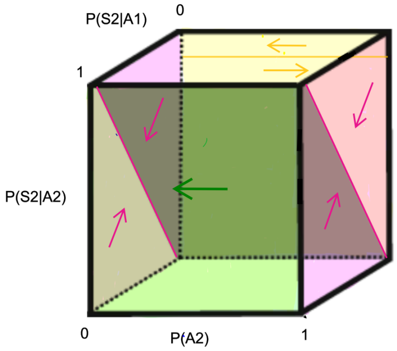


Figure 2: Deliberative evidentialist reasoning under Huttegger's framework

Recall the conditional expected utility formulae of evidential decision theory. That is,

$$EU_{evid}(A1) = D(S1\&A1)P(S1|A1) + D(S2\&A1)P(S2|A1)$$

$$EU_{evid}(A2) = D(S1\&A2)P(S1|A2) + D(S2\&A2)P(S2|A2)$$

Given these formulae and the logical fact that $P(S1|A1)+P(S2|A1) = 1$ and $P(S1|A2)+P(S2|A2) = 1$ (one or other state of the world must obtain given our act), we may discern the movement of $P(A2)$ on the faces of the cube by calculating the expected utility of both acts. First, let us address the front face, indicated in green, where $P(S2|A1) = 1$. The top edge is where $P(S2|A2) = 1$. Here we have $EU(A1) = 10$ and $EU(A2) = -1$. Since $EU(A2) < EU(A1)$, by the adaptive dynamics, $P(A2)$ decreases. Similarly, on the bottom edge of the front face, where $P(S2|A2) = 0$, $EU(A2) < EU(A1)$.

It can be verified that all points in between the edges also lead to a final choice probability of $P(A2) = 0$ on the front face of the cube. This is intuitive as, if $P(S2|A1) = 1$, Tereza can outsmart Death. That is, if the probability of Death being in Aleppo given that Tereza stays in Damascus is 1, she should surely stay in Damascus and not pay the extra 1 util to flee.

Now consider the back face, indicated in yellow, where $P(S2|A1) = 0$. The top edge is where $P(S2|A2) = 1$. Here we have $EU(A1) = 0$ and $EU(A2) = -1$. Again $P(A2)$ decreases. However, on the bottom edge of the back face, the dynamics look different. Here, $P(S2|A2) = 0$, so $EU(A1) = 0$ and $EU(A2) = 9$. Since $EU(A2) > EU(A1)$, $P(A2)$ increases. The exact point at which Tereza prefers fleeing over staying will be explored in the next section using what I call the *plane of indifference*.

However, we have not yet considered the operation of the independence dynamics on the left and right faces, indicated in pink. This leads us to what Huttegger calls the Eells-Jeffrey manifold, represented by the grey diagonal face in the cube, and consists of all points where $P(S2|A2) = P(S2) = P(S2|A1)$, in other words, where there is state-act independence. Movement toward the Eells-Jeffrey manifold is given by the evolving metatickle which screens off states from acts during an agent's deliberation. If our metatickle is sufficient to reach full state-act independence, we must determine the movement on the manifold itself.

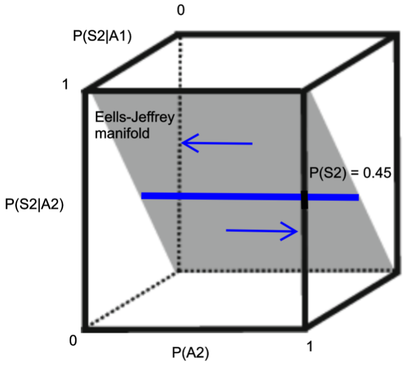


Figure 3: Evidentialist reasoning on the Eells-Jeffrey manifold

All areas above the bold blue line move to $P(A2) = 0$ and all areas below it move to $P(A2) = 1$ by the adaptive dynamics. The bold blue line is where $P(S2|A2) = P(S2|A1) = 0.45$. Here, $EU(A1) = EU(A2) = 4.5$ so there is no movement in $P(A2)$ as per our specification of the adaptive dynamics. I have not yet discussed the dynamical movement in much of the interior of the cube, which is the subject of the next section, but first it is worth noting the following facts.

Here, we have multiple equilibria represented by the bold blue line. All of these choice probabilities of $P(A2)$ render the expected utility of staying equal to that of fleeing, despite the fact that the unconditional probability of Death being in Damascus is 0.45.[16] However, this is also the case for the deliberative causalist. Though the mixed act of fleeing with probability 0.474 is the end point of deliberation, at this point, all other acts have equal expected utility so all are equally permissible ([15]). Here, one might inquire what then renders the mixed act the correct answer. The reason is that this is the uniquely ratifiable act (should one have the option to execute it using a chance device that represents this probability distribution). That is, it is the only act where, upon knowledge that one has chosen it, one would not

[16] It should be noted that such lines of equilibria in general exhibit structural instability. That is, they are sensitive to changes in the dynamical rule ([25]

prefer otherwise.[17]

In Sections 5 and 6, I show that the prescription of the mixed act under Huttegger's framework hinges upon two further conditions: (i) the independence dynamics does not take the "shortest path" to state-act independence, and (ii) the relative strength of the adaptive and independence dynamics must be such that they reach the Eells-Jeffrey manifold exactly where $P(A2) = 0.474$. Since these conditions imply that deliberation must proceed via a very specific route to the precise choice probability, it will not deliver reconciliation under many plausible specifications of the deliberative process. First, I consider what happens under one plausible specification of the independence dynamics.

5 Shortest-path independence and the plane of indifference

In this section, I offer an original analysis of the Huttegger's deliberative framework given a plausible version of the dynamical process, which I call the *shortest-path independence dynamics*. I prove the existence of what I call the *plane of indifference* which determines why the framework is irresolute in the case of Death in Damascus and not in Newcomb's problem. I then show that, under Huttegger's framework, this plane of indifference exists in all two-act, two-state decision problems which exhibit the basic mathematical structure of either Newcomb or decision instability problems. The upshot is that the precise specification of the independence dynamics matters for reconciliation only depending on the positioning of this plane of indifference. This provides a principled way of knowing *ex-ante* whether a reconciliation of evidential and causal decision theory is plausible for a wide range of decision problems under this framework.

Informally, the independence dynamics drives the agent's conditional probabilities toward one another over time, though the exact way in which this occurs is left open in Huttegger's work. One way the independence dynamics could operate is by adjusting one starting conditional probability to match the other. For example, if Tereza's initial value of $P(S2|A2)$ is 0.99 and her initial value of $P(S2|A1)$ is 0.01, she adjusts up the value of $P(S2|A1)$ until it also equals 0.99. However, this does not seem particularly rational. Given the description of the decision problem, both of her initial conditional probabilities reflect the Death's reliability in predicting her action, so there appears no reason to count one rather than the other as more viable for informing her unconditional credence in the state of the world.

A more plausible version of the independence dynamics would be one that concludes at the average across her two initial conditional probabilities. Since a movement in the direction of the manifold for one conditional probability then implies an equal movement in the direction of the manifold for the other, the independence dynamics decrees – absent its interaction with the adaptive dynamics – that Tereza's conditional probabilities move in the straight line that captures the shortest path to the manifold. This is illustrated in Figure 4, which represents a slice through the dynamical cube and the diagonal line represents the manifold.[18]

To see what this means for our deliberative process, first we must return to an important feature of the dynamical cube previously overlooked. In our earlier illustration, the line of equilibria on the manifold represented a situation where there was no movement prescribed by the adaptive dynamics; any choice probability of $P(A2)$ was acceptable since all mixtures of acts had equal expected utility. Moving off the

[17] This is also supported by consideration of the fact that the mixed act would constitute the Nash equilibrium of a normal form game with Death and Tereza as players. For discussion of the connection between ratifiability in deliberative decision theory and Nash equilibria in game theory, see [25]; [8]; [17]; and [27].

[18] Since the analysis is qualitative, this may extend to sufficiently similar independence dynamics, though this has not yet been considered.

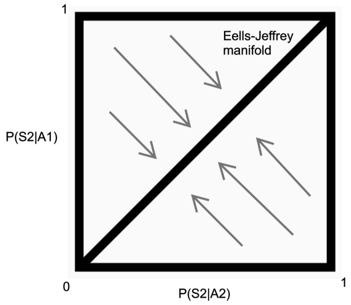


Figure 4: Shortest path to Eells-Jeffrey manifold

Eells-Jeffrey manifold, we see that this is not only a feature existing at state-act independence but, as I will show, there exists a whole plane on which the adaptive dynamics prescribes no change in $P(A2)$. This occurs where the two conditional probabilities of state given act, $P(S2|A1)$ and $P(S2|A2)$ sum to 0.9. The fact that this is a plane of the cube follows from the fact that two axes of the 3-dimensional space represent these conditional probabilities. The fact that the adaptive dynamics decrees no change in choice probability on this plane can be seen from the following.

Let $P(S2|A1)+P(S2|A2)=0.9$ and note it is true by definition that $P(S1|A1)=1-P(S2|A1)$ and $P(S1|A2)=1-P(S2|A2)$. Then

$$\begin{aligned} EU(A1) &= 0P(S1|A1)+10P(S2|A1) \\ &= 10P(S2|A1) \end{aligned}$$

And

$$\begin{aligned} EU(A2) &= 9P(S1|A2)-1P(S2|A2) \\ &= 10P(S2|A1) \end{aligned}$$

Since the expected utility of both acts are equal as defined in terms of $P(S2|A1)$, the adaptive dynamics prescribes no movement on the plane given by $P(S2|A1)+P(S2|A2)=0.9$. Figure 5 illustrates what I call the *plane of indifference*.

The key feature of this plane is that if one begins deliberation on the plane, since $P(A2)$ does not change, one simply moves by the independence dynamics toward the line of equilibria and ends deliberation with the same choice probability as she began with. Of utmost interest is what happens when we begin deliberation either below or above the plane of indifference. It turns out that if Tereza begins at any point below the plane, where $P(S2|A1)+P(S2|A2)<0.9$, Tereza's deliberation concludes that she should flee to Aleppo with probability 1. If she begins above the plane, where $P(S2|A1)+P(S2|A2)>0.9$, Tereza concludes she must stay in Damascus, and flee to Aleppo with probability 0.

For example, consider $P(S2|A1)+P(S2|A2)=1$. Here, we have a 2-dimensional plane which sits above the plane of indifference. All initial choice probabilities will lead Tereza to staying. To see this, note that since we have imposed the constraint $P(S2|A2)+P(S2|A1)=1$, and by logical fact, $P(S1|A1)+P(S2|A1)=1$ and $P(S1|A2)+P(S2|A2)=1$, our constraint implies $P(S1|A2)+P(S1|A1)=1$. Given these formulae, we may calculate our expected utilities. First, consider the top edge of the plane,

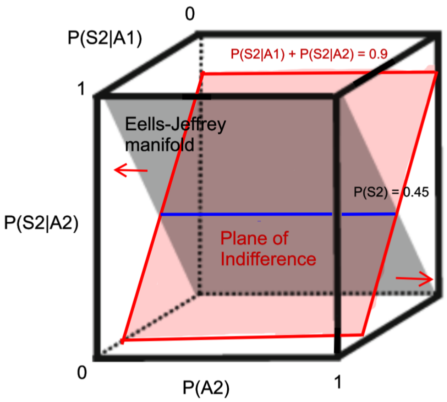


Figure 5: The plane of indifference

where $P(S2|A2) = 1$. We see that $EU(A1) = 0$ and $EU(A2) = -1$. Since $EU(A2) < EU(A1)$, by the adaptive dynamics, $P(A2)$ must reduce. Similarly, on the bottom edge of the plane where $P(S2|A2) = 0$, $EU(A2) < EU(A1)$. Since $P(S2|A2) + P(S2|A1) = 1$, shortest-path independence dynamics drives her unconditional probability $P(S2)$ to 0.5. In the middle of the plane on its intersection with the Eells-Jeffrey manifold, $EU(A1) = 5$ and $EU(A2) = 4$ so, again, $EU(A2) < EU(A1)$. As a result, deliberation moves Tereza toward staying in Damascus until we reach a stable equilibrium point where $P(A2) = 0$ and $P(S2|A2) = P(S2) = P(S2|A1) = 0.5$. Analogous reasoning applies when we begin on the other side of the plane and $P(S2|A1) + P(S2|A2) < 0.9$.

In what follows, I will prove that the adaptive dynamics is governed by whether we are below or above the plane of indifference for a general payoff table representing a wide range of decision instability problems. Let a denote the utility assigned to survival and b the utility assigned to death. Since we consider an asymmetric payoff table, let c denote the cost of fleeing. Our payoff table represents a general version of a wide range of asymmetric decision instability problems where $a > b$ and $c \leq a - b$. Other problems with a similar structure are the Murder Lesion problem and the Psychopath Button ([5]; [2]; [15]).

	S1	S2
A1	b	a
A2	a - c	b - c

Table 3: Generalised payoff table for asymmetric decision instability problem

The plane of indifference can be defined in terms of the utilities in the payoff table. Recall that the adaptive dynamics prescribes no movement in $P(A2)$ when $EU(A1) = EU(A2)$. This is when

$$bP(S1|A1) + aP(S2|A1) = (a-c)P(S1|A2) + (b-c)P(S2|A2)$$

By substitution and rearranging, we get

$$P(S2|A1) + P(S2|A2) = \frac{a-b-c}{a-b}$$

We must prove that the sum is defined and that it is greater than or equal to 0 and less than or equal to 2 in order for it to appropriately represent an agent's conditional probabilities. First, by definition of the payoff table $a > b$, so the denominator is positive and the expression is defined. Second, $\frac{a-b-c}{a-b} \geq 0$ entails that the numerator is also positive. Note that since $a > b$, this will be satisfied as long as $c \leq a-b$. Of course, this is true from the definition of the asymmetric decision instability problem. If the cost of fleeing was greater than the difference between survival and death, we would not be in a case of asymmetric Death in Damascus as it would never be preferable to flee. Finally, $\frac{a-b-c}{a-b} \leq 2 = a-b-c \leq 2(a-b) = -c \leq a-b$. This is satisfied by definition of the payoff table again, as c is positive and $a > b$ so the left hand side is negative whilst the right is positive.

From this equation for the plane of indifference, we can see that as the cost of fleeing increases, the right hand side of the equation reduces, meaning the plane of indifference will move downwards in the diagonal space of the dynamical cube. This decreases the area of the cube where Tereza's deliberation leads her to flee. In other words, the greater the cost of fleeing, the more sure Tereza must be that Death is in Damascus than that he is in Aleppo in order that rationality decree she purchases the ticket to flee.[19] Now that we have proved the existence of an indifference plane, we can demonstrate how the adaptive dynamics will operate either side of it in a general setting.

Since $a-b$ is positive (the utility of living exceeds that of dying) we can easily replace our equalities in the above existence proof with inequalities. The direction of the inequality does not change throughout the proof. It follows that:

$$P(S2|A1) + P(S2|A2) > \frac{a-b-c}{a-b} \iff EU(A1) > EU(A2)$$

$$P(S2|A1) + P(S2|A2) < \frac{a-b-c}{a-b} \iff EU(A1) < EU(A2)$$

This means that if the agent begins deliberation above the plane, she will end deliberation with $P(A2) = 0$ and if she begins below it, she will end deliberation with $P(A2) = 1$.

Here, one might ask whether her dynamical deliberation could cross over the plane. In principle, it could. However, this would be to violate the plausible stipulation we have made that the ideal deliberator approaches the Eells-Jeffrey manifold via the shortest-path independence dynamics. By definition of how I have specified the shortest-path dynamics, the path toward the manifold is perpendicular to the manifold. This can be seen in Figure 4. We can also prove that the indifference plane is perpendicular to the manifold by showing that the dot product of the normal vectors of both planes is 0. Since the normal vector of a plane is perpendicular to it, it is sufficient to show that the normal vectors are perpendicular to each other in order to show that the planes are perpendicular. The plane of indifference is given by $P(S2|A2) + P(S2|A1) = \frac{a-b-c}{a-b}$ and the Eells-Jeffrey manifold is given by $P(S2|A2) - P(S2|A1) = 0$. The normal vectors are therefore $A = \langle 1, 1 \rangle$ and $B = \langle 1, -1 \rangle$. The dot product is thus $A \cdot B = 0$. The planes are therefore perpendicular and this will hold for any value of $\frac{a-b-c}{a-b}$.

It is clear, therefore, that the shortest-path dynamics decrees dynamical adjustments of conditional probabilities that run parallel to the plane of indifference and do not cross it. Given this feature, one's initial starting point entirely determines the ending point of deliberation. This is true of more general cases than the one considered here, as long as the payoff table bears the same mathematical relationship to the one presented above, where $a > b$ and $c \leq a-b$, and raises important questions for the reconciliation of causal and evidential decision theory for problems of decision instability in Huttegger's deliberative framework.

[19] If $c = 0$, we are in a symmetric decision instability problem where the plane of indifference intersects the Eells-Jeffrey manifold at $P(S2) = 0.5$.

Now let us consider why this problem does not arise in Newcomb's problem. In short, the reason is that the structure of the payoff table renders the plane of indifference *parallel* to the Eells-Jeffrey manifold. This means that, above or below the plane, shortest-path independence dynamics will necessarily pass through it to the Eells-Jeffrey manifold where adaptive dynamics dictates that Tomas takes both boxes. Consider the following generalised payoff table where $a > b$ and $c \leq a - b$. Other problems with a similar structure are the Cholesterol problem, Smoking problem, and Solomon's problem ([22]; [6]; [3]).

	S1	S2
A1	b	a
A2	b + c	a + c

Table 4: Generalised payoff table for Newcomb's Problem

As above, the plane of indifference is found where $EU(A1) = EU(A2)$. This is when

$$bP(S1|A1) + aP(S2|A1) = (b+c)P(S1|A2) + (a+c)P(S2|A2)$$

By substituion and rearranging, we get

$$P(S2|A1) - P(S2|A2) = \frac{c}{a-b}$$

We must prove that the difference is defined and that it lies between -1 and 1 inclusive in order for it to appropriately represent an agent's conditional probabilities. First, by definition of the payoff table $a > b$, so the denominator is positive and the expression is defined. Second, $-1 \leq \frac{c}{a-b} = b - a \leq c$. This is satisfied by definition of the Newcomb payoff table, since if c was strictly less than $b - a$, c would be negative, and there would be no benefit to two-boxing. Finally, $\frac{c}{a-b} \leq 1 = c \leq a - b$. This is again satisfied by the definition of Newcomb payoffs, since if c were strictly greater than $a - b$, this would mean $c + b > a$ and it would therefore always be better to two-box.

Notice here that the relationship that defines the plane is not a sum but a difference. This means that the plane is parallel to the Eells-Jeffrey manifold. This is easily proved by taking the ratio of the components of their normal vectors and showing that they are the same. Indeed, they are both 1. This will hold for any value of $\frac{c}{a-b}$. It will be illuminating to rewrite the above condition as $P(S2|A2) = P(S2|A1) - \frac{c}{a-b}$ so we see the indifference plane sits below the manifold. This is illustrated in Figure 6.

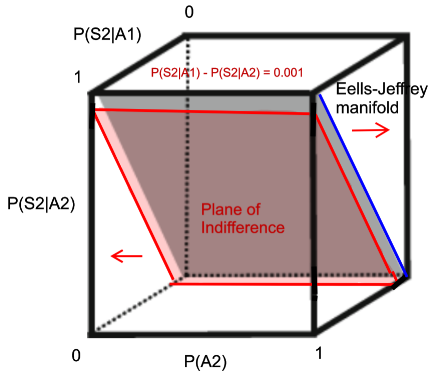


Figure 6: The plane of indifference for the Newcomb problem

The movement decreed by the adaptive dynamics on either side of the plane in the Newcomb problem is given by examining the following biconditional statements. As before, the proof proceeds straightforwardly from the existence proof replacing the equalities with inequalities without any change in direction, as the term $a-b$ is positive.

$$P(S2|A1) > P(S2|A2) + \frac{c}{a-b} \iff EU(A1) > EU(A2)$$

$$P(S2|A1) < P(S2|A2) + \frac{c}{a-b} \iff EU(A1) < EU(A2)$$

We can see from Figure 6 that when $P(S2|A1) > P(S2|A2)$, we are below the Eells-Jeffrey manifold. So when $P(S2|A1) > P(S2|A2) + \frac{c}{a-b}$, we are below the plane of indifference. Here, the biconditional statements above reveal that the rational act according to our adaptive dynamics is to one-box. By analogous reasoning, all points above the indifference plane end deliberation in two-boxing. As the independence dynamics moves the agent towards the Eells-Jeffrey manifold, and the Eells-Jeffrey manifold lies above the indifference plane, the adaptive dynamics decrees that the agent ought to two-box in Newcomb's problem, corroborating Huttegger's conclusion.

As the value of c, the monetary sum under the transparent box, increases, the plane of indifference shifts downward in diagonal space away from the Eells-Jeffrey manifold. As a result, the region of the cube where Tomas should rationally one-box reduces. This is intuitive as, by description of the problem, the agent only receives the value c when he two-boxes, so the greater the value of c, the greater the incentive to two-box. The denominator $a-b$ captures the difference between the contents of the opaque box in the two states of the world. If this difference is large, the plane shifts upwards, expanding the region of points which decree as rational one-boxing. This again is intuitive, as the greater the incentive to one-box, the less sure the agent need be that the predictor put a there in order for him to rationally choose it. Note that when $c = 0$ the plane of indifference is exactly equivalent to the Eells-Jeffrey manifold. It might be tempting to think that if there is nothing under the transparent box, the agent should one-box, but this is not the correct answer. Recall that when we have reached state-act independence, Tomas does not see his act as evidence about the state of the world, so he is rationally indifferent between one-boxing and two-boxing. The causalist answer is the same, as the payoffs are the same under both states of the world.

The preceding discussion has shown that it is the plane of indifference which determines rational action in both decision problems. The crucial difference, however, is that regardless of the exact specification of the independence dynamics, the agent's trajectory of deliberation in Newcomb's problem may pass through the indifference plane to the Eells-Jeffrey manifold, since the two are parallel. This means that where one begins deliberation does not determine where one ends in the same way that it does in the Death in Damascus problem. Here, if we accept the plausibility of the shortest-path independence dynamics, movement toward the manifold never crosses the indifference plane, since the independence path and plane of indifference are parallel to one another. This analysis shows that the relatively straightforward reconciliation of causal and evidential deliberation for Newcomb's problem under Huttegger's deliberative framework is not so straightforwardly achieved in problems of decision instability. Much more would have to be said on the nature of the independence dynamics in order to determine whether we may cross the plane of indifference and end deliberation with a resolute answer. In the next section, I turn to these further requirements.

6 On the possibility of reconciliation

Recall that, under Huttegger's framework, deliberation ends when the adaptive dynamics prescribes no further movement and when we reach state-act independence. In this section, I show that this will only lead to a reconciliation under two very specific conditions: (i) the independence dynamics must be specified such that it does not take the shortest path to the manifold, and (ii) the adaptive dynamics and independence dynamics must have a relative speed such that they reach the Eells-Jeffrey manifold at precisely the point of reconciliation.

As we saw from the previous section, if we take the shortest-path independence dynamics to be true, whether Tereza begins above or below the plane of indifference determines where she will end deliberation. The only time, therefore, where she could end deliberation with $P(A2) = 0.474$ is when she begins with deliberation with her choice probability at $P(A2) = 0.474$ and her conditional probabilities precisely on the plane of indifference (where they sum to 0.9). In this case, shortest-path independence will move her directly to the line of equilibria without any change in her choice probability. This is a case where there appears to be no deliberation at all driving her conclusion, and is therefore implausible as a reconciliation of evidential and causal decision theory via deliberation.

Of course, there may be viable independence dynamics other than shortest-path independence so let us relax this assumption. However, even if we allow violation of shortest-path independence, it must be the case that the relative speed of the adaptive and independence dynamics is such that the agent reaches the Eells-Jeffrey manifold precisely at the point where it intersects the plane of indifference at $P(A2) = 0.474$. If Tereza reaches the manifold on the equilibrium line at any point to the left or right of this, $P(A2) \neq 0.474$ and $\frac{dP(A2)}{dt} = 0$ so we do not achieve reconciliation. If Tereza reaches the manifold at any other point above or below the equilibrium line, the adaptive dynamics leads her to $P(A2) = 0$ or 1 depending on whether this is above or below the plane of indifference. It is only if the two conditions I have specified obtain that we may witness trajectories such as those depicted in Figure 7, but the reconciliation here appears forced.

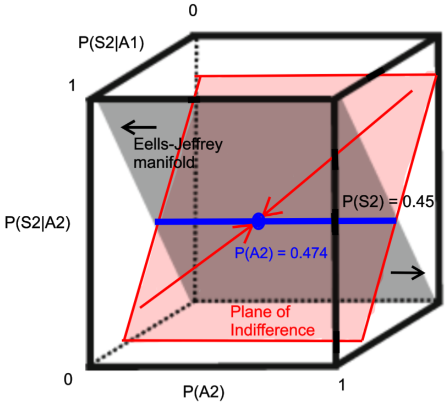


Figure 7: Diagrammatic portrayal of the deliberative evidentialist reasoning under Huttegger's framework. The red arrows represent possible trajectories to reconciliation. Both trajectories cross the plane of indifference where the upper red arrow begins above it and the lower red arrow begins below it.

Again, it is important to recognise that this was not a issue in the case of Newcomb's problem. Here, regardless of the specification of the independence dynamics, since the Eells-Jeffrey manifold lies on the side of the indifference plane where two-boxing is rational, as long as deliberation leads us to state-act independence, the framework will always prescribe the correct answer. The relative strength of the independence and adaptive dynamics may lead Tomas to different points on the line of equilibria where the Eells-Jeffrey manifold intersects the right face of the cube, but this does not change Tomas' ultimate action, as $P(A2) = 1$. Where he concludes deliberation only determines his beliefs about his winnings. That is, he believes himself to be more fortunate if he ends deliberation where the probability of the \$1,000,000 being there, $P(S2)$, is high, and less fortunate if he ends deliberation where it is low.

The analysis I have offered in this section therefore represents a principled way to delineate when the specification of the independence dynamics matters for the reconciliation of evidential and causal decision theory under Huttegger's framework. In particular, it depends on whether the plane of indifference intersects the Eells-Jeffrey manifold or not. If it does not, implying it lies entirely to one side of the Eells-Jeffrey manifold, the specification of the independence dynamics does not matter. Any independence dynamics that moves the agent in the direction of state-act independence over time will lead to the same answer. As is shown from the generalised proofs, for any problem representing the mathematical structure of the generalised Newcomb's problem, the plane of indifference will not intersect the Eells-Jeffrey manifold. For any problem representing the mathematical structure of the generalised Death in Damascus problem, the plane of indifference will be perpendicular to the Eells-Jeffrey manifold, and the specification of the independence dynamics as well as its strength relative to that of the adaptive dynamics, matters for where the agent concludes deliberation. We therefore have a robust way of determining *ex-ante* whether reconciliation of evidential and causal decision theory is plausible for a wide range of two-state, two-act decision problems under this framework.

Note that what is important is not whether the plane of indifference is perpendicular or parallel to the Eells-Jeffrey manifold, but whether it *intersects* the manifold, meaning that the analysis here could in principle be extended to other decision problems, where the angle of the plane of indifference relative to the manifold differs, in order to determine whether specification of the independence dynamics matters in these problems. Furthermore, we would expect the key result – that the relative strength of the adaptive and independence dynamics matters for reconciliation – to hold in larger (*nxn*) decision problems, though this has not as yet been investigated.

7 Conclusion

The prescriptions of evidential and causal decision theory come apart in two general classes of problems known as Newcomb problems and decision instability problems. Huttegger ([11]) has developed a framework for evidential deliberation building on Eells' ([3]) metatickle approach and Skyrms' ([24]) deliberation dynamics which reconciles the prescriptions of the evidentialist and causalist in Newcomb's problem. Since deliberation results in increasing awareness of our beliefs and desires (and these are the mechanisms by which our action is determined), our acts no longer provide information about the state of the world. That is, deliberation screens off the state-act correlation which previously caused the evidentialist to choose the dominated act in Newcomb's problem. Huttegger's more sophisticated, deliberative evidentialist agent agrees with the causalist in preferring two-boxing.

In this paper, I have extended Huttegger's framework to consider an asymmetric case of decision instability: the Death in Damascus problem. I have shown that, in this context, Skyrms' adaptive dynamics and Huttegger's independence dynamics are insufficient to recommend a decisive answer. In Section 5, I consider a plausible version of the independence dynamics, shortest-path independence, and explore the particular features of the deliberative process that this independence dynamics decrees in Death in Damascus. We find that the dynamics decrees different answers for different initial starting points of deliberation. I prove the statements made here are applicable to a more general class of problems of decision instability, as long as the payoff table accords with some simple mathematical relationships. In particular, I show that there exists what I call a *plane of indifference* where either act is equally acceptable, and this plane of indifference entails that where one concludes deliberation depends entirely on where one begins deliberation. This, however, is not true of the Newcomb case.

There are three upshots to this work. First, whilst application of the Eellsian metatickle to deliberation could straightforwardly lead to the correct answer in Newcomb's problem, this notion is not so easily extended to problems of decision instability, and the reconciliation requires assumptions that appear forced. Second, the proof of the plane of indifference for all two-state, two-act problems whose payoff tables exhibit the basic mathematical relationships in Section 5 provides us with a principled way of delineating those cases where the specification of the independence dynamics matters for a reconciliation of evidential and causal decision theory within this framework. Specifically, if the plane of indifference never intersects the Eells-Jeffrey manifold, the specification of the independence dynamics does not matter for reconciliation. If it does, reconciliation requires additional, and potentially questionable, assumptions about the exact specification of the adaptive and independence dynamics.

Finally, this work shows that the metatickle approach has so far failed to reconcile evidential and causal decision theory. Eells' and Jeffrey's original ideas were widely criticised for not providing details of how an agent arrives at knowledge of their own beliefs and desires, involving implicit assumptions, or idealisations that limit the metatickle approach ([11]; [1]; [10]; [14]; [18]; [23]; [24]). Attempts to resolve this using the theory of deliberation have shown it does not result in a reconciliation, rather the evidentialist is left in a state of indecision in Newcomb's problem ([25]). Eells' amendment ([4]) to his original idea then introduced spurious assumptions about other features of the agent, such as her felt urgency to act, which are marked deviations from traditional evidential decision theory ([11]). In this paper, I have shown that the most recent attempt to salvage Eells' idea, owing to Huttegger ([11]) also fails to deliver a reconciliation of evidential and causal decision theory in problems of decision instability. Future work on reconciliation would need to pay heed to the fact that our results will depend heavily on the interaction of the adaptive and independence dynamics, and any attempt at reconciliation would need to specify their relative strength such that evidential decision theory agrees with causal decision theory in both Newcomb and decision instability problems.

References

[1] A. Ahmed (2014): *Evidence, Decision and Causality.* Cambridge University Press, doi:10.1017/CBO9781139107990.

[2] F. Arntzenius (2008): *No Regrets, or: Edith Piaf Revamps Decision Theory.* Erkenntnis 68(2), pp. 277–297, doi:10.1007/s10670-007-9084-8.

[3] E. Eells (1982): *Rational Decision and Causality.* Cambridge University Press.

[4] E. Eells (1984): *Metatickles and the dynamics of deliberation.* Theory and Decision 17, pp. 71–95, doi:10.1007/BF00140057.

[5] A. Egan (2007): *Some Counterexamples to Causal Decision Theory.* Philosophical Review 116(1), pp. 93–114, doi:10.1215/00318108-2006-023.

[6] A. Gibbard & W. Harper (1978[1981]): *Counterfactuals and Two Kinds of Expected Utility.* In C. A. Hooker, J. L. Leach & E. F. McClennan, editors: Foundations and Applications of Decision Theory, University of Western Ontario Series in Philosophy of Science, D. Reidel, pp. 125–162, doi:10.1007/978-94-009-9789-9_-5.

[7] W. Harper: *Ratifiability and Causal Decision Theory: Comments on Eells and Seidenfeld*. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Volume Two: Symposia and Invited Papers*, pp. 213–228.

[8] W. Harper (1986): *Mixed Strategies and Ratifiability in Causal Decision Theory*. *Erkenntnis* 24(1), pp. 25–36, doi:10.1007/BF00183199.

[9] T. Horgan (1981): *Counterfactuals and Newcomb's Problem*. *The Journal of Philosophy* 78(6), pp. 331–356, doi:10.2307/2026128.

[10] P. Horwich (1985): *Decision Theory in the Light of Newcomb's Problem*. *Philosophy of Science* 52(3), pp. 431–450, doi:10.1086/289259.

[11] S. Huttegger (forthcoming): *Reconciling Evidential and Causal Decision Theory*. *Philosopher's Imprint*.

[12] R. C. Jeffrey (1981): *The Logic of Decision Defended*. *Synthese* 48(3), pp. 473–492, doi:10.1007/BF01063989.

[13] R. C. Jeffrey (1983 [1965]): *The Logic of Decision*. University of Chicago Press.

[14] J. Joyce (1999): *The Foundations of Causal Decision Theory*. Cambridge University Press, doi:10.1017/CBO9780511498497.

[15] J. Joyce (2012): *Regret and Instability in Causal Decision Theory*. *Synthese* 187(1), pp. 123–145, doi:10.1007/s11229-011-0022-6.

[16] J. Joyce (2018): *Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems*. In A. Ahmed, editor: *Newcomb's Problem*, Cambridge University Press, pp. 138–159, doi:10.1017/9781316847893.008.

[17] J. Joyce & A. Gibbard (1998): *Causal Decision Theory*. In S. Barbera, P. Hammond & C. Seidl, editors: *Handbook of Utility Theory (Volume 1: Principles)*, Kluwer Acaemic Publishers, pp. 627–666.

[18] D. Lewis (1981): *Causal Decision Theory*. *Australasian Journal of Philosophy* 59(1), pp. 5–30, doi:10.1080/00048408112340011.

[19] R. Nozick (1969): *Newcomb's Problem and Two Principles of Choice*. In N. Rescher, editor: *Essays in Honor of Carl G. Hempel*, Reidel, pp. 114–146, doi:10.1007/978-94-017-1466-2_7.

[20] H. Reichenbach (1959): *Modern Philosophy of Science*. Routledge and Kegan Paul.

[21] L. Savage (1954): *The Foundations of Statistics*. Wiley.

[22] B. Skyrms (1980): *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press.

[23] B. Skyrms (1982): *Causal Decision Theory*. *The Journal of Philosophy* 79(2), pp. 695–711, doi:10.2307/2026547.

[24] B. Skyrms (1984): *Pragmatics and Empiricism*. Yale University Press.

[25] B. Skyrms (1990): *The Dynamics of Rational Deliberation*. Harvard University Press.

[26] R. C. Stalnaker (1968): *A Theory of Conditionals*. In N. Rescher, editor: *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*, Blackwell, pp. 98–112.

[27] P. Weirich: *Causal Decision Theory*. In E. N. Zalta, editor: *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*. Available at `https://plato.stanford.edu/archives/win2016/entries/decision-causal/`.

Tableaux for the Logic of Strategically Knowing How

Yanjun Li

College of Philosophy, Nankai University, Tianjin, China

The logic of goal-directed *knowing how* proposed in [5] extends the standard epistemic logic with an operator of *knowing how*. The *knowing how* operator is interpreted as that there exists a strategy such that the agent knows that the strategy can make sure that φ. This paper presents a tableau procedure for the multi-agent version of the logic of strategically *knowing how* and shows the soundness and completeness of this tableau procedure. This paper also shows that the satisfiability problem of the logic can be decided in PSPACE.

1 Introduction

Epistemic logic proposed by von Wright and Hintikka (see [24, 11]) is a logical formalism for reasoning about knowledge of agents. It deals with propositional knowledge, that is, the knowledge expressed as *knowing that* φ is true. In recent years, other patterns of knowledge besides knowing that are attracting increasing attention in logic community, such as *knowing whether* [8, 4], *knowing who* [3], *knowing the value* [2, 6], and *knowing why* [28] (see a survey in [27]). Motivated by different scenarios in philosophy and AI, reasoning about *knowing how* assertions are particularly interesting [23].

The discussion about formalizing the notion of *knowing how* can date back to [16, 17]. Currently, there are two main approaches of formalizing *knowing how*. One of them is connecting *knowing how* with logics of *knowing that* and *ability* (see e.g. [13, 10]). However, the main difficulty of this approach is that a simple combination of the modalities of *knowing that* and *ability* does not seem to capture a natural understanding of *knowing how* (see a discussion in [12, 9]). Instead of expressing *knowing how* by *knowing that* and *ability* modalities, the other approach first adopted in [25] is expressing *knowing how* in modal languages with a new modality of *knowing how*.

Inspired by the idea of automated planning under uncertainty in artificial intelligence, The author of [25, 26] proposed a logical framework of *knowing how* which includes a binary modality of *knowing how*. In [1], a new semantics for this *knowing how* modality was given, which is based on an indistinguishability relation between plans. It is shown in [1] that the satisfiability problem of this *knowing how* logic is NP-complete, but this logic does not include the modality for *knowing that*.

Inspired by the tradition of coalition logic, the authors of [18, 21] introduced a logic to capture *knowing how* of coalitions. A coalition C knows how to achieve φ if and only if there is a joint action a for C such that it is distributed knowledge for C that doing a can make sure that φ. Variants of the basic framework were proposed and discussed. In [20], a logic of *knowing how* under the assumption of perfect recall was studied. In [22], a logic of *knowing how* with the degree of uncertainty was discussed. In [19], a logic of second-order *knowing how* was proposed, for the case that one coalition knows what the joint action of another coalition is. The topic of complexity is not covered in these literatures.

Along with the idea of formalizing *knowing how* based on planning, the authors of [5] proposed a single-agent logic of *knowing how* via strategies. A strategy is a partial function from the set of agents' belief states into the set of actions. The agent knows how to achieve φ if and only if there is a strategy such that all executions of the strategy will terminate on φ-states. Besides strategies, there are other

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 379–391, doi:10.4204/EPTCS.379.30

types of plans, such as simple plans (i.e. a single action), linear plans (i.e. a sequence of actions), and so on. The authors of [15] proposed a unified logical framework to incorporate logics of *knowing how* via different notions of plans. They used a PDL-style programming language to syntactically specify various types of plans and discussed ten types of *knowing how* based on ten different notions of plans. It is shown in [15] that the ten notions of plans lead to the same *konwing how* logic, but over finite models, the *konwing how* logic based on knowledge-based plans requires an extra axiom, which leads to the same logic as the logic of strategically *knowing how*.

In [14], a tableau-based decision procedure was proposed for the logic of *knowing how* via simple plans. This paper develops the method and presents a tableau procedure for the *knowing how* logic via strategies proposed in [5]. Strategically *knowing how* can not be handled by the original method, since strategies are much more complicated than simple plans. This paper also shows that the satisfiability problem of the logic of strategically *knowing how* is in PSPACE. With other known results, this leads to the result that the satisfiability problem of the logic of strategically *knowing how* is PSPACE-complete.

The structure of this paper is as follows: Section 2 recalls the logic of strategically *knowing how*; Section 3 presents a tableau procedure for the logic and proves its soundness; Section 4 shows the completeness of the tableau procedure and proves that the complexity of this logic is PSPACE-complete. Section 5 concludes with some remarks.

2 The logic of strategically knowing how

This section presents the multi-agent version of the logic of strategically *knowing how* from [5].

Let $\mathbf{P}$ be a set of propositional letters and $\mathbf{I}$ be a set of agents where $|\mathbf{I}| \geq 2$.

Definition 1 (ELKh_n Language)**.** *The Epistemic Language $\mathscr{L}_{\mathsf{ELKh}_n}$ of* Knowing how *is defined by the following BNF where $p \in \mathbf{P}$ and $i \in \mathbf{I}$:*

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid Kh_i\varphi.$$

We use $\top, \vee, \to$ as usual abbreviations. The formula $K_i\varphi$ means that the agent i knows that φ, and the formula $Kh_i\varphi$ means that the agent knows how to achieve the goal that φ.

Definition 2 (ELKh_n Models)**.** *A model M is a quintuple $\langle W, \{\sim_i \mid i \in \mathbf{I}\}, \{\mathbf{A}_i \mid i \in \mathbf{I}\}, \{R_a \mid a \in \mathbf{A}_i, i \in \mathbf{I}\}, V\rangle$ where:*

- *W is a non-empty set of states,*
- *$\sim_i \subseteq W \times W$ is an equivalence relation for each $i \in \mathbf{I}$,*
- *$\mathbf{A}_i$ is a set of actions for each $i \in \mathbf{I}$,*
- *$R_a \subseteq W \times W$ is a binary relation for each $a \in \mathbf{A}$ where $\mathbf{A} = \bigcup_{i \in \mathbf{I}} \mathbf{A}_i$,*
- *$V : W \to 2^{\mathbf{P}}$ is a valuation function.*

Given $s \in W$, we use $[s]^i$ to denote the equivalence class of s over $\sim_i$, i.e., $[s]^i = \{t \in W \mid s \sim_i t\}$, and use $[W]^i$ to denote the set of all equivalence classes of states in W over $\sim_i$, namely $[W]^i = \{[s]^i \mid s \in W\}$. We say that the action $a \in \mathbf{A}_i$ is *executable* at s if $(s,t) \in R_a$ for some t. We use $[s]^i \xrightarrow{a} [t]^i$ to denote that there are some $s' \in [s]^i$ and some $t' \in [t]^i$ such that $(s',t') \in R_a$.

Definition 3 (Strategies)**.** *Given a model M, a* uniformly executable strategy *(or simply called* strategy*) for agent i in M is a partial function $\sigma : [W]^i \to \mathbf{A}_i$ such that $\sigma([s]^i)$ is executable at all $s' \in [s]^i$. Particularly, the empty function is also a strategy, the* empty strategy.

We use $\mathtt{dom}(\sigma)$ to denote the domain of σ.

Definition 4 (Executions)**.** *Given a strategy σ of agent i in M, a* possible execution *of σ is a possibly infinite sequence of equivalence classes $\delta = [s_0]^i[s_1]^i \cdots$ such that $[s_k]^i \xrightarrow{\sigma([s_k]^i)} [s_{k+1}]^i$ for all $0 \leq k < |\delta|$. Particularly, $[s]^i$ is a possible execution if $[s]^i \notin \mathtt{dom}(\sigma)$. If the execution ρ is a finite sequence $[s_0]^i \cdots [s_k]^i$, we call $[s_k]^i$ the end node of ρ. A possible execution of σ is* complete *if it is infinite or its end node is not in $\mathtt{dom}(\sigma)$.*

Given an i-strategy σ, we use $\mathtt{ECE}(\sigma, [s]^i)$ to denote the set of all end nodes of all σ's complete executions starting from $[s]^i$.

Definition 5 (ELKh_n Semantics)**.** *The satisfaction relation $\vDash$ between a pointed model (M,s) and a formula φ is defined as follows:*

$$\begin{array}{lcl}
M,s \vDash p & \Longleftrightarrow & s \in V(p) \\
M,s \vDash \neg\varphi & \Longleftrightarrow & M,s \nvDash \varphi \\
M,s \vDash \varphi \wedge \psi & \Longleftrightarrow & M,s \vDash \varphi \text{ and } M,s \vDash \psi \\
M,s \vDash K_i\varphi & \Longleftrightarrow & \text{for all } s' : \text{ if } s \sim_i s' \text{ then } M,s' \vDash \varphi \\
M,s \vDash Kh_i\varphi & \Longleftrightarrow & \text{there exists a strategy } \sigma \text{ for agent } i \text{ such that} \\
 & & 1. \text{ all } \sigma\text{'s complete executions starting from } [s]^i \text{ are finite, and} \\
 & & 2. [t]^i \subseteq \{s' \in W \mid M,s' \vDash \varphi\} \text{ for all } [t]^i \in \mathtt{ECE}(\sigma, [s]^i).
\end{array}$$

Proposition 6. *The following formulas are valid.*

(1). $K_i\varphi \to Kh_i\varphi$

(2). $Kh_i\varphi \to K_iKh_i\varphi$

(3). $\neg Kh_i\varphi \to K_i\neg Kh_i\varphi$

(4). $Kh_i\varphi \to Kh_iK_i\varphi$

(5). $Kh_iKh_i\varphi \to Kh_i\varphi$

Proof. For (1), it is due to the empty strategy. For (2), (3), and (4), it follows from the semantics. For (5), see [5]. □

3 Tableaux

This section presents a tableau procedure for the logic ELKh_n and shows the soundness of the tableau procedure.

Given φ, let $sub^+(\varphi)$ be the set $\{\psi, \neg\psi \mid \psi$ is a subformula of $\varphi\} \cup \{K_i\psi, \neg K_i\psi \mid Kh_i\psi$ is a subformula of $\varphi\}$.

3.1 Tableau procedure

A tableau is a rooted tree in which each node is labeled with a set of prefixed formulas. A prefixed formula is a pair $\langle \sigma, \varphi \rangle$ where the prefix σ is an alternative sequence of natural numbers and agents or Kh-formulas, such as $\langle 1i2Kh_ip3, \varphi \rangle$. The prefixes represent states in models. The agent symbols and Kh-formulas occurring in σ indicate some epistemic information and action information on the current state. For example, the prefix $1i2Kh_ip3$ indicates the following informations: there are three states 1, 2, and 3; $1 \sim_i 2$; there is an action a such that a is a good plan for $2 \vDash Kh_ip$ and $(2,3) \in R_a$.

Definition 7 (Tableaux)**.** *A tableau for φ_0 is a labeled tree that is defined as follows:*

A. Create the root node and label it with $\langle 0, \varphi_0 \rangle$;

B. Extend the tree by rules in Table 1.

$$(R\neg)\ \frac{\langle \sigma, \neg\neg\varphi \rangle}{\langle \sigma, \varphi \rangle}$$

$$(R\vee)\ \frac{\langle \sigma, \neg(\varphi_1 \wedge \varphi_2) \rangle}{\begin{array}{c|c|c} \langle \sigma, \neg\varphi_1 \rangle & \langle \sigma, \neg\varphi_1 \rangle & \langle \sigma, \varphi_1 \rangle \\ \langle \sigma, \neg\varphi_2 \rangle & \langle \sigma, \varphi_2 \rangle & \langle \sigma, \neg\varphi_2 \rangle \end{array}} \qquad (R\wedge)\ \frac{\langle \sigma, \varphi_1 \wedge \varphi_2 \rangle}{\begin{array}{c} \langle \sigma, \varphi_1 \rangle \\ \langle \sigma, \varphi_2 \rangle \end{array}}$$

$$(\mathrm{Cut}\neg K)\ \frac{\langle \sigma, \neg K_i\varphi \rangle}{\langle \sigma, \neg\varphi \rangle \mid \langle \sigma, \varphi \rangle} \qquad (CutK)\ \frac{\langle \sigma, K_i\varphi \rangle}{\langle \sigma, \varphi \rangle}$$

$$(\mathrm{Cut}Kh)\ \frac{\langle \sigma, Kh_i\varphi \rangle}{\langle \sigma, \neg K_i\varphi \rangle \mid \langle \sigma, K_i\varphi \rangle} \qquad (\mathrm{Cut}\neg Kh)\ \frac{\langle \sigma, \neg Kh_i\varphi \rangle}{\langle \sigma, \neg K_i\varphi \rangle}$$

$$(R\neg K)\ \frac{\langle \sigma, \neg K_i\varphi \rangle}{\langle \sigma i n', \neg\varphi \rangle}\ n' \text{ is new} \qquad (RK)\ \frac{\langle \sigma, K_i\varphi \rangle}{\langle \sigma i n', \varphi \rangle}\ \sigma i n' \text{ is used}$$

$$(RK4)\ \frac{\langle \sigma, K_i\varphi \rangle}{\langle \sigma i n', K_i\varphi \rangle}\ \sigma i n' \text{ is used} \qquad (RK5)\ \frac{\langle \sigma, \neg K_i\varphi \rangle}{\langle \sigma i n', \neg K_i\varphi \rangle}\ \sigma i n' \text{ is used}$$

$$(RKh4)\ \frac{\langle \sigma, Kh_i\varphi \rangle}{\langle \sigma i n', Kh_i\varphi \rangle}\ \sigma i n' \text{ is used} \qquad (RKh5)\ \frac{\langle \sigma, \neg Kh_i\varphi \rangle}{\langle \sigma i n', \neg Kh_i\varphi \rangle}\ \sigma i n' \text{ is used}$$

$$(RKh)\ \frac{\begin{array}{c} \langle \sigma, Kh_i\varphi \rangle \\ \langle \sigma, \neg K_i\varphi \rangle \end{array}}{\langle \sigma Kh_i\varphi n', K_i\varphi \rangle}\ n' \text{ is new} \qquad (R\neg Kh)\ \frac{\begin{array}{c} \langle \sigma, \neg Kh_i\varphi \rangle \\ \langle \sigma, Kh_i\psi \rangle \\ \langle \sigma, \neg K_i\psi \rangle \end{array}}{\begin{array}{c} \langle \sigma Kh_i\psi n', K_i\psi \rangle \\ \langle \sigma Kh_i\psi n', \neg Kh_i\varphi \rangle \end{array}}\ n' \text{ is new}$$

Table 1: Tableau rules

Next, we will give a procedure to construct a tableau (Definition 8 below). In Section 3.2, we will show that the procedure is sound, and we in Section 4.1 will show that it is complete, and we in Section 4.2 will show that it runs in polynomial space. Before that, we first introduce some auxiliary notations below.

Let $\mathbf{A}$ be a set of actions. We use $\mathbf{A}^+$ to denote the set $\mathbf{A} \cup \{\varepsilon\}$. Let Γ be a set of formulas. We use $\Gamma|K_i$, $\Gamma|\neg K_i$ and $\Gamma|Kh_i$ to respectively denote the set $\{\varphi \in \Gamma \mid \varphi \text{ is of the form } K_i\psi\}$, $\{\varphi \in \Gamma \mid \varphi \text{ is of the form } \neg K_i\psi\}$ and $\{\varphi \in \Gamma \mid \varphi \text{ is of the form } Kh_i\psi\}$. We say that a formula set Γ is *blatantly inconsistent* iff either $\varphi, \neg\varphi \in \Gamma$ for some formula φ or $\bot \in \Gamma$.

A labeled tree $\mathscr{T}$ is a triple $\langle N, E, L \rangle$, where $\langle N, E \rangle$ is a rooted tree with the node set N and the edge set E and $L : N \cup E \to \mathscr{P}(\mathscr{L}_{\mathsf{ELKh}_n}) \cup \mathbf{I} \cup \mathbf{A}^+$ is a label function such that each node is labeled a formula set and each edge is labeled an agent $i \in \mathbf{I}$ or an action $a \in \mathbf{A}^+$. A node sequence $n_1 \cdots n_{h+1}$ is a *path* in $\mathscr{T}$ if $(n_k, n_{k+1}) \in E$ for all $1 \leq k \leq h$. If the node sequence $n_1 \cdots n_{h+1}$ is a path in $\mathscr{T}$ and the label $L(n_k, n_{k+1})$ is either $i \in \mathbf{I}$ or ε for all $1 \leq k \leq h$, we then say that n_1 is an *i-ancestor* of n_{h+1}.

A tree is called an *and-or* tree if each non-leaf node is marked as “and” node or “or” node. A subtree of an and-or tree is called *complete* if it contains the root node, and each “and” non-leaf node has all its child nodes, and each “or” non-leaf node has at least one child node.

Now we are ready to give a procedure to construct a tableau. We remark that, strictly speaking, the tree constructed by the following procedure is not really a tableau. Rather, it is a tree in which the desired tableau is embedded. Such trees are called *pre-tableaux* in [7]. Since in the remaining paper, we will work only on the following procedure and show that the following procedure is sound and complete and runs in polynomial space, it does not matter what we call it. So, in the remaining paper, we will call the tree constructed by the following procedure a tableau.

Definition 8 (Tableaux construction). *A tableau for φ_0 is a labeled and-or tree $\mathscr{T}_{\varphi_0}$ which is constructed by the following steps:*

(I). Construct a tree consisting of a single node n_0 (i.e. the root node), and label the root node the formula set $\{\varphi_0\}$.

(II). Repeat until none of (1)-(2) below applies:

(1). Forming a subformula-closed propositional tableau: if n is an unblocked leaf node and $L(n)$ is not blatantly inconsistent, then mark n as an “or” node and check the first unchecked formula $\varphi \in L(n)$ at n by the following:

(a). If φ is of the form $\neg\neg\psi$ and ψ is not in $L(n)$, then create a successor node n' of n, set

$$\begin{aligned} L(n') &= L(n) \cup \{\psi\}, \\ L(n,n') &= \varepsilon, \end{aligned}$$

and mark φ and all checked formulas at n as “checked” at n'.

(b). If φ is of the form $\varphi_1 \wedge \varphi_2$ and either φ_1 or φ_2 is not in $L(n)$, then create a successor node n' of n, set

$$\begin{aligned} L(n') &= L(n) \cup \{\varphi_1, \varphi_2\}, \\ L(n,n') &= \varepsilon, \end{aligned}$$

and mark φ and all checked formulas at n as “checked” at n'

(c). If φ is of the form $\neg(\varphi_1 \wedge \varphi_2)$ and none of the three sets $\{\neg\varphi_1, \neg\varphi_2\}$, $\{\neg\varphi_1, \neg\varphi_2\}$, $\{\neg\varphi_1, \neg\varphi_2\}$ is a subset of $L(n)$, then create three successors n_1, n_2, n_3 of n, set

$$\begin{aligned} L(n_1) &= L(n) \cup \{\neg\varphi_1, \neg\varphi_2\}, \\ L(n_2) &= L(n) \cup \{\neg\varphi_1, \varphi_2\}, \\ L(n_3) &= L(n) \cup \{\varphi_1, \neg\varphi_2\}, \\ L(n,n_1) &= L(n,n_2) = L(n,n_3) = \varepsilon, \end{aligned}$$

and mark φ and all checked formulas at n as “checked” at n_1, n_2, n_3

(d). If φ is of the form $K_i\psi$ and ψ is not in $L(n)$, then create a successor node n' of n, and set

$$\begin{aligned} L(n') &= L(n) \cup \{\psi\}, \\ L(n,n') &= \varepsilon, \end{aligned}$$

and mark φ and all checked formulas at n as “checked” at n'

(e). If φ is of the form $\neg K_i\psi$ and neither $\neg\psi$ nor ψ is in $L(n)$, then create two successors n_1, n_2 of n, set

$$\begin{aligned}&L(n_1) = L(n) \cup \{\neg\psi\},\\ &L(n_2) = L(n) \cup \{\psi\},\\ &L(n,n_1) = L(n,n_2) = \varepsilon\end{aligned}$$

and mark φ and all checked formulas at n as "checked" at n_1 and n_2

(f). If φ is of the form $Kh_i\psi$ and neither $\neg K_i\psi$ nor $K_i\psi$ is in $L(n)$, then create two successors n_1, n_2 of n, set

$$\begin{aligned}&L(n_1) = L(n) \cup \{\neg K_i\psi\},\\ &L(n_2) = L(n) \cup \{K_i\psi\},\\ &L(n,n_1) = L(n,n_2) = \varepsilon,\end{aligned}$$

and mark φ and all checked formulas at n as "checked" at n_1 and n_2.

(g). If φ is of the form $\neg Kh_i\psi$ and $\neg K_i\psi$ is not in $L(n)$, then create a successor n' of n, set

$$\begin{aligned}&L(n') = L(n) \cup \{\neg K_i\psi\},\\ &L(n,n') = \varepsilon,\end{aligned}$$

and mark φ and all checked formulas at n as "checked" at n'

(h). if none of (a)-(g) above applies, then mark φ as "checked".

(2). Creating successors for $\neg K_i, Kh_i, \neg Kh_i$ formulas: if n is an unblocked leaf node, $L(n)$ is not blatantly inconsistent, and each formula in $L(n)$ is marked as "checked", then mark (or re-mark) n as an "and" node and do the following:

(i). For each formula in $L(n)$ of the form $\neg K_i\psi$, if there is no i-ancestor of n' such that $L(n') = \Sigma(n, \neg K_i\psi)$, then create a successor $n_{\neg K_i\psi}$ of n and set

$$\begin{aligned}&L(n_{\neg K_i\psi}) = \Sigma(n, \neg K_i\psi),\\ &L(n, n_{\neg K_i\psi}) = i\end{aligned}$$

in which

$$\Sigma(n, \neg K_i\psi) = \{\neg\psi\} \cup (L(n)|K_i) \cup (L(n)|\neg K_i) \cup (L(n)|Kh_i) \cup (L(n)|\neg Kh_i).$$

(j). For each pair on $L(n)$ of the form $(Kh_i\psi, \neg K_i\psi)$, create a successor $n_{Kh_i\psi}$ of n, and set

$$\begin{aligned}&L(n_{Kh_i\psi}) = \{K_i\psi\},\\ &L(n, n_{Kh_i\psi}) = a_{Kh_i\psi}.\end{aligned}$$

(k). For each triple on $L(n)$ of the form $(\neg Kh_i\chi, Kh_i\psi, \neg K_i\psi)$, create a successor $n_{(\neg Kh_i\chi, Kh_i\psi)}$, and set

$$\begin{aligned}&L(n_{(\neg Kh_i\chi, Kh_i\psi)}) = \{K_i\psi, \neg Kh_i\chi\},\\ &L(n, n_{(\neg Kh_i\chi, Kh_i\psi)}) = a_{Kh_i\psi}.\end{aligned}$$

Moreover, if there is an ancestor n' of $n_{(\neg Kh_i\chi, Kh_i\psi)}$ such that $L(n') = L(n_{(\neg Kh_i\chi, Kh_i\psi)})$, then mark $n_{(\neg Kh_i\chi, Kh_i\psi)}$ as blocked, and we say that $n_{(\neg Kh_i\chi, Kh_i\psi)}$ is blocked by n'.

Definition 9. *A subtree of a tableau is* closed *if there is some node n in it such that $L(n)$ is blatantly inconsistent. Otherwise, it is called* open. *A tableau is* closed *iff all its complete subtrees are closed.*

3.2 Soundness

In this subsection, we will show that the procedure of Definition 8 is sound.

Definition 10 (Interpretations)**.** *Given a model M and a subtree $\mathscr{T}' = \langle N', E', L' \rangle$ of the tableau $\mathscr{T}_{\varphi_0}$, let f be a function from N' to W. We say that f is an* interpretation *of $\mathscr{T}'$ if and only if $M, f(n) \vDash \varphi$ for all $\varphi \in L(n)$ and all $n \in N'$.*

Lemma 11. *If $M, s \vDash \varphi_0$, then there exists an interpretation of some complete subtree of any $\mathscr{T}_{\varphi_0}$.*

Proof. Let f_0 be the function $f_0 = \{n_0 \mapsto s\}$ which maps the root n_0 of $\mathscr{T}_{\varphi_0}$ to the state s. It is obvious that $M, f_0(n_0) \vDash \varphi$ for all $\varphi \in L(n_0)$.

Then we only need to show the following statement:

> If n is a leaf node of a subtree $\mathscr{T}$, f is an interpretation of $\mathscr{T}$, and n is in the domain of f, then
> (A): for each construction steps (a)-(g), there is an interpretation of $\mathscr{T}$ extending with one child node of n, and
> (B): for each construction steps (i)-(k), there is an interpretation of $\mathscr{T}$ extending with all child nodes of n.

For (A), firstly it is obvious for the steps (a)-(c). For the step (d), it follows that $\sim_i$ is a reflexive relation. For the steps (e) and (f), it follows from the fact that a formula is either true or false on a state. For the step (g), it follows from the fact that $\neg Kh_i\psi \to \neg K_i\psi$ is valid (see Proposition 6).

Next, we will show that (B) holds.

For the step (i), for each n's child node $n_{\neg K_i\psi}$, we know that $\neg K_i\psi \in L(n)$ and $L(n_{\neg K_i\psi}) = \{\neg\psi\} \cup (L(n)|K_i) \cup (L(n)|\neg K_i) \cup (L(n)|Kh_i) \cup (L(n)|\neg Kh_i)$. Since f is an interpretation of $\mathscr{T}$ including n, it follows that $M, f(n) \vDash \neg K_i\psi$. Hence, there exists a state $t_{\neg K_i\psi} \in [f(n)]^i$ such that $M, t_{\neg K_i\psi} \vDash \neg\psi$. Moreover, since $\sim_i$ is an equivalence relation, $(M, t_{\neg K_i\psi})$ satisfies all the K_i-formulas and $\neg K_i$-formulas that are true at $f(n)$. Furthermore, by Proposition 6, we have that $(M, t_{\neg K_i\psi})$ satisfies all the Kh_i-formulas and $\neg Kh_i$-formulas that are true at $f(n)$. Let f' be the f-extension $f \cup \{n_{\neg K_i\psi} \mapsto t_{\neg K_i\psi} \mid n_{\neg K_i\psi}$ is a child node of $n\}$. Therefore, we have that $M, f'(n_{\neg K_i\psi}) \vDash L(n_{\neg K_i\psi})$.

For the step (j), for each n's child node $n_{Kh_i\psi}$, we know that $Kh_i\psi \in L(n)$ and $L(n_{\neg K_i\psi}) = \{K_i\psi\}$. Since f is an interpretation of $\mathscr{T}$ including n, it follows that $M, f(n) \vDash Kh_i\psi$. So, by the semantics, there exist an i-strategy σ and a state $t_{Kh_i\psi} \in \mathtt{ECE}(\sigma, [f(n)]^i)$ such that $M, t_{Kh_i\psi} \vDash K_i\psi$. Let f' be the f-extension $f \cup \{n_{Kh_i\psi} \mapsto t_{Kh_i\psi} \mid n_{Kh_i\psi}$ is a child node of $n\}$. Therefore, we have that $M, f'(n_{Kh_i\psi}) \vDash L(n_{Kh_i\psi})$.

For the step (k), for each n's child node $n_{(\neg Kh_i\chi, Kh_i\psi)}$, we know that $\neg Kh_i\chi, Kh_i\psi \in L(n)$ and $L(n_{(\neg Kh_i\chi, Kh_i\psi)}) = \{K_i\psi, \neg Kh_i\chi\}$. Since f is an interpretation of $\mathscr{T}$ including n, it follows that $M, f(n) \vDash \neg Kh_i\chi \wedge Kh_i\psi$. Due to $M, f(n) \vDash Kh_i\psi$, it follows by the semantics that there exist an i-strategy σ such that $M, t \vDash K_i\psi$ for all $[t]^i \in \mathtt{ECE}(\sigma, [f(n)]^i)$. Moreover, it must be the case that there exists $t_{(\neg Kh_i\chi, Kh_i\psi)}$ such that $[t_{(\neg Kh_i\chi, Kh_i\psi)}]^i \in \mathtt{ECE}(\sigma, [f(n)]^i)$ and $M, t_{(\neg Kh_i\chi, Kh_i\psi)} \vDash \neg Kh_i\chi$. Otherwise, if $M, t \vDash Kh_i\chi$ for all $[t]^i \in \mathtt{ECE}(\sigma, [f(n)]^i)$, this implies $M, f(n) \vDash KhKh_i\chi$. By Proposition 6, it follows that $M, f(n) \vDash Kh_i\chi$, which is contradictory with the fact that $M, f(n) \vDash \neg Kh_i\chi$. Hence, there exists $t_{(\neg Kh_i\chi, Kh_i\psi)}$ such that $[t_{(\neg Kh_i\chi, Kh_i\psi)}]^i \in \mathtt{ECE}(\sigma, [f(n)]^i)$ and $M, t_{(\neg Kh_i\chi, Kh_i\psi)} \vDash \neg Kh_i\chi$, which implies that $M, t_{(\neg Kh_i\chi, Kh_i\psi)} \vDash \neg Kh_i\chi \wedge K_i\psi$. Let f' be the f-extension $f \cup \{n_{(\neg Kh_i\chi, Kh_i\psi)} \mapsto t_{(\neg Kh_i\chi, Kh_i\psi)} \mid n_{(\neg Kh_i\chi, Kh_i\psi)}$ is a child node of $n\}$. Therefore, we have that $M, f'(n_{(\neg Kh_i\chi, Kh_i\psi)}) \vDash L(n_{(\neg Kh_i\chi, Kh_i\psi)})$. □

The soundness below follows from Lemma 11 above.

Theorem 12 (Soundness)**.** *If φ_0 is satisfiable, then $\mathscr{T}_{\varphi_0}$ is not closed.*

4 Completeness and complexity

4.1 Completeness

In this subsection, we will show that the procedure of Definition 8 is complete.

Recall that $sub^+(\varphi)$ is the set $\{\psi, \neg\psi \mid \psi$ is a subformula of $\varphi\} \cup \{K_i\psi, \neg K_i\psi \mid Kh_i\psi$ is a subformula of $\varphi\}$. From the construction of $\mathscr{T}_{\varphi_0}$, it follows that $L(n) \subseteq sub^+(\varphi_0)$ for each node n. Moreover, each formula in $L(n)$ is "inherited" from n's ancestors, that is, if $\psi \in L(n)$ and n' is an ancestor of n then there exists $\varphi \in L(n')$ such that $\psi \in sub^+(\varphi)$.

A path $n_1 \cdots n_h$ of $\mathscr{T}_{\varphi_0}$ is called an ε-path iff $L(n_k, n_{k+1}) = \varepsilon$ for all $1 \leq k < h$. Particularly, a path with length 1 is an ε-path. An ε-path $n_1 \cdots n_h$ is *maximal* iff there are no such nodes n and n' that either $nn_1 \cdots n_h$ or $n_1 \cdots n_h n'$ is an ε-path. Given a path $\rho = n_1 \cdots n_h$, we use $ini(\rho)$, $end(\rho)$ and $L(\rho)$ to respectively denote n_1, n_h and $L(n_h)$. We use $\rho \xrightarrow{x} \rho'$ to means that $\rho\rho'$ is a path and $L(end(\rho), ini(\rho')) = x$.

From the construction of $\mathscr{T}_{\varphi_0}$ we know that if a node n is blocked then there is a unique node n' which blocks n and n itself is a maximal ε-path. Moreover, if ρ is a maximal ε-path and $end(\rho)$ is blocked, then ρ is a single node, i.e. $\rho = n$ where the node n is blocked.

For each maximal ε-path ρ of $\mathscr{T}_{\varphi_0}$, by the construction, we know that if ρ is not blocked and $L(\rho)$ is not blatantly inconsistent then each formula in $L(\rho)$ is marked as "checked". This means that $L(\rho)$ is closed over sub^+-formulas, that is, if $\varphi \in L(\rho)$ then either ψ or $\sim \psi$ is in $L(\rho)$ for all $\psi \in sub^+(\varphi)$, where $\sim \psi = \chi$ if $\psi = \neg\chi$, otherwise $\sim \psi = \neg\psi$.

Definition 13 ($M^{\mathscr{T}}$). *Let $\mathscr{T}$ be a complete subtree of $\mathscr{T}_{\varphi_0}$. The model induced by $\mathscr{T}$, denoted by $M^{\mathscr{T}}$, is defined as follows:*

- $W = \{\rho \mid \rho$ *is a maximal ε-path of $\mathscr{T}$, and ρ is not blocked*$\}$,
- $\sim_i$ *is the minimal equivalence relation X on W such that* $\{(\rho, \rho') \mid \rho \xrightarrow{i} \rho'\} \subseteq X$,
- $\mathbf{A}_i = \{a_{Kh_i\psi} \mid$ *there exists an edge (n, n') in $\mathscr{T}$ such that* $L(n, n') = a_{Kh_i\psi}\}$.
- *for each $a \in \mathbf{A}_i$,* $R_a = \{(\rho, \rho') \mid \rho \xrightarrow{a} \rho'$, *or* $\rho \xrightarrow{a} \rho''$ *where ρ'' is a maximal ε-path blocked by* $\rho'\}$.
- $f(\rho) \in V(p)$ *iff* $p \in L(f(\rho))$.

Please note that by the definition $\mathbf{A}_i$, we know that if $i \neq j$ then $\mathbf{A}_i \cap \mathbf{A}_j = \emptyset$.

The following proposition states that all paths in the same $\sim_i$-closure share the same formulas in the forms $K_i\varphi$, $\neg K_i\varphi$, $Kh_i\varphi$, or $\neg Kh_i\varphi$.

Proposition 14. *Let $\mathscr{T}$ be a complete and open subtree of $\mathscr{T}_{\varphi_0}$. If $\rho_1 \sim_i \rho_2$ in $M^{\mathscr{T}}$, then $L(\rho_1)|x = L(\rho_2)|x$ where $x \in \{K_i, \neg K_i, Kh_i, \neg Kh_i\}$.*

Proof. Besides $\rho_1 = \rho_2$, there are three possible cases: (1) $end(\rho_1)$ is an i-ancestor of $end(\rho_2)$, or (2) $end(\rho_2)$ is an i-ancestor of $end(\rho_1)$, or (3) there is some $\rho \in W$ such that $end(\rho)$ is an i-ancestor of both $end(\rho_1)$ and $end(\rho_2)$.

For (1), firstly we will show that $L(\rho_1)|x \subseteq L(\rho_2)|x$. In the construction step (i) all x-formulas are inherited from the i-ancestor. Moreover, in the steps (a)-(f) all formulas labeled on the parent node are inherited by each child node. Thus, $L(\rho_1)|x \subseteq L(\rho_2)|x$.

Secondly, we will show that $L(\rho_2)|x \subseteq L(\rho_1)|x$. Since $end(\rho_1)$ is an ancestor of $end(\rho_2)$, we know from the construction that for each formula $\psi \in L(\rho_2)$ there exists $\varphi \in L(\rho_1)$ such that $\psi \in sub^+(\varphi)$. Moreover, since ρ_1 is a non-blocked maximal ε-path and $L(\rho_1)$ is not blatantly inconsistent, it follows that all formulas in $L(\rho_1)$ are marked as "checked". So, $L(\rho_1)$ is closed over sub^+-formulas. Hence, for each x-formula $\chi \in L(\rho_2)$, either χ or $\sim\chi$ is in $L(\rho_1)$. If $\sim\chi \in L(\rho_1)$, since $\sim\chi$ will be inherited to $L(\rho_2)$, it means that $L(\rho_2)$ will be blatantly inconsistent, which is contradictory with the fact that $\mathscr{T}$ is open. Therefore, it only can be that $\chi \in L(\rho_1)$. Hence, $L(\rho_2)|x \subseteq L(\rho_1)|x$.

For (2), it can be shown by the same process as for (1).

For (3), by (1) we know that $L(\rho)|x = L(\rho_1)|x$ and $L(\rho)|x = L(\rho_2)|x$. Thus, $L(\rho_1)|x = L(\rho_2)|x$. □

Before we show the truth lemma, we need the following auxiliary proposition.

Proposition 15. *Let $\mathscr{T}$ be a complete subtree of $\mathscr{T}_{\varphi_0}$, σ be an i-strategy in $M^{\mathscr{T}}$, and $\delta = [\rho_1]^i \cdots [\rho_h]^i$ be a σ-execution in $M^{\mathscr{T}}$. If $\neg Kh_i\chi \in L(\rho_h)$ and δ is not complete, then there exists some $\rho \in W$ such that $\delta[\rho]^i$is a σ-execution and $\neg Kh_i\chi \in L(\rho)$.*

Proof. Since $[\rho_1]^i \cdots [\rho_h]^i$ is not complete, it follows that $[\rho_h]^i \in \mathtt{dom}(\sigma)$. Let $\sigma([\rho_h]^i) = a_{Kh_i\psi}$ for some $a_{Kh_i\psi} \in \mathbf{A}_i$. Since σ is a uniformly executable strategy, this means that a_{Kh_i} is executable on ρ_h. By the definition of $M^{\mathscr{T}}$ in Definition 13, we know that there exists a maximal ε-path ρ' of $\mathscr{T}$ such that $\rho_h \xrightarrow{a_{Kh_i\psi}} \rho'$, that is, $\rho_h\rho'$ is a path of $\mathscr{T}$ and $L(end(\rho_h), ini(\rho')) = a_{Kh_i\psi}$. Note that such labels can only be added by the construction step (j) or (k).

From the construction steps (j) and (k), we then have that $Kh_i\psi, \neg K_i\psi \in L(\rho_h)$. Since we also have that $\neg Kh_i\chi \in L(\rho_h)$, by the construction step (k), there is a node $n_{(\neg Kh_i\chi, Kh_i\psi)}$ in $\mathscr{T}_{\varphi_0}$ such that $L(n_{(\neg Kh_i\chi, Kh_i\psi)}) = \{K_i\psi, \neg Kh_i\chi\}$ and $L(end(\rho_h), n_{(\neg Kh_i\chi, Kh_i\psi)}) = a_{Kh_i\psi}$.

Since $\mathscr{T}$ is a complete subtree of $\mathscr{T}_{\varphi_0}$, it follows that $n_{(\neg Kh_i\chi, Kh_i\psi)}$ is also a node in $\mathscr{T}$. There are two cases: $n_{(\neg Kh_i\chi, Kh_i\psi)}$ is blocked or not.

If $n_{(\neg Kh_i\chi, Kh_i\psi)}$ is not blocked, let ρ be the maximal ε-path in $\mathscr{T}$ that begins with the node $n_{(\neg Kh_i\chi, Kh_i\psi)}$. By the definition of $M^{\mathscr{T}}$, we know that $(\rho_h, \rho) \in R_{a_{Kh_i\psi}}$. This means that $[\rho_1]^i \cdots [\rho_h]^i[\rho]^i$ is a σ-execution. Moreover, due to $\neg Kh_i\chi \in L(ini(\rho))$, it follows that $\neg Kh_i\chi \in L(\rho)$.

If $n_{(\neg Kh_i\chi, Kh_i\psi)}$ is blocked by its some node n', then n' is an ancestor of n. So, we have that n' is a node in $\mathscr{T}$. Let ρ'' be the maximal ε-path that contains n'. Since block nodes are leaf nodes, it follows that ρ'' is not blocked. Thus, $\rho'' \in W$. By the definition of $M^{\mathscr{T}}$, we have that $(\rho_h, \rho'') \in R_{a_{Kh_i\psi}}$. This means that $[\rho_1]^i \cdots [\rho_h]^i[\rho'']^i$ is a σ-execution. Moreover, due to $\neg Kh_i\chi \in L(n)$, $L(n) = L(n')$ and $L(n') \subseteq L(\rho)$, it follows that $\neg Kh_i\chi \in L(\rho)$. □

Lemma 16 (Truth lemma)**.** *If $\mathscr{T}$ is a complete and open subtree of $\mathscr{T}_{\varphi_0}$, then for each $\varphi \in sub^+(\varphi_0)$, we have that $M^{\mathscr{T}}, \rho \vDash \varphi$ iff $\varphi \in L(\rho)$.*

Proof. We prove it by induction on φ. Since $L(\rho)$ is closed over sub^+-formulas, the atomic case and Boolean cases are straightforward. Next, we will focus on the cases of $K_i\psi$ and $Kh_i\psi$.

For the case $K_i\psi \in L(\rho)$, given $\rho' \in [\rho]^i$, we will show that $M^{\mathscr{T}}, \rho' \vDash \psi$. By Proposition 14, we have that $K_i\psi \in L(\rho')$. Moreover, since $\mathscr{T}$ is open and ρ' is not blocked, this means that all formulas in $L(\rho')$ are marked as "checked". Let $\rho' = n_1 \cdots n_h$. From the construction, we know that there is some node n_k, where $1 \leq k \leq h$, such that the first time in ρ' the formula $K_i\psi$ is marked as "checked", and this only can be done by the construction step (d). Thus, we have that $\psi \in L(n_k)$, and then $\psi \in L(\rho')$. By IH, we have that $M^{\mathscr{T}}, \rho' \vDash \psi$.

For the case $K_i\psi \notin L(\rho)$, since $L(\rho)$ is closed over sub^+-formulas, it follows that $\neg K_i\psi \in L(\rho)$. Let $end(\rho)$ be the node n. Since ρ is a non-blocked maximal ε-path and $\mathscr{T}$ is open, this means that all formulas in $L(n)$ is marked as "checked". From the construction, we know that the construction step (i) will be triggered on n. Thus, either there is an i-ancestor node n' such that $\neg\psi \in L(n')$ or there is an n's i-child node n'' such that $\neg\psi \in L(n'')$. Since $\mathscr{T}$ is a complete subtree of $\mathscr{T}_{\varphi_0}$, it follows that no matter n'

or n'' will be included in $\mathscr{T}$. Thus, there is $\rho' \in [\rho]^i$ such that $\neg\psi \in L(\rho')$. Since $\mathscr{T}$ is open, it follows that $\psi \notin L(\rho')$. By IH, we have that $M^{\mathscr{T}}, \rho \not\vDash \psi$. Hence, $M^{\mathscr{T}}, \rho \not\vDash K_i\psi$.

For the case $Kh_i\psi \in L(\rho)$, since $L(\rho)$ is closed over sub^+-formulas, it follows that either $K_i\psi \in L(\rho)$ or $\neg K_i\psi \in L(\rho)$.

If $K_i\psi \in L(\rho)$, by the proof of the K_i-case above, we know that $M^{\mathscr{T}}, \rho \vDash K_i\psi$. By Proposition 6, we have that $M^{\mathscr{T}}, \rho \vDash Kh_i\psi$.

If $\neg K_i\psi \in L(\rho)$, we will show that σ is a good strategy for $M^{\mathscr{T}}, \rho \vDash Kh_i\psi$ where σ is the function $\{[\rho]^i \mapsto a_{Kh_i\psi}\}$.

If $\rho' \in [\rho]^i$ and $(\rho', \rho'') \in R_{a_{Kh_i\psi}}$, then either $\rho' \xrightarrow{a_{Kh_i\psi}} \rho''$, or $\rho' \xrightarrow{a_{Kh_i\psi}} \rho'''$ where ρ''' is blocked by ρ''. For the first case, it is obvious that $\rho'' \notin [\rho]^i$. For the second case, we also have that $\rho'' \notin [\rho]^i$. The reason is that nodes can only be blocked in the construction step (k), so for the second case we have $K_i\psi \in L(\rho''')$ and $L(\rho''') \subseteq L(\rho'')$. Due to $\neg K_i\psi \in L(\rho)$ and Proposition 14, Therefore, for the second case we also have that $\rho'' \notin [\rho]^i$. Thus, we have shown that if $\rho' \in [\rho]^i$ and $(\rho', \rho'') \in R_{a_{Kh_i\psi}}$, then $\rho'' \notin [\rho]^i$, which means that $[\rho']^i[\rho'']^i$ is a complete σ-execution from $[\rho]^i$. Hence, to show that $M^{\mathscr{T}}, \rho \vDash Kh_i\psi$, we only need to show that for each $\rho' \in [\rho]^i$, (1) the action $a_{Kh_i\psi}$ is executable at ρ', (which means that σ is a uniform executable strategy,) and (2) $M^{\mathscr{T}}, \rho'' \vDash K_i\psi$ for each ρ'' with $(\rho', \rho'') \in R_{a_{Kh_i\psi}}$.

For (1), due to $Kh_i\psi, \neg K_i\psi \in L(\rho)$ and Proposition 14, we know that $Kh_i\psi, \neg K_i\psi \in L(\rho')$. Thus, by the construction step (j), there is a node $n_{Kh_i\psi}$ such that $L(n_{Kh_i\psi}) = \{K_i\psi\}$ and $L(end(\rho'), n_{Kh_i\psi}) = a_{Kh_i\psi}$. Since $\mathscr{T}$ is a complete subtree of $\mathscr{T}_{\varphi_0}$, it follows that $n_{Kh_i\psi}$ is a node in $\mathscr{T}$. Let ρ''' be the maximal ε-path that begins with $n_{Kh_i\psi}$. We have that $\rho' \xrightarrow{a_{Kh_i\psi}} \rho'''$, and then $(\rho', \rho''') \in R_{a_{Kh_i\psi}}$. Thus, the action $a_{Kh_i\psi}$ is executable at ρ'.

For (2), if $(\rho', \rho'') \in R_{a_{Kh_i\psi}}$, it means that either $L(end(\rho'), ini(\rho'')) = a_{Kh_i\psi}$ or $L(end(\rho'), n') = a_{Kh_i\psi}$ where n' is blocked by ρ''. From the construction, we know that only the steps (j) and (k) can add such labels. From these steps, we know that $K_i\psi \in L(ini(\rho''))$ or $K_i\psi \in L(n')$. In either case, we have that $K_i\psi \in L(\rho'')$. By the proof of the K_i-case above, we know that $M^{\mathscr{T}}, \rho'' \vDash K_i\psi$.

For the case $Kh_i\psi \notin L(\rho)$, assume that $M^{\mathscr{T}}, \rho \vDash Kh_i\psi$. By the semantics, there is an i-strategy σ such that all compete σ-executions from $[\rho]^i$ are finite and $M^{\mathscr{T}}, \rho' \vDash K_i\psi$ for all $[\rho']^i \in \mathtt{ECE}(\sigma, [\rho]^i)$. Due to $Kh_i\psi \notin L(\rho)$, we have that $\neg Kh_i\psi \in L(\rho)$. By Proposition 15, we know that there is a complete σ-execution $[\rho]^i \cdots [\rho']^i$ such that $\neg Kh_i\psi \in L(\rho')$. By the construction step (g), we know that $\neg K_i\psi \in L(\rho')$. Due to $Kh_i\psi \in sub^+(\varphi_0)$, it follows that $K_i\psi \in sub^+(\varphi_0)$. By proof of the K_i-case above, we have that $M^{\mathscr{T}}, \rho' \vDash \neg K_i\psi$. Contradiction! Thus, $M^{\mathscr{T}}, \rho \not\vDash Kh_i\psi$. □

The completeness below follows from Lemma 16 above.

Theorem 17 (Completeness)**.** *If $\mathscr{T}_{\varphi_0}$ is not closed, then φ_0 is satisfiable.*

4.2 Complexity

In this section, we will show that the procedure of Definition 8 runs in polynomial space.

Definition 18 (Depth)**.** *The* depth *of a formula, denoted by $dep(\varphi)$, is the depth of nesting of K or Kh operators, which is defined in Table 2.*

The reason that $dep(Kh_i\psi) > dep(K_i\psi)$ is that in the construction steps (j) and (k) we need to label n's child node with the formula $K_i\psi$ for $Kh_i\psi \in L(n)$.

Proposition 19. *The height of the tableau tree $\mathscr{T}_{\varphi_0}$ is bounded by a polynomial function of the size of the set $sub^+(\varphi_0)$.*

$$
\begin{array}{rcl}
dep(\bot) & = & 0 \\
dep(p) & = & 0 \\
dep(\neg\psi) & = & dep(\psi) \\
dep(\psi\wedge\chi) & = & max\{dep(\psi), dep(\chi)\} \\
dep(K_i\psi) & = & dep(\psi)+1 \\
dep(Kh_i\psi) & = & dep(\psi)+2
\end{array}
$$

Table 2: Depth of formulas

Proof. Let the size of $sub^+(\varphi_0)$ be m. In the construction of the tableau tree $\mathcal{T}_{\varphi_0}$ in Definition 8, each step of (a)-(k) might add the height of the tree with 1 degree. The steps (a)-(g) can be consecutively executed at most m times to get a sub^+-formula closure.

Now consider a $\mathcal{T}_{\varphi_0}$ path that starts from the root node, and we contract it by seeing the consecutive nodes whose edges are labeled ε as one single node. Let n be a node in this contracted path b.

If the node n's child node is added by the step (j), then the greatest depth of formulas labeled on the child node is strictly less than the greatest depth of formulas labeled on n.

If the node n's child node is added by the step (i), although the greatest depth of formulas labeled on the child node might be the same as the greatest depth of formulas labeled on n, but, such descendant nodes with the same greatest depth with n can be consecutively added by the step (i) at most m times to achieve a node whose K_i ancestor has the same labeled formulas. After that, if a descendant node n' is added by the step (i) again, then it must be a K_j descendant where $i \neq j$. Thus, the greatest depth will be strictly shrunk. If the descendant node n' is added by the step (k), the greatest depth might still keep the same with n. However, the step (k) can be executed on one path at most m^2 times before it adds a blocked node.

Hence, there are at most m^4 consecutive nodes in b that have the same greatest depth before the greatest depth becomes strictly less. Therefore, the length of the contracted path b is at most m^5, and the length of the original path is at most m^6. It follows that the height of $\mathcal{T}_{\varphi_0}$ is bounded by m^6. □

Lemma 20. *There is an algorithm that runs in polynomial space for deciding whether $\mathcal{T}_{\varphi_0}$ is closed.*

Proof. Let the size of $sub^+(\varphi_0)$ be m. In the construction step (i), at most m successors are added. In the step (j), at most m^2 successors are added, and in the step (k) at most m^3 successors are added. This means that each node has at most $m+m^2+m^3$ child nodes. Therefore, the tableau tree $\mathcal{T}_{\varphi_0}$ is an $m+m^2+m^3$-ary tree. By Proposition 19, we know that the height of the tree is bounded by m^6.

We can mark the node to check whether the tree is closed. Note that how a node is marked can be completely determined by its label and how its successors are marked. Once we have determined how a node is marked, we never have to consider the subtree below that node again. Thus, a depth-first search of the tree that runs in polynomial space can decide whether the tree is closed [7]. □

Since the satisfiability problem of epistemic logic with no less than 2 agents is PSPACE-hard (see [7]), and it is a fragment of ELKh_n, it follows that the satisfiability problem of ELKh_n is also PSPACE-hard. Together with Lemma 20, we have the following result.

Theorem 21. *The problem of the satisfiability of ELKh_n formulas is PSPACE-complete.*

5 Conclusion

This paper presented a tableau procedure for the multi-agent version of the logic of strategically *knowing how*. The tableau method presented in this paper is developed from the tableau method for epistemic logic in [7] and the tableau method for *knowing how* logic via simple plans [14]. This paper showed the soundness, the completeness, and the complexity of this tableau procedure. Since the procedure runs in polynomial space, it follows that the satisfiability problem of the logic of strategically *knowing how* can be decided in PSPACE. Moreover, since the *knowing how* logic based on PDL-style knowledge-based plans over finite models in [15] is the same as the logic of strategically *knowing how*, it means that the satisfiability problem of that logic also can be decided in PSPACE.

Acknowledgement The author is grateful to the anonymous reviewers of TARK-2023. Their detailed comments helped the author to improve the presentation of the paper. The author also thanks the support from the National Social Science Foundation of China (Grant No. 18CZX062).

References

[1] Carlos Areces, Raul Fervari, Andrés R. Saravia & Fernando R. Velázquez-Quesada (2021): *Uncertainty-Based Semantics for Multi-Agent Knowing How Logics*. Electronic Proceedings in Theoretical Computer Science 335, Open Publishing Association, pp. 23–37, doi:10.4204/EPTCS.335.3.

[2] Alexandru Baltag (2016): *To Know is to Know the Value of a Variable*. In Lev D. Beklemishev, Stéphane Demri & András Maté, editors: Advances in Modal Logic, 11, College Publications, pp. 135–155. Available at http://www.aiml.net/volumes/volume11/Baltag.pdf.

[3] Sophia Epstein & Pavel Naumov (2021): *Epistemic Logic of Know-Who*. Proceedings of the AAAI Conference on Artificial Intelligence 35(13), pp. 11479–11486, doi:10.1609/aaai.v35i13.17367.

[4] Jie Fan, Yanjing Wang & Hans van Ditmarsch (2015): *Contingency and Knowing Whether*. Rev. Symb. Log. 8(1), pp. 75–107, doi:10.1017/S1755020314000343.

[5] Raul Fervari, Andreas Herzig, Yanjun Li & Yanjing Wang (2017): *Strategically knowing how*. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pp. 1031–1038, doi:10.24963/ijcai.2017/143.

[6] Tao Gu & Yanjing Wang (2016): *"Knowing value" logic as a normal modal logic*. In Lev D. Beklemishev, Stéphane Demri & András Maté, editors: Advances in Modal Logic, 11, College Publications, pp. 362–381. Available at http://www.aiml.net/volumes/volume11/Gu-Wang.pdf.

[7] Joseph Y Halpern & Yoram Moses (1992): *A guide to completeness and complexity for modal logics of knowledge and belief*. Artificial Intelligence. 54(2), pp. 319–379, doi:10.1016/0004-3702(92)90049-4.

[8] Sergiu Hart, Aviad Heifetz & Dov Samet: *"Knowing Whether," "Knowing That," and The Cardinality of State Spaces* 70(1), pp. 249–256. doi:10.1006/jeth.1996.0084.

[9] Andreas Herzig (2015): *Logics of knowledge and action: critical analysis and challenges*. Autonomous Agents and Multi-Agent Systems 29(5), pp. 719–753, doi:10.1007/s10458-014-9267-z.

[10] Andreas Herzig & Nicolas Troquard (2006): *Knowing How to Play: Uniform Choices in Logics of Agency*. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '06, Association for Computing Machinery, New York, NY, USA, p. 209–216, doi:10.1145/1160633.1160666.

[11] J Hintikka (1962): *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca N.Y.

[12] Wojciech Jamroga & Thomas Ågotnes (2007): *Constructive knowledge: what agents can achieve under imperfect information*. *Journal of Applied Non-Classical Logics* 17(4), pp. 423–475, doi:10.3166/jancl.17.423-475.

[13] Yves Lespérance, Hector J. Levesque, Fangzhen Lin & Richard B. Scherl (2000): *Ability and Knowing How in the Situation Calculus*. *Stud Logica* 66(1), pp. 165–186, doi:10.1023/A:1026761331498.

[14] Yanjun Li (2021): *Tableau-Based Decision Procedure for Logic of Knowing-How via Simple Plans*. *Lecture Notes in Computer Science* 13040, Springer, pp. 266–283, doi:10.1007/978-3-030-89391-0_15.

[15] Yanjun Li & Yanjing Wang (2021): *Planning-based knowing how: A unified approach*. *Artificial Intelligence* 296, p. 103487, doi:10.1016/j.artint.2021.103487.

[16] J McCarthy (1979): *First Order Theories of Individual Concepts and Propositions*. *Machine Intelligence* 9., pp. 129–147.

[17] Robert C Moore (1984): *A formal theory of knowledge and action*. Technical Report, DTIC Document.

[18] Pavel Naumov & Jia Tao (2017): *Together We Know How to Achieve: An Epistemic Logic of Know-How (Extended Abstract)*. *Electronic Proceedings in Theoretical Computer Science* 251, pp. 441–453, doi:10.4204/EPTCS.251.32.

[19] Pavel Naumov & Jia Tao (2018): *Second-Order Know-How Strategies*. In: *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '18)*. Available at http://dl.acm.org/citation.cfm?id=3237444.

[20] Pavel Naumov & Jia Tao (2018): *Strategic Coalitions with Perfect Recall*. *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*, doi:10.1609/aaai.v32i1.11579.

[21] Pavel Naumov & Jia Tao (2018): *Together we know how to achieve: An epistemic logic of know-how*. *Artificial Intelligence* 262, pp. 279–300, doi:10.1016/j.artint.2018.06.007.

[22] Pavel Naumov & Jia Tao (2019): *Knowing-how under uncertainty*. *Artificial Intelligence* 276, pp. 41–56, doi:10.1016/j.artint.2019.06.007.

[23] Carlotta Pavese (2022): *Knowledge How*. In Edward N. Zalta & Uri Nodelman, editors: *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition, Metaphysics Research Lab, Stanford University. Available at https://plato.stanford.edu/archives/fall2022/entries/knowledge-how/.

[24] G H Von Wright (1951): *An Essay in Modal Logic*. North Holland, Amsterdam.

[25] Yanjing Wang (2015): *A Logic of Knowing How*. In: *Proceedings of LORI 2015*, pp. 392–405, doi:10.1007/978-3-662-48561-3_32.

[26] Yanjing Wang (2018): *A logic of goal-directed knowing how*. *Synthese* 195, p. 4419–4439, doi:10.1007/s11229-016-1272-0.

[27] Yanjing Wang (2018): *Beyond Knowing That: A New Generation of Epistemic Logics*. In Hans van Ditmarsch & Gabriel Sandu, editors: *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, Springer International Publishing, Cham, pp. 499–533, doi:10.1007/978-3-319-62864-6_21.

[28] Chao Xu, Yanjing Wang & Thomas Studer (2021): *A logic of knowing why*. *Synthese* 198(2), pp. 1259–1285, doi:10.1007/s11229-019-02104-0.

Epistemic Syllogistic: First Steps

Yipu Li
Department of Philosophy, Peking University, CHINA
luxemburg@stu.pku.edu.cn

Yanjing Wang
Department of Philosophy, Peking University, CHINA
y.wang@pku.edu.cn

Aristotle's discussions on modal syllogistic have often been viewed as error-prone and have garnered significant attention in the literature due to historical and philosophical interests. However, from a contemporary standpoint, they also introduced natural fragments of first-order modal logic, warranting a comprehensive technical analysis. In this paper, drawing inspiration from the natural logic program, we propose and examine several variants of modal syllogistic within the epistemic context, thereby coining the term *Epistemic Syllogistic*. Specifically, we concentrate on the *de re* interpretation of epistemic syllogisms containing non-trivial yet natural expressions such as "all things known to be A are also known to be not B." We explore the epistemic apodeictic syllogistic and its extensions, which accommodate more complex terms. Our main contributions include several axiomatizations of these logics, with completeness proofs that may be of independent interest.

1 Introduction

Although modal logic is regarded as a relatively young field, its origins can be traced back to Aristotle, who explored syllogistic reasoning patterns that incorporated modalities. However, in contrast to his utterly successful assertoric syllogistic, Aristotle's examination of modal syllogisms is often viewed as error-prone and controversial, thus receiving less attention from logicians. In the literature, a large body of research on Aristotle's modal syllogistic primarily centers on the possibility of a coherent interpretation of his proposed modal systems grounded by his philosophy on necessity and contingency (see, e.g., [11, 5, 12]).

We adopt a more liberal view on Aristotle's modal syllogistic, considering it as a source of inspiration for formalizing natural reasoning patterns involving modalities, rather than scrutinizing the coherence of the original systems. Our approach is encouraged by the fruitful research program of *natural logic*, which explores "light" logic systems that admit intuitive reasoning patterns in natural languages while balancing expressivity and computational complexity [1, 8]. In particular, various extensions of the assertoric syllogistic have been proposed and studied [8].

In this paper, we propose a systematic study on *epistemic syllogistic* to initiate our technical investigations of (extensions of) modal syllogistic. The choice for the epistemic modality is intentional for its ubiquitous use in natural languages. Consider the following syllogism:

$$\frac{\text{All } C \text{ are } B \qquad \text{Some } C \text{ is } \textit{known} \text{ to be } A}{\text{Some } B \text{ is } \textit{known} \text{ to be } A}$$

Taking the intuitive *de re* reading, the second premise and the conclusion above can be formalized as $\exists x(Cx \wedge \mathsf{K}Ax)$ and $\exists x(Bx \wedge \mathsf{K}Ax)$ respectively in first-order modal logic (FOML).[1] It then becomes apparent that this syllogism is valid under the standard semantics of FOML. One objective of our investigation into epistemic syllogistic is to explore various natural fragments of FOML following the general structure of syllogisms.

[1] The *de dicto* reading of the second premise would be $\mathsf{K}(\exists x(Cx \wedge Ax))$, which we do not discuss here.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 392–406, doi:10.4204/EPTCS.379.31

Aristotle's original apodeictic syllogistic only allows a single occurrence of a necessity modality at a particular position in each sentence of assertoric syllogistic. However, from a modern perspective, we can greatly extend it and express interesting epistemic statements involving multiple agents and nested knowledge, such as "Everything known to be *A* by *i* is also known to be *A* by *j*". Moreover, it is also interesting to allow nested knowledge such as "Something *i* knows that *j* knows to be *A* is also known to be *B* by *i*". The general idea is to extend the language of terms but keep the pattern of "Some t is g" and "All t are g", as proposed in [10].

In this paper, we begin by presenting preliminaries about assertoric syllogisms in Section 2. We then proceed to examine the epistemic version of Aristotle's apodeictic syllogistic in Section 3 and provide a complete axiomatization. In Section 4, we significantly expand the language of terms in a compositional manner to allow for nesting of modalities with respect to multiple agents. The completeness of the proposed proof systems is demonstrated in Section 5. We conclude with a discussion of future work in the final section.

2 Preliminaries

In this section, we familiarize the readers with the basics of Aristotle's syllogistic. Let us first consider the language of *Assertoric Syllogistic*.

Definition 1 (Language L_{AS}) *Given a countable set of predicates U, the language of Assertoric Syllogistic is defined by the following grammar:*

$$\varphi ::= \mathsf{All}(\mathsf{t},\mathsf{g}) \mid \mathsf{Some}(\mathsf{t},\mathsf{g}), \quad \mathsf{t} ::= A, \quad \mathsf{g} ::= A \mid \neg A$$

where $A \in U$. For the ease of presentation, we also write $\neg\mathsf{All}(\mathsf{A},\mathsf{B}) := \mathsf{Some}(\mathsf{A},\neg\mathsf{B})$, $\neg\mathsf{All}(\mathsf{A},\neg\mathsf{B}) := \mathsf{Some}(\mathsf{A},\mathsf{B})$, $\neg\mathsf{Some}(\mathsf{A},\mathsf{B}) := \mathsf{All}(\mathsf{A},\neg\mathsf{B})$ *and* $\neg\mathsf{Some}(\mathsf{A},\neg\mathsf{B}) := \mathsf{All}(\mathsf{A},\mathsf{B})$.

The semantics for L_{AS} is based on first-order structures.

Definition 2 (Semantics for L_{AS}) *A model of L_{AS} is a pair $\mathscr{M} = (D,\rho)$ where D is a non-empty domain and $\rho : U \to \mathscr{P}(D)$ is an interpretation function. The satisfaction relation is defined as below where the third column shows the equivalent clauses in the first-order language.*

$\mathscr{M} \models_{AS} \mathsf{All}(\mathsf{A},\mathsf{B})$	$\iff$	$\rho(A) \subseteq \rho(B)$	$\mathscr{M} \Vdash \forall x(Ax \to Bx)$
$\mathscr{M} \models_{AS} \mathsf{All}(\mathsf{A},\neg\mathsf{B})$	$\iff$	$\rho(A) \cap \rho(B) = \emptyset$	$\mathscr{M} \Vdash \forall x(Ax \to \neg Bx)$
$\mathscr{M} \models_{AS} \mathsf{Some}(\mathsf{A},\mathsf{B})$	$\iff$	$\rho(A) \cap \rho(B) \neq \emptyset$	$\mathscr{M} \Vdash \exists x(Ax \wedge Bx)$
$\mathscr{M} \models_{AS} \mathsf{Some}(\mathsf{A},\neg\mathsf{B})$	$\iff$	$\rho(A) \not\subseteq \rho(B)$	$\mathscr{M} \Vdash \exists x(Ax \wedge \neg Bx)$

Note that since we wish to generalize the ideas of the syllogistics from the modern perspective, the interpretation of a predicate can be an empty set, in contrast with the Aristotelian non-emptiness assumption.

Following the study of Corcoran [3] and Martin [6], we present the following deduction system $\mathbb{S}_{AS}$. Note that our system is slightly different from that of Corcoran's and Martin's, as they are loyal to Aristotle's non-emptiness assumption.[2]

[2] Cf. [7] for a direct proof system that replaces RAA rule by the explosion rule. Moss' work is targeted at a stronger language, which allows complement terms in the antecedent. e.g. $\mathsf{All}(\neg\mathsf{A},\neg\mathsf{B})$.

$$\mathsf{All}(\mathsf{A},\mathsf{A}) \qquad \frac{\mathsf{Some}(\mathsf{A},\mathsf{B})}{\mathsf{Some}(\mathsf{B},\mathsf{A})}\,\text{Conversion} \qquad \frac{\dfrac{[\neg\varphi]}{\psi} \quad \dfrac{[\neg\varphi]}{\neg\psi}}{\varphi}\,\text{RAA}$$

$$\frac{\mathsf{All}(\mathsf{A},\mathsf{B}) \qquad \mathsf{All}(\mathsf{B},\mathsf{g})}{\mathsf{All}(\mathsf{A},\mathsf{g})}\,\text{Barbara-Celarent} \qquad \frac{\mathsf{Some}(\mathsf{A},\mathsf{g})}{\mathsf{Some}(\mathsf{A},\mathsf{A})}\,\text{Existence}$$

With a slight modification of Corcoran's result in Section 4 of [3], it follows that the above system is sound and complete.

Theorem 3 $\mathbb{S}_{AS}$ *is sound and strongly complete w.r.t. the semantics.*

3 Epistemic Apodeictic Syllogistic

Inspired by apodeictic syllogistic, we introduce the first language of *Epistemic Syllogistic*.

Definition 4 (Language L_{EAS}) *Given a countable set of predicates U, the language of Epistemic Apodeictic Syllogistic is generated by the following grammar of formulas (φ) and terms (t,g):*

$$\varphi ::= \mathsf{All}(\mathsf{t},\mathsf{g}) \mid \mathsf{Some}(\mathsf{t},\mathsf{g}), \quad \mathsf{t} ::= A, \quad \mathsf{g} ::= A \mid \neg A \mid \mathsf{K}A \mid \mathsf{K}\neg A$$

where $A \in U$. We collect all the g as the set of (categorical) terms $Term^{EAS}(U)$.

Note that the formulas should be read *de re*. For example, $\mathsf{All}(\mathsf{A},\mathsf{K}\neg\mathsf{B})$ says "all A are known to be not B", expressing the logical form $\forall x(Ax \to \mathsf{K}\neg Bx)$. Formulas without modalities are called non-modal formulas.

L_{EAS} is interpreted on first-order Kripke models with a constant domain.

Definition 5 (Models for L_{EAS}) *A model for L_{EAS} a tuple $\mathcal{M} = (W,R,D,\rho)$. W is the set of possible worlds, $R \subseteq W \times W$ is a reflexive relation, D is the non-empty domain, and $\rho : W \times U \to \mathscr{P}(D)$ is the interpretation function. We also write $\rho_w(A)$ for $\rho(w,A)$.*

Note that further frame conditions such as transitivity and Euclidean property do not play a role here since the syntax does not allow nested modalities, which will be relaxed in the next section.

To ease the presentation of the semantics, we extend the interpretation ρ to any term.

Definition 6 $\rho^+ : W \times Term^{ES}(U) \to \mathscr{P}(D)$ *is defined as:*

$$\rho^+_w(A) = \rho_w(A), \quad \rho^+_w(\neg A) = D - \rho_w(A) \quad \rho^+_w(\mathsf{K}A) = \bigcap_{wRv} \rho_v(A) \quad \rho^+_w(\mathsf{K}\neg A) = \bigcap_{wRv} (D - \rho_v(A))$$

Definition 7 (Semantics for L_{EAS}) *Given a pointed model $\mathcal{M},w$, the satisfaction relation is defined as follows where the third column lists the corresponding first-order modal formulas.*

$$\begin{array}{|llll|}\hline \mathcal{M},w \models_{ES} \mathsf{All}(\mathsf{A},\mathsf{g}) & \iff & \rho_w(A) \subseteq \rho^+_w(g) & \mathcal{M},w \Vdash \forall x(Ax \to g(x)) \\ \mathcal{M},w \models_{ES} \mathsf{Some}(\mathsf{A},\mathsf{g}) & \iff & \rho_w(A) \cap \rho^+_w(g) \neq \emptyset & \mathcal{M},w \Vdash \exists x(Ax \wedge g(x)) \\ \hline \end{array}$$

where we abuse the notation and let $g(x)$ be a modal predicate formula defined as follows:

$$\begin{array}{llll} g(x) = Ax & \textit{if } g = A, & g(x) = \neg Ax & \textit{if } g = \neg A \\ g(x) = \mathsf{K}Ax & \textit{if } g = \mathsf{K}A, & g(x) = \neg\mathsf{K}Ax & \textit{if } g = \neg\mathsf{K}A \end{array}$$

where K is the modal operator and Ax is an atomic formula.

We propose the following proof system $\mathbb{S}_{EAS}$:

$$\mathsf{All}(\mathsf{A},\mathsf{A}) \qquad\qquad \dfrac{\dfrac{[\neg\varphi]}{\psi} \quad \dfrac{[\neg\varphi]}{\neg\psi}}{\varphi}\ \text{RAA (given non-modal } \varphi,\psi)$$

$$\dfrac{\mathsf{Some}(\mathsf{A},\mathsf{Kg})}{\mathsf{Some}(\mathsf{A},\mathsf{g})}\ \text{E-Truth} \qquad\qquad \dfrac{\mathsf{All}(\mathsf{A},\mathsf{Kg})}{\mathsf{All}(\mathsf{A},\mathsf{g})}\ \text{A-Truth}$$

$$\dfrac{\mathsf{Some}(\mathsf{A},\mathsf{B})}{\mathsf{Some}(\mathsf{B},\mathsf{A})}\ \text{Conversion} \qquad\qquad \dfrac{\mathsf{All}(\mathsf{A},\mathsf{B}) \quad \mathsf{All}(\mathsf{B},\mathsf{g})}{\mathsf{All}(\mathsf{A},\mathsf{g})}\ \text{Barbara/Celarent}$$

$$\dfrac{\mathsf{Some}(\mathsf{A},\mathsf{B}) \quad \mathsf{All}(\mathsf{B},\mathsf{g})}{\mathsf{Some}(\mathsf{A},\mathsf{g})}\ \text{Darii/Ferio} \qquad\qquad \dfrac{\mathsf{All}(\mathsf{C},\mathsf{B}) \quad \mathsf{Some}(\mathsf{C},\mathsf{Kg})}{\mathsf{Some}(\mathsf{B},\mathsf{Kg})}\ \text{Disamis/Bocardo}$$

$$\dfrac{\mathsf{Some}(\mathsf{A},\mathsf{g})}{\mathsf{Some}(\mathsf{A},\mathsf{A})}\ \text{Existence 1} \qquad\qquad \dfrac{\mathsf{Some}(\mathsf{B},\mathsf{KA})}{\mathsf{Some}(\mathsf{A},\mathsf{KA})}\ \text{Existence 2}$$

We say a set of formulas is *consistent* if it cannot derive a contradiction in system $\mathbb{S}_{EAS}$. Note that the RAA rule is restricted to non-modal formulas, as formulas with K in L_{EAS} do not have negations expressible in the language.

Theorem 8 (Completeness) *If $\Sigma \models_{ES} \varphi$, then $\Sigma \vdash_{\mathbb{S}_{EAS}} \varphi$.*

Due to the lack of space, we only sketch the idea of the (long) proof in Appendix A.

4 Multi-agent Syllogistic with Nested Knowledge

The language L_{EAS} has an asymmetry in the grammar such that the first term is simpler than the second. In this section, we restore the symmetry of the two terms. Moreover, the terms are now fully compositional using modalities and negations, thus essentially allowing nested modalities in both $\Box$ and $\Diamond$ shapes, also in a multi-agent setting. It can be viewed as a modal extension of the language of Syllogistic Logic with Complement in [7], or a fragment of the language of Aristotelian Modal Logic in [10].

Definition 9 (Language L_{NES}) *Given a countable set of predicates U and a set of agents I, the language L_{NES} is defined by the following grammar:*

$$\varphi ::= \mathsf{All}(\mathsf{g},\mathsf{g}) \mid \mathsf{Some}(\mathsf{g},\mathsf{g}), \quad \mathsf{g} ::= A \mid \mathsf{K}_i g \mid \neg g$$

Where $A \in U$ and $i \in I$. The set of terms g is denoted as $Term^{NES}(U)$.

As before, we define $\neg\mathsf{All}(\mathsf{g}_1,\mathsf{g}_2) := \mathsf{Some}(\mathsf{g}_1,\neg\mathsf{g}_2)$ and $\neg\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2) := \mathsf{All}(\mathsf{g}_1,\neg\mathsf{g}_2)$. Moreover, let $\widehat{\mathsf{K}}_i g$ be an abbreviation for $\neg\mathsf{K}_i\neg g$. With this powerful language L_{NES}, we can express the following: "Everything i knows to be A, j also knows" by $\mathsf{All}(\mathsf{K_iA},\mathsf{K_jA})$; "According to i, something known to be B is possible to be also A" by $\mathsf{Some}(\mathsf{K_iB},\widehat{\mathsf{K}}_\mathsf{i}\mathsf{A})$; "Everything i knows that j knows to be A is also known to be B by i" by $\mathsf{All}(\mathsf{K_iK_jA},\mathsf{K_iB})$.

L_{NES} is also interpreted on first-order Kripke models with a constant domain and multiple relations $(W,\{R_i\}_{i\in I},D,\rho)$. We say the model is a T/S4/S5 model if each R_i is a reflexive/reflexive and transitive /equivalence relation, respectively. Now we define ρ^+, the interpretation function for terms.

Definition 10 $\rho^+ : W \times Term^{NES}(U) \to \mathscr{P}(D)$ *is defined recursively as:*

$$\rho^+_w(A) = \rho_w(A) \qquad \rho^+_w(\neg g) = D - \rho^+_w(g) \qquad \rho^+_w(\mathsf{K}_i g) = \bigcap_{wR_iv} \rho^+_v(g)$$

It is easy to see that $\rho^+_w(\widehat{\mathsf{K}}_i g) = \rho^+_w(\neg\mathsf{K}_i\neg g) = \bigcup_{wR_iv}\rho^+_v(g)$.

Definition 11 (Semantics for L_{NES}) *The third column is the corresponding FOML formulas.*

$\mathcal{M},w \models_{NES} \mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)$	$\iff$	$\rho^+_w(g_1) \subseteq \rho^+_w(g_2)$	$\mathcal{M},w \Vdash \forall x(g_1(x) \to g_2(x))$
$\mathcal{M},w \models_{NES} \mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2)$	$\iff$	$\rho^+_w(g_1) \cap \rho^+_w(g_2) \neq \emptyset$	$\mathcal{M},w \Vdash \exists x(g_1(x) \land g_2(x))$

A simple induction would show the FOML formulas above are indeed equivalent to our L_{NES} formulas. For $x \in \{T, S4, S5\}$, we write $\Sigma \models_{x-NES} \varphi$ if for all x-model such that $\mathcal{M},w \models_{NES} \Sigma$, $\mathcal{M},w \models_{NES} \varphi$.

Here is an observation playing an important role in later proofs.

Proposition 12 *For any $g \in Term^{NES}(U)$,* $\mathsf{All}(\mathsf{g},\neg\mathsf{g})$ *and* $\mathsf{Some}(\mathsf{g},\mathsf{g})$ *are both invalid over S5 models (thus also invalid over T, S4 models).*

PROOF [Sketch] First note that $\mathsf{Some}(\mathsf{g},\mathsf{g})$ is equivalent to $\neg\mathsf{All}(\mathsf{g},\neg\mathsf{g})$. We just need to show $\mathsf{All}(\mathsf{g},\neg\mathsf{g})$ and its negation are both satisfiable for all g. Note that a model with a singleton domain $\{a\}$ can be viewed as a Kripke model for propositional modal logic, where a predicate A can be viewed as a propositional letter: it holds on a world w iff $a \in \rho_w(A)$. Then a term g can be viewed as an equivalent modal formula. Since there is only one a in the domain, $\mathsf{All}(\mathsf{g},\neg\mathsf{g})$ is equivalent to $\neg g$, viewed as a modal formula, by the semantics. We just need to show each $\neg g$ and g has singleton S5 models. It is easy to see that each g and $\neg g$ (as modal formula) can be rewritten into an equivalent negative normal form (NNF) using K_i and $\widehat{\mathsf{K}}_i$ to push the negation to the innermost propositional letter, e.g., $\neg\mathsf{K}_i\neg\mathsf{K}_j\neg\mathsf{K}_i A$ can be rewritten as $\widehat{\mathsf{K}}_i\mathsf{K}_j\widehat{\mathsf{K}}_i\neg A$. Now it is easy to satisfy such formulas by a Kripke model with a single world w and the reflexive relations for all R_i: make A true on w iff the NNF of g or $\neg g$ ends up with the literal A instead of $\neg A$. Then we can turn this model into a first-order Kripke model by setting $\rho_w(A) = \{a\}$ iff A is true on w.

■

We propose the following proof system $\mathbb{T}_{NES}$:

$$\mathsf{All}(\mathsf{g},\mathsf{g}),\qquad \mathsf{All}(\mathsf{K_i g},\mathsf{g}),\qquad \mathsf{All}(\mathsf{g},\neg\neg\mathsf{g}),\qquad \mathsf{All}(\neg\neg\mathsf{g},\mathsf{g})$$

$$\frac{\mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)\qquad \mathsf{All}(\mathsf{g}_2,\mathsf{g}_3)}{\mathsf{All}(\mathsf{g}_1,\mathsf{g}_3)}\text{ Barbara}\qquad \frac{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2)}{\mathsf{Some}(\mathsf{g}_2,\mathsf{g}_1)}\text{ Conversion}\qquad \frac{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2)}{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_1)}\text{ Existence}$$

$$\frac{\mathsf{All}(\mathsf{g},\neg\mathsf{g})}{\mathsf{Some}(\neg\mathsf{g},\neg\mathsf{g})}\text{ Non-emptiness}\qquad \frac{\dfrac{[\varphi]}{\psi}\qquad \dfrac{[\varphi]}{\neg\psi}}{\neg\varphi}\text{ RAA}\qquad \frac{\vdash \mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)}{\vdash \mathsf{All}(\mathsf{K_i g}_1,\mathsf{K_i g}_2)}\text{ K}$$

Clearly, $\mathsf{All}(\mathsf{K_i g},\mathsf{g})$ is the counterpart of the usual T axiom in modal logic. The premise of Non-emptiness makes sure that nothing is g, since the FOML model has the nonempty domain, it follows that there is some $\neg g$. Note that the K-rule is restricted to provable formulas, as in the case of the monotonicity rule in modal logic. We define $\mathbb{S}4_{NES}$ to be $\mathbb{T}_{NES} + \mathsf{All}(\mathsf{K_i g},\mathsf{K_i K_i g})$, and $\mathbb{S}5_{NES}$ to be $\mathbb{S}4_{NES} + \mathsf{All}(\neg\mathsf{K_i g},\mathsf{K_i}\neg\mathsf{K_i g})$. It is straightforward to establish soundness if we read the formulas as their first-order modal counterparts.

Theorem 13 (Soundness) $\Sigma \vdash_{\mathbb{T}_{NES}} \varphi$ *implies* $\Sigma \models_{TNES} \varphi$. $\Sigma \vdash_{\mathbb{S}4_{NES}} \varphi$ *implies* $\Sigma \models_{S4NES} \varphi$. $\Sigma \vdash_{\mathbb{S}5_{NES}} \varphi$ *implies* $\Sigma \models_{S5NES} \varphi$.

Below are some derived rules and theorems that will play a role in the later proofs.

Proposition 14 *The following are derivable in $\mathbb{T}_{NES}$ (and thus in $\mathbb{S}4_{NES}, \mathbb{S}5_{NES}$).*

$$\dfrac{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2) \qquad \mathsf{All}(\mathsf{g}_2,\mathsf{g}_3)}{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_3)}\,\text{Darii} \qquad \dfrac{\mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)}{\mathsf{All}(\neg\mathsf{g}_2,\neg\mathsf{g}_1)}\,\text{Contrapositive} \qquad \dfrac{\mathsf{All}(\mathsf{g},\neg\mathsf{g})}{\mathsf{All}(\mathsf{t},\neg\mathsf{g})}\,\text{NonExistence}$$

$$\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g},\widehat{\mathsf{K}}_\mathsf{i}\mathsf{g}) \qquad \vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K}_\mathsf{i}\mathsf{g},\widehat{\mathsf{K}}_\mathsf{i}\mathsf{g}) \qquad \vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K}_\mathsf{i}\neg\neg\mathsf{g},\mathsf{K}_\mathsf{i}\mathsf{g}) \qquad \vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K}_\mathsf{i}\mathsf{g},\mathsf{K}_\mathsf{i}\neg\neg\mathsf{g})$$

PROOF
Darii

$$\dfrac{\dfrac{\dfrac{\mathsf{All}(\mathsf{g}_2,\mathsf{g}_3) \qquad [\mathsf{All}(\mathsf{g}_3,\neg\mathsf{g}_1)]}{\mathsf{All}(\mathsf{g}_2,\neg\mathsf{g}_1)}\,\text{Barbara} \qquad \dfrac{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2)}{\mathsf{Some}(\mathsf{g}_2,\mathsf{g}_1)}\,\text{Conversion}}{\mathsf{Some}(\mathsf{g}_3,\mathsf{g}_1)}\,\text{RAA}}{\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_3)}\,\text{Conversion}$$

Contrapositive

$$\dfrac{\dfrac{[\mathsf{Some}(\neg\mathsf{g}_2,\mathsf{g}_1)] \qquad \mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)}{\mathsf{Some}(\neg\mathsf{g}_2,\mathsf{g}_2)}\,\text{Darii} \qquad \dfrac{}{\mathsf{All}(\neg\mathsf{g}_2,\neg\mathsf{g}_2)}}{\mathsf{All}(\neg\mathsf{g}_2,\neg\mathsf{g}_1)}\,\text{RAA}$$

Non-Existence

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[\mathsf{Some}(\mathsf{t},\neg\neg\mathsf{g})] \qquad \dfrac{}{\mathsf{All}(\neg\neg\mathsf{g},\mathsf{g})}}{\mathsf{Some}(\mathsf{t},\mathsf{g})}\,\text{Darii}}{\mathsf{Some}(\mathsf{g},\mathsf{t})}\,\text{Conversion}}{\mathsf{Some}(\mathsf{g},\mathsf{g})}\,\text{Existence} \qquad \mathsf{All}(\mathsf{g},\neg\mathsf{g})}{\mathsf{Some}(\mathsf{g},\neg\mathsf{g})}\,\text{Darii} \qquad \dfrac{}{\mathsf{All}(\mathsf{g},\mathsf{g})}}{\mathsf{All}(\mathsf{t},\neg\mathsf{g})}\,\text{RAA}$$

$\mathsf{All}(\mathsf{g},\widehat{\mathsf{K}}_\mathsf{i}\mathsf{g})$ can be proved based on the T-axiom $\mathsf{All}(\mathsf{K}_\mathsf{i}\mathsf{g},\mathsf{g})$ and Contrapositive above. $\mathsf{All}(\mathsf{K}_\mathsf{i}\mathsf{g},\widehat{\mathsf{K}}_\mathsf{i}\mathsf{g})$ follows by Barbara.

$\mathsf{All}(\mathsf{K}_\mathsf{i}\neg\neg\mathsf{g},\mathsf{K}_\mathsf{i}\mathsf{g})$ and $\mathsf{All}(\mathsf{K}_\mathsf{i}\mathsf{g},\mathsf{K}_\mathsf{i}\neg\neg\mathsf{g})$ can be shown by applying K principle on $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g},\neg\neg\mathsf{g})$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\neg\neg\mathsf{g},\mathsf{g})$. ■

Recall that Σ is inconsistent iff it can derive a contradiction. We can show:

Proposition 15 *A set of formulas Σ is inconsistent iff $\Sigma \vdash \mathsf{Some}(\mathsf{g},\neg\mathsf{g})$.*

PROOF $\Leftarrow$: $\Sigma \vdash \mathsf{All}(\mathsf{g},\mathsf{g})$ since it is an axiom. But by assumption, $\Sigma \vdash \neg\mathsf{All}(\mathsf{g},\mathsf{g}) = \mathsf{Some}(\mathsf{g},\neg\mathsf{g})$. $\Rightarrow$: Without loss of generality, assume $\Sigma \vdash \mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2), \mathsf{All}(\mathsf{g}_1,\neg\mathsf{g}_2)$, then by conversion and Darii, $\Sigma \vdash \mathsf{Some}(\mathsf{g}_2,\neg\mathsf{g}_2)$. ■

5 Completeness

Now we proceed to prove that $\mathbb{T}_{NES}$ is strongly complete w.r.t. reflexive frames. The result can be easily generalized to show the completeness of $\mathbb{S}4_{NES}$ and $\mathbb{S}5_{NES}$ w.r.t. their corresponding classes of frames, to which we will come back at the end of the section.

The completeness proof is based on the canonical (Kripke) model construction, similar to the case of modal logic. However, the language L_{NES} is significantly weaker than the full language of FOML, which introduces some difficulties. In particular, L_{NES} is essentially *not* closed under subformulas: if we view our $\mathsf{Some}(\mathsf{g}_1,\mathsf{g}_2)$ and $\mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)$ as $\exists x(g_1(x) \wedge g_2(x))$ and $\forall x(g_1(x) \to g_2(x))$, then $g_1(x)$ and $g_2(x)$ are not expressible as *formulas* in L_{NES}. Therefore in constructing the canonical model, we need to supplement each maximal consistent set Δ with a proper "maximal consistent set" of terms for each object, which can be viewed as a description of the object. Inspired by [7], we define some notion of *types* to capture such descriptions, which closely resembles the concept of *points* in [7],[3] in the setting of the orthoposet-based algebraic semantics for a (non-modal) syllogistic logic.[4]

Moreover, to prove the truth lemma eventually, we need Lemma 23 which asserts that a set of existential sentences is consistent iff each single one of them is consistent. The lemma is equivalent to the assertion that in $\mathbb{T}_{NES}$, every existential sentence brings no new universal consequences. The seemingly obvious statement is actually non-trivial since our system allows RAA and hence does not allow an easy inductive proof on deduction steps. We leave it to future work for finding an alternative direct proof system without RAA. For now, we need to construct a simpler canonical model to show Lemma 23 in the coming subsection, which also leads to the weak completeness of $\mathbb{T}_{NES}$.

5.1 Satisfiability of Existential Formulas and Weak Completeness

Inspired by the notion of *point* in [7], we first define the *types* as maximal descriptions of objects using terms. Obviously, an object must respect the universal formulas, and be either g or not g but not both for every term g. This will give us some basic properties of types.

Definition 16 (Type) *A* type $\mathscr{X}$ *is a subset of* $Term^{NES}(U)$ *s.t.*

- *If* $g_1 \in \mathscr{X}$ *and* $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}_1,\mathsf{g}_2)$*, then* $g_2 \in \mathscr{X}$*. (Respects Provably Barbara)*
- *For all* $g \in Term^{NES}(U)$*, either* $g \in \mathscr{X}$ *or* $\neg g \in \mathscr{X}$*. (Completeness)*
- *For all* $g \in Term^{NES}(U)$*,* $g, \neg g$ *are not both in* $\mathscr{X}$*. (Consistency)*

Denote the set of all types by $\mathbb{W}$*.*

Definition 17 *A collection* $\mathscr{Y}$ *of terms is said to be* possible *if for all* $g_1, g_2 \in \mathscr{Y}$*,* $\nvdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}_1,\neg\mathsf{g}_2)$.

Note that all types are possible: If $g_1, g_2 \in \mathscr{X} \in \mathbb{W}$ satisfies $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}_1,\neg\mathsf{g}_2)$, then since $\mathscr{X}$ respects provably Barbara, $\neg g_2 \in \mathscr{X}$, contradicting the consistency of $\mathscr{X}$.

Lemma 18 (Witness Lemma for Possible Collection of Terms) *If* $\mathscr{X}_0$ *is possible, then there is a type* $\mathscr{X} \in \mathbb{W}$ *extending it.*

PROOF Enumerate all terms in $Term^{NES}(U)$ as $s_0, s_1, \ldots$. We will construct a series of subsets of $Term^{NES}(U)$, $\mathscr{X}_0 \subseteq \mathscr{X}_1 \subseteq \mathscr{X}_2 \ldots$ s.t.

- For all $t_1, t_2 \in \mathscr{X}_n$, $\nvdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}_1,\neg\mathsf{t}_2)$. ($\mathscr{X}_n$ is possible)
- $\mathscr{X}_{n+1}$ is $\mathscr{X}_n \cup \{s_{n+1}\}$ or $\mathscr{X}_n \cup \{\neg s_{n+1}\}$.

[3] It is also called a *quantum state* in [2].

[4] The completeness of the (non-modal) syllogistic logic in [7] was proved via a representation theorem of orthoposets. Our proofs below are self-contained and do not rely on the results of orthoposets.

Now we show that each possible $\mathscr{X}_n$ can be extended into a possible $\mathscr{X}_{n+1}$. Given $\mathscr{X}_n$ that is possible, prove that at least one of $s_{n+1}, \neg s_{n+1}$ can be added to $\mathscr{X}_n$ to form $\mathscr{X}_{n+1}$ that is possible. Assume $\mathscr{X}_n \cup \{s_{n+1}\}$ is not possible, then $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}, \neg \mathsf{t}')$ for some $t, t' \in \mathscr{X} \cup \{s_{n+1}\}$. We need to show that $\nvdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \neg \mathsf{g}')$ for all $g, g' \in \mathscr{X} \cup \{\neg s_{n+1}\}$. Suppose not, then $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \neg \mathsf{g}')$ for some $g, g' \in \mathscr{X} \cup \{\neg s_{n+1}\}$. Since $\mathscr{X}_n$ is possible, *at least one* of t and t' must be s_{n+1}, and *at least one* of g and g' must be $\neg s_{n+1}$. Furthermore, by Proposition 12 and soundness, $\nvdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{u}, \neg \mathsf{u})$. Therefore *exactly one of* t and t' is s_{n+1}, and *exactly one of* g and g' is $\neg s_{n+1}$. In the following we derive contradictions from $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \neg \mathsf{g}')$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}, \neg \mathsf{t}')$ based on four cases.

Let us consider the case when $t' = s_{n+1}$ and $g' = \neg s_{n+1}$, thus $t, g \in \mathscr{X}_n$. By double negation axiom, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \mathsf{s}_{\mathsf{n}+1})$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}, \neg \mathsf{s}_{\mathsf{n}+1})$. Then we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}, \neg \mathsf{g})$ by contrapositive and Barbara. Then it contradicts to the assumption that $\mathscr{X}_n$ is possible and we are done. The case when $t = s_{n+1}$ and $g = \neg s_{n+1}$ can be proved similarly using contrapostive and double negation.

Now let us consider the case when $t = s_{n+1}$ and $g' = \neg s_{n+1}$, then we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \mathsf{s}_{\mathsf{n}+1})$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{s}_{\mathsf{n}+1}, \neg \mathsf{t}')$. By Barbara, we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \neg \mathsf{t}')$, contradicting to the assumption that $\mathscr{X}_n$ is possible. Similar for the case when $t' = s_{n+1}$ and $g = \neg s_{n+1}$.

Consequently, at least one of $s_{n+1}, \neg s_{n+1}$ can be added to $\mathscr{X}_n$ to form $\mathscr{X}_{n+1}$ that is possible.

Let $\mathscr{X} = \bigcup_{n \in \mathbb{N}} \mathscr{X}_n$. Note that each $t \in \mathscr{X}$ has to be added or "readded" at some finite step $\mathscr{X}_k$ thus any two $t_1, t_2 \in \mathscr{X}$ must be included in some $\mathscr{X}_j$. Therefore $\not\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}_1, \neg \mathsf{t}_2)$ since all the $\mathscr{X}_n$ are possible.

Finally, we prove that $\mathscr{X}$ is a type. It is complete since one of $s_n, \neg s_n$ is added at some $\mathscr{X}_n$. It is consistent since if $t, \neg t \in \mathscr{X}$, but by axiom double negation we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}, \neg\neg \mathsf{t})$, contradicting the fact that $\mathscr{X}$ is possible. Now for provably Barbara: If $t_1 \in \mathscr{X}$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}_1, \mathsf{t}_2)$, then $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t}_1, \neg\neg \mathsf{t}_2)$, hence $\neg t_2 \notin \mathscr{X}$ since $\mathscr{X}$ is possible. By its completeness, $t_2 \in \mathscr{X}$. ■

In the following, we build a canonical model for consistent sets of *existential formulas*. Note that we use a fixed set $\mathbb{N}$ as the domain and assign a type to each number in $\mathbb{N}$ on each world, i.e., a world is simply a function from natural numbers to types. The accessibility relation is defined as usual in modal logic.

Definition 19 (Canonical Model for Existential Formulas) *The canonical model for existential formulas of $\mathbb{T}_{NES}$ is defined as $\mathscr{M}^E = (W^E, \{R_i^E\}_{i \in I}, D^E, \rho^E)$, where:*

- $W^E = \mathbb{W}^{\mathbb{N}}$. *That is: a world w is a map from $\mathbb{N}$ to types.*
- $w_1 R_i^E w_2$ *iff* $\mathsf{K}_i g \in w_1(n)$ *entails* $g \in w_2(n)$ *for all* $n \in \mathbb{N}$, $g \in Term^{NES}(U)$.
- $D^E = \mathbb{N}$
- $\rho^E(w, A) = \{n \mid A \in w(n)\}$.

Proposition 20 (Reflexivity) *The canonical model for existential formulas of $\mathbb{T}_{NES}$ is reflexive.*

PROOF For arbitrary $g \in Term^{NES}(U)$, $w \in W^E$, if $\mathsf{K}_i g \in w(n)$, then since $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i g}, \mathsf{g})$, and $w(n)$ respects provably Barbara, $g \in w(n)$. Hence $w R_i^E w$. ■

To show that the canonical model satisfies the desired existential formulas, the key is to show that $\rho^{E^+}(w, g) = \{n \in \mathbb{N} \mid g \in w(n)\}$. That is: an object has property g if g is in the type it corresponds to. Similar to the proof of truth lemma in propositional modal logic, we have to prove an existence lemma for the induction step for K_i. The existence lemma reads: if in w, an object is not known to be g, then w must see a world where the object is not g.

Lemma 21 (Existence Lemma) *For all w, $m \in \mathbb{N}$, $t \in Term^{NES}(U)$ s.t. $\neg\mathsf{K}_i t \in w(m)$, there is w' s.t. $wR_i^E w'$ and $\neg t \in w'(m)$.*

Proof Consider the set $\mathscr{Y} = \{g \mid \mathsf{K}_i g \in w(m)\} \cup \{\neg t\}$. Prove that it is possible. Towards a contradiction, suppose $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t_1}, \neg\mathsf{t_2})$ for some $t_1, t_2 \in \{g \mid \mathsf{K}_i g \in w(m)\} \cup \{\neg t\}$. By K principle we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i t_1}, \mathsf{K_i}\neg\mathsf{t_2})$. There are three cases to be considered.

Consider the case where $\mathsf{K}_i t_1, \mathsf{K}_i t_2 \in w(m)$. By Proposition 14, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i}\neg\mathsf{t_2}, \widehat{\mathsf{K}}_\mathsf{i}\neg\mathsf{t_2})$ and Barbara, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i t_1}, \neg\mathsf{K_i t_2})$, contradicting the fact that $w(m)$ is possible.

Suppose $t_1 = t_2 = \neg t$, by Proposition 14, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i}\neg\mathsf{t}, \mathsf{K_i t})$ which entails $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i}\neg\mathsf{t}, \widehat{\mathsf{K}}_\mathsf{i}\mathsf{t})$ but this is not possible by soundness, since it is not valid over T-models according to Proposition 12.

If $\mathsf{K}_i t_1 \in w(m)$ and $t_2 = \neg t$, then $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t_1}, \neg\neg\mathsf{t})$ entails $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i t_1}, \mathsf{K_i t})$, which contradicts the fact that $\neg\mathsf{K}_i t \in w(m)$ and $w(m)$ is possible. If $\mathsf{K}_i t_2 \in w(m)$ and $t_1 = \neg t$, it leads to a contradiction as well since from $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i t_1}, \neg\mathsf{K_i t_2})$, we have the symmetric $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i t_2}, \neg\mathsf{K_i t_1})$ by contrapostive.

Consequently $\not\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t_1}, \neg\mathsf{t_2})$ for all $t_1, t_2 \in \mathscr{Y}$. By Lemma 18, $\mathscr{Y} = \{g \mid \mathsf{K}_i g \in w(m)\} \cup \{\neg t\}$ can be extended to a type. Denote it by $\mathscr{X}_m$. Clearly, by repeating the reasoning in the above first case, for each $n \neq m \in \mathbb{N}$ we can find an $\mathscr{X}_n \in \mathbb{W}$ such that $\{g \mid \mathsf{K}_i g \in w(n)\} \in \mathscr{X}_n$. Let w' then be defined by $w'(n) = \mathscr{X}_n$ for each n. Then $\neg t \in w'(m)$ and $wR_i^E w'$. ■

Lemma 22 (Truth Lemma for Terms) $\rho^{E^+}(w, g) = \{m \mid g \in w(m)\}$ *for all* $g \in Term^{NES}(U)$.

Proof Apply an induction on terms. The base case is true by definition.

Case 1: For $\neg g$, $\rho^{E^+}(w, \neg g) = D^E - \rho^{E^+}(w, g) = D^E - \{m \mid g \in w(m)\} = \{m \mid \neg g \in w(m)\}$. The last equality holds because types are consistent and complete.

Case 2: For $\mathsf{K}_i g$, $\rho^{E^+}(w, \mathsf{K}_i g) = \bigcap_{wR_i^E w'} \rho^{E^+}(w, g) = \bigcap_{wR_i^E w'} \{m \mid g \in w(m)\}$, which equals $\{m \mid \mathsf{K}_i g \in w(m)\}$ by the following reasoning:

$\supseteq$ side is easy to see, since if $m \in \{m \mid \mathsf{K}_i g \in w(m)\}$ and $wR_i^E w'$, then $\mathsf{K}_i g \in w(m)$ entails $g \in w'(m)$ by definition. Hence $m \in \bigcap_{wR_i^E w'} \{m \mid g \in w'(m)\}$.

$\subseteq$ side. Assume $i \in \bigcap_{wR_i^E w'} \{m \mid g \in w'(m)\}$, then $m \notin \bigcup_{wR_i^E w'} \{m \mid g \notin w'(m)\}$, by the completeness and consistency of $w(m)$, $i \notin \bigcup_{wR_i^E w'} \{m \mid \neg g \in w'(m)\}$. By Contrapositive of Existence Lemma, $m \notin \{m \mid \neg\mathsf{K}_i g \in w(m)\}$. Consequently, $m \in \{m \mid \mathsf{K}_i g \in w(m)\}$. ■

Now we can show a set of consistent existential sentences is satisfiable.

Lemma 23 (Sets of Consistent Existential Sentences are Satisfiable) *For a set of existential sentence Σ_{Some}, if $\not\vdash_{\mathbb{T}_{NES}} \neg\varphi_{Some}$ for all $\varphi_{Some} \in \Sigma_{Some}$, then Σ_{Some} is satisfiable (thus $\mathbb{T}_{NES}$-consistent).*

Proof Enumerate sentences in Σ_{Some} as $\varphi_0, \varphi_1, \ldots$. For each n, suppose $\varphi_n = \mathsf{Some}(\mathsf{g_1}, \mathsf{g_2})$, we show that $\{g_1, g_2\}$ is possible. First note that since $\neg\varphi$ is an abbreviation, the assumption $\not\vdash_{\mathbb{T}_{NES}} \neg\varphi_{Some}$ says $\not\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g_1}, \neg\mathsf{g_2})$. By contrapostive, $\not\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g_2}, \neg\mathsf{g_1})$. By Proposition 12, $\mathsf{All}(\mathsf{g_1}, \neg\mathsf{g_1}), \mathsf{All}(\mathsf{g_2}, \neg\mathsf{g_2})$ are not valid, thus cannot be proved in $\mathbb{T}_{NES}$ by soundness. Therefore $\{g_1, g_2\}$ is possible and can be extended as a type by Lemma 18; call it $\mathscr{X}_n$.

Now we can define a $w \in W^E$. If Σ_{Some} is infinite, let $w(n) = \mathscr{X}_n$ for all $n \in \mathbb{N}$; if not, let $w(n) = \mathscr{X}_n$ for $n \leq |\Sigma_{Some}|$, and $w(n) = \mathscr{X}_0$ for $n > |\Sigma_{Some}|$. Now we can show $\mathscr{M}^E, w \models_{NES} \Sigma_{Some}$ since each $\varphi_n \in \Sigma_{some}$ is at least witnessed by n due to our construction of w and Lemma 22. Consistency of Σ_{some} follows by soundness. ■

The weak completeness follows from the above lemma.

Theorem 24 (Weak Completeness) *If* $\models_{NES} \varphi$*, then* $\vdash_{\mathbb{T}_{NES}} \varphi$.

Proof By Proposition 12 and the validity of the rule of Existence, we have $\not\models_{NES} \varphi_{Some}$ for any existential sentence φ_{Some}. Hence it suffices to prove that for all universal sentence φ_{All}, if $\models_{NES} \varphi_{All}$, then $\vdash_{\mathbb{T}_{NES}} \varphi_{All}$. Which is equivalent to showing if $\not\vdash_{\mathbb{T}_{NES}} \varphi_{All}$, then $\not\models_{NES} \varphi_{All}$. Hence it suffices to show that for all existential sentence φ_{Some}, if $\not\vdash_{\mathbb{T}_{NES}} \neg\varphi_{Some}$, then φ_{Some} is satisfiabe, which follows from Lemma 23 w.r.t. a singleton set. ■

5.2 Strong Completeness

Normally, a weak completeness result naturally leads to strong completeness if the logic is compact. However, even though L_{NES} is indeed compact as it is a fragment of FOML, strong completeness does not easily follow and requires an argument based on Lemma 23. That is because in syllogistic, formulas are not closed under conjunction. Consequently, weak completeness does not lead to the satisfiability of every finite consistent formula set. Now we proceed to give a proof of strong completeness, again by building a (more complicated) canonical models, but for arbitrary maximal consistent sets.

Again, inspired by the notion of *point* in [7], we define the Δ-*type* to describe the sets of maximal properties an object may exemplify given the maximal consistent set Δ.

Definition 25 (Δ-type) *Given an MCS* Δ*, a* Δ-type*, denoted by* $\mathscr{X}$ *is a subset of* $Term^{NES}(U)$ *s.t.*

- *If* $g_1 \in \mathscr{X}$ *and* $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}_1, \mathsf{g}_2)$*, then* $g_2 \in \mathscr{X}$*. (Respects Barbara)*
- *For all* $g \in Term^{NES}(U)$*, either* $g \in \mathscr{X}$ *or* $\neg g \in \mathscr{X}$*. (Completeness)*
- *For all* $g \in Term^{NES}(U)$*,* $g, \neg g$ *are not both in* $\mathscr{X}$*. (Consistency)*

Denote the set of all Δ*-types by* $\mathbb{W}(\Delta)$.

Given an existential sentence $\mathsf{Some}(\mathsf{g}_1, \mathsf{g}_2) \in \Delta$, we expect there to be some type $\mathscr{X}$ exemplifying both g_1, g_2. To show this, we first generalize the notion of a *possible* set of terms w.r.t. a maximal consistent set Δ.

Definition 26 *Given a maximal consistent set* Δ*, call a set of terms* $\mathscr{Y}$ Δ-possible*, if for all* $t_1, t_2 \in \mathscr{Y}$*,* $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{Some}(\mathsf{t}_1, \mathsf{t}_2)$.

It is easy to see that the Δ-types are Δ-possible based on the fact that Δ is an MCS. The following lemma is the counterpart of Lemma 11.2 in [7] in the setting of orthoposet-based algebraic semantics. We present the following direct proof in our setting.

Lemma 27 (Witness Lemma for Δ-Possible Collection) *Each set of terms* $\mathscr{X}_0$ *that is* Δ*-possible can be extended to a* Δ*-type* $\mathscr{X} \in \mathbb{W}(\Delta)$.

Proof Enumerate all terms in $Term^{NES}(U) - \mathscr{X}_0$ as $\{s_n\}$. Construct a series of subsets of $Term^{NES}(U)$ s.t. $\mathscr{X}_0 \subseteq \mathscr{X}_1 \subseteq \mathscr{X}_2 \ldots$ and:

- $\mathscr{X}_n$ is Δ-possible: For all $t_1, t_2 \in \mathscr{X}_n$, $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{Some}(\mathsf{t}_1, \mathsf{t}_2)$.
- $\mathscr{X}_{n+1} = \mathscr{X}_n \cup \{g\}$, where $g = s_n$ or $\neg s_n$.

Now we show by induction that such a sequence can be constructed.

By assumption $\mathscr{X}_0$ is Δ-possible.

Given $\mathscr{X}_n$ s.t. for all $t_1, t_2 \in \mathscr{X}_n$, $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{Some}(\mathsf{t}_1, \mathsf{t}_2)$, and s_{n+1}, prove that at least one of $s_{n+1}, \neg s_{n+1}$ can be added to $\mathscr{X}_n$ to form $\mathscr{X}_{n+1}$ s.t. it remains Δ-possible. Essentially, we have to show either (1)

$\mathsf{Some}(\mathsf{t},\mathsf{s_{n+1}}) \in \Delta$ for all $t \in \mathscr{X}_n$ and $\mathsf{Some}(\mathsf{s_{n+1}},\mathsf{s_{n+1}}) \in \Delta$, or (2) $\mathsf{Some}(\mathsf{t},\neg\mathsf{s_{n+1}}) \in \Delta$ for all $t \in \mathscr{X}_n$ and $\mathsf{Some}(\neg\mathsf{s_{n+1}},\neg\mathsf{s_{n+1}}) \in \Delta$.

We prove that not (1) leads to (2). If (1) is not the case, there are two cases. Case 1: $\mathsf{Some}(\mathsf{s_{n+1}},\mathsf{s_{n+1}}) \notin \Delta$. Then $\mathsf{All}(\mathsf{s_{n+1}},\neg\mathsf{s_{n+1}}) \in \Delta$ since Δ is maximal. By derived rule nonexistence, $\mathsf{All}(\mathsf{g},\neg\mathsf{s_{n+1}}) \in \Delta$ for all $g \in Term^{NES}(U)$. For each $t \in \mathscr{X}_n$, $\mathsf{Some}(\mathsf{t},\mathsf{t}) \in \Delta$, hence $\mathsf{Some}(\mathsf{t},\neg\mathsf{s_{n+1}}) \in \Delta$ by Darii. For $\neg s_{n+1}$, by rule non-emptiness and that $\mathsf{All}(\mathsf{s_{n+1}},\neg\mathsf{s_{n+1}}) \in \Delta$, $\mathsf{Some}(\neg\mathsf{s_{n+1}},\neg\mathsf{s_{n+1}}) \in \Delta$. Hence (2) holds.

Case 2: Suppose $\mathsf{Some}(\mathsf{t},\mathsf{s_{n+1}}) \notin \Delta$ for some $t \in \mathscr{X}_n$, we need to show that (2) holds. For $\neg s_{n+1}$, if $\mathsf{Some}(\neg\mathsf{s_{n+1}},\neg\mathsf{s_{n+1}}) \notin \Delta$, then $\mathsf{All}(\neg\mathsf{s_{n+1}},\mathsf{s_{n+1}}) \in \Delta$. Since $\mathsf{Some}(\mathsf{t},\mathsf{s_{n+1}}) \notin \Delta$, $\mathsf{All}(\mathsf{t},\neg\mathsf{s_{n+1}}) \in \Delta$, then $\mathsf{All}(\mathsf{t},\mathsf{s_{n+1}}) \in \Delta$ by Barbara, but since $t \in \mathscr{X}_n$, $\mathsf{Some}(\mathsf{t},\mathsf{t}) \in \Delta$. This leads to $\mathsf{Some}(\mathsf{t},\mathsf{s_{n+1}}) \in \Delta$, a contradiction to the assumption. We still need to show $\mathsf{Some}(\mathsf{t}',\neg\mathsf{s_{n+1}}) \in \Delta$ for all $t' \in \mathscr{X}_n$. Assume towards a contradiction that $\mathsf{Some}(\mathsf{t}',\neg\mathsf{s_{n+1}}) \notin \Delta$ for some $t' \in \mathscr{X}_n$, then $\mathsf{All}(\mathsf{t}',\neg\neg\mathsf{s_{n+1}}) \in \Delta$. Since $\mathsf{Some}(\mathsf{t},\mathsf{s_{n+1}}) \notin \Delta$ then $\mathsf{All}(\mathsf{t},\neg\mathsf{s_{n+1}}) \in \Delta$. The following deduction shows that $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{t},\neg\mathsf{t}')$, contradicting $\mathsf{Some}(\mathsf{t},\mathsf{t}') \in \Delta$, which follows from our induction assumption that $\mathscr{X}_n$ is Δ-possible.

$$\cfrac{\mathsf{All}(\mathsf{t},\neg\mathsf{s_{n+1}}) \qquad \cfrac{\cfrac{\mathsf{All}(\mathsf{t}',\neg\neg\mathsf{s_{n+1}}) \qquad \cfrac{}{\mathsf{All}(\neg\neg\mathsf{s_{n+1}},\mathsf{s_{n+1}})}}{\mathsf{All}(\mathsf{t}',\mathsf{s_{n+1}})}\text{Barbara}}{\mathsf{All}(\neg\mathsf{s_{n+1}},\neg\mathsf{t}')}\text{Contrapositive}}{\mathsf{All}(\mathsf{t},\neg\mathsf{t}')}\text{Barbara}$$

Consequently, either (1) or (2) holds and at least one of $s_{n+1}, \neg s_{n+1}$ can be added to $\mathscr{X}_n$ to form $\mathscr{X}_{n+1}$ that is Δ-possible.

Let $\mathscr{X} = \bigcup_{n\in\mathbb{N}} \mathscr{X}_n$. Then $\Delta \vdash_{\mathbb{T}_{NES}} \mathsf{Some}(\mathsf{t_1},\mathsf{t_2})$ for all $t_1, t_2 \in \mathscr{X}$.

Finally, we prove that $\mathscr{X}$ is a Δ-type. It is complete since one of $s_n, \neg s_n$ is added at each step, and all predicates in U are eventually visited. It is consistent since Δ is consistent, so $\Delta \nvdash_{\mathbb{T}_{NES}} \mathsf{Some}(\mathsf{t},\neg\mathsf{t})$ for all t, hence $t, \neg t$ can't both be in $\mathscr{X}$. Finally we show that it respects Barbara: If $t_1 \in \mathscr{X}$ and $\mathsf{All}(\mathsf{t_1},\mathsf{t_2}) \in \Delta$, then $\neg t_2 \notin \mathscr{X}$, otherwise we have $\mathsf{Some}(\mathsf{t_1},\neg\mathsf{t_2}) \in \Delta$, contradicting the consistency of Δ. By completeness, $t_2 \in \mathscr{X}$. ■

Now we start to construct a canonical model for $\mathbb{T}_{NES}$, and show that every maximal consistent set is satisfiable in it. Compared to the previous construction, we now need to take the maximal consistent sets (MCS) into consideration. A world w is a pair of an MCS Δ and a map from $\mathbb{N}$ to $\mathbb{W}(\Delta)$. By abusing the notation, as in the previous subsection, we write $w(m)$ for $f(m)$ if $w = \langle \Delta, f \rangle$.

Definition 28 (Canonical Model for $\mathbb{T}_{NES}$) *The canonical model for $\mathbb{T}_{NES}$ is defined as*

$$\mathscr{M}^* = (W^*, \{R_i^*\}_{i\in I}, D^*, \rho^*)$$

where:

- $W^* = \bigcup_{\Delta\in MCS}\{\langle\Delta, f\rangle \mid f \in \mathbb{W}(\Delta)^{\mathbb{N}}\}$.
- wR_i^*w' *iff* $\mathsf{K}_i g \in w(m)$ *entails* $g \in w'(m)$ *for all* $m \in \mathbb{N}$, $g \in Term^{NES}(U)$.
- $D^* = \mathbb{N}$
- $\rho^*(\langle\Delta, f\rangle, A) = \{m \in \mathbb{N} \mid A \in f(m)\}$ *for all* $A \in U$.

It is not hard to show reflexivity as in the previous subsection.

Lemma 29 (Reflexivity) *The canonical model for $\mathbb{T}_{NES}$ is reflexive.*

PROOF Take arbitrary $g \in Term^{NES}(U)$, $w \in W^*$. Assume Δ is the maximal consistent set behind w. If $\mathsf{K}_i g \in w(m)$, then since $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i g}, \mathsf{g})$, $\mathsf{All}(\mathsf{K_i g}, \mathsf{g}) \in \Delta$. Then $g \in w(m)$ since $w(m)$ respects Barbara. Hence $wR_i^* w$. ■

Lemma 30 (Existence Lemma) *For all w, $m \in \mathbb{N}$, $t \in Term^{NES}(U)$ s.t. $\neg \mathsf{K}_i t \in w(m)$, there is w' s.t. $wR_i^* w'$ and $\neg t \in w'(m)$.*

PROOF Assume $w = \langle \Delta, f \rangle$ where Δ is a maximal consistent set. Consider $\Sigma = \{\mathsf{Some}(\mathsf{g}, \neg \mathsf{t}) \mid \mathsf{K}_i g \in w(m)\} \cup \bigcup_{n \in \mathbb{N}} \{\mathsf{Some}(\mathsf{g_1}, \mathsf{g_2}) \mid \mathsf{K}_i g_1, \mathsf{K}_i g_2 \in w(n)\}$, where the second part of the union is to make sure we can obtain the right types. We show Σ is consistent. Note that Σ is made up of existential sentences only, thus by Lemma 23, it suffices to prove that $\not\vdash_{\mathbb{T}_{NES}} \neg \varphi$ for all $\varphi \in \Sigma$.

Given $\varphi = \mathsf{Some}(\mathsf{g}, \neg \mathsf{t}) \in \Sigma$ for some $\mathsf{K}_i g \in w(m)$, assume for contradiction that $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g}, \mathsf{t})$. Then by K principle, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i g}, \mathsf{K_i t})$, hence $\mathsf{All}(\mathsf{K_i g}, \mathsf{K_i t}) \in \Delta$ and $\mathsf{K}_i t \in f(m)$ since $f(m)$ respects Barbara, but $\neg \mathsf{K}_i t \in w(m)$, contradicting consistency of $w(m)$.

Given $\varphi = \mathsf{Some}(\mathsf{g_1}, \mathsf{g_2}) \in \Sigma$ for some $\mathsf{K}_i g_1, \mathsf{K}_i g_2 \in w(n)$, assume for contradiction $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{g_1}, \neg \mathsf{g_2})$. Again by K principle, $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i g_1}, \mathsf{K_i} \neg \mathsf{g_2})$. By Proposition 14, we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i} \neg \mathsf{g_2}, \widehat{\mathsf{K}}_\mathsf{i} \neg \mathsf{g_2})$ and $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\widehat{\mathsf{K}}_\mathsf{i} \neg \mathsf{g_2}, \neg \mathsf{K_i g_2})$. Now by Barbara, we have $\vdash_{\mathbb{T}_{NES}} \mathsf{All}(\mathsf{K_i g_1}, \neg \mathsf{K_i g_2})$. Then $\mathsf{All}(\mathsf{K_i g_1}, \neg \mathsf{K_i g_2}) \in \Delta$ and hence $\neg \mathsf{K}_i g_2 \in w(n)$ since $w(n)$ respects Barbara. This is a contradiction to $\mathsf{K}_i g_2 \in w(n)$ and that $w(n)$ is consistent.

Since Σ is consistent, we can expand Σ to a maximal consistent set Δ' by a Lindenbaum-like argument. Observe that for all $n \neq m$, $\mathsf{K}_i g_1, \mathsf{K}_i g_2 \in w(n)$, we have $\mathsf{Some}(\mathsf{g_1}, \mathsf{g_2}) \in \{\mathsf{Some}(\mathsf{g_1}, \mathsf{g_2}) \mid \mathsf{K}_i g_1, \mathsf{K}_i g_2 \in w(n)\} \subseteq \Delta'$, hence $\{g \mid \mathsf{K}_i g \in w(n)\}$ is Δ'-possible and can be expanded to a Δ'-type by Lemma 27, denote it by $\mathscr{X}_n$. Similarly, as $\{g \mid \mathsf{K}_i g \in w(m)\} \cup \{\neg t\}$ is possible too, it can be expanded to a Δ'-type, denote it by $\mathscr{X}_m$. Let h be a function from $\mathbb{N}$ to $\mathbb{W}(\Delta')$ s.t. $h(n) = \mathscr{X}_n$. It is clear that $(\Delta, f) R_i^* (\Delta', h)$ and $\neg t \in h(m)$. ■

Now we can establish the truth lemma similar to the one in the previous section.

Lemma 31 (Truth Lemma) $\rho^{*^+}(w, g) = \{m \in \mathbb{N} \mid g \in w(m)\}$ *for all* $g \in Term^{NES}(U)$.

Finally, we can show the strong completeness of $\mathbb{T}_{NES}$.

Theorem 32 (Strong Completeness for $\mathbb{T}_{NES}$) *$\mathbb{T}_{NES}$ is strongly complete w.r.t. the class of reflexive frames.*

PROOF As usual, we show each consistent Σ for the $\mathbb{T}_{NES}$ is satisfiable on a reflexive model.

We first expand Σ to a maximal consistent set Δ, and enumerate the existential sentences in Δ as $\psi_0, \psi_1 \ldots$. For each n, suppose $\psi_n = \mathsf{Some}(\mathsf{g_1}, \mathsf{g_2})$, $\{g_1, g_2\}$ is thus it is Δ-possible since $\mathsf{Some}(\mathsf{g_1}, \mathsf{g_2})$, $\mathsf{Some}(\mathsf{g_2}, \mathsf{g_1})$, $\mathsf{Some}(\mathsf{g_1}, \mathsf{g_1})$, $\mathsf{Some}(\mathsf{g_2}, \mathsf{g_2}) \in \Delta$ by rules of Conversion and Existence. Hence, it can be extended to a Δ-type $\mathscr{X}_n$ in $\mathbb{W}(\Delta)$. Take $f : \mathbb{N} \to \mathbb{W}(\Delta)$ s.t. $f(n) = \mathscr{X}_n$ for all n. Show that $\mathscr{M}^*, \langle \Delta, f \rangle \models_{NES} \Delta$.

For $\mathsf{All}(\mathsf{g_1}, \mathsf{g_2}) \in \Delta$: Assume $n \in \rho_w^{*^+}(g_1)$, then by Truth Lemma $g_1 \in w(n)$, then since $w(n)$ respects Barbara, $g_2 \in w(n)$ hence by truth lemma $n \in \rho_w^{*^+}(g_2)$. Consequently $\mathscr{M}^*, w \models_{NES} \mathsf{All}(\mathsf{g_1}, \mathsf{g_2})$.

For $\mathsf{Some}(\mathsf{g_1}, \mathsf{g_2}) \in \Delta$: Suppose it is enumerated as φ_n. By construction of w, $g_1, g_2 \in \mathscr{X}_n = w(n)$. By truth lemma $n \in \rho_{\Delta, f}^{*^+}(g_1) \cap \rho_{\Delta, f}^{*^+}(g_2)$. Consequently $\mathscr{M}^*, w \models_{NES} \mathsf{Some}(\mathsf{g_1}, \mathsf{g_2})$. ■

It is straightforward to adapt the completeness proof with extra axioms enforcing certain frame conditions in the canonical model.

Theorem 33 *$\mathbb{S}4_{NES}$ and $\mathbb{S}5_{NES}$ are strongly complete w.r.t. the class of reflexive and transitive frames and the class of frames with equivalence relations respectively.*

6 Conclusions and future work

In this paper, we have taken the initial steps towards developing an epistemic syllogistic framework. We provided complete axiomatizations with respect to two epistemic syllogistic languages featuring *de re* knowledge. The same techniques can be applied to *belief* instead of knowledge. In fact, for systems concerning consistent belief over serial models, we only need to replace the counterpart of axiom T with D: $\mathsf{All}(\mathsf{Kg}, \widehat{\mathsf{K}}\mathsf{g})$. Adding counterparts of axioms 4 and 5 will yield a complete system of KD45 belief. So far, the usual axioms can all enforce the canonical frame to adopt the desired structure as their modal logic counterparts. If we proceed without seriality, an additional rule is required: from $\mathsf{Some}(\mathsf{Kg}_1, \mathsf{K}\neg\mathsf{g}_1)$, infer $\mathsf{All}(\mathsf{g}_2, \mathsf{Kg}_3)$, to capture the scenario where the current world has no successor. It is evident that syllogisms can be studied in modal contexts other than the epistemic setting as well.

As for other future work, we will consider the axiomatization problem of the full language of the so-called Aristotelian Modal Logic [10], and also consider the *de dicto* readings of the modal operators. It is also interesting to study the computational properties of these logics. One observation is that, like the cases of epistemic logics of know-wh [14], these epistemic syllogistic languages that we considered are one-variable fragments of FOML that are decidable in general. We will also explore the technical connections to other natural logics extending syllogistics such as [4], and to the bundled fragments of first-order modal logic where quantifiers and modalities are also packed to appear together [13, 9].

Acknowledgement This work is supported by NSSF grant 19BZX135 awarded to Yanjing Wang. The authors would like to thank Larry Moss for useful pointers and thank the three anonymous reviewers for their valuable comments that improved the presentation of the paper.

References

[1] Johan van Benthem (2008): *Natural Logic: A View from the 1980s*. In M. K. Chakraborty et al., editors: *Logic, Navya-Nyaya and Applications: Homage to Bimal Krishna Matilal*, College, London, pp. 21–42.

[2] Cristian S. Calude, Peter H. Hertling & Karl Svozil (1999): *Embedding Quantum Universes in Classical Ones*. *Foundations of Physics* 29, p. 349–379, doi:10.1023/A:1018862730956.

[3] John Corcoran (1972): *Completeness of an Ancient Logic*. *J. Symb. Log.* 37(4), pp. 696–702, doi:10.2307/2272415.

[4] Alex Kruckman & Lawrence S. Moss (2021): *Exploring the Landscape of Relational Syllogistic Logics*. *The Review of Symbolic Logic* 14(3), p. 728–765, doi:10.1017/S1755020320000386.

[5] Marko Malink (2013): *Aristotle's Modal Syllogistic*. Harvard University Press, Cambridge, MA, doi:10.4159/harvard.9780674726352.

[6] John N. Martin (1997): *Aristotle's Natural Deduction Reconsidered*. *History and Philosophy of Logic* 18(1), pp. 1–15, doi:10.1080/01445349708837269.

[7] Larry Moss (2011): *Syllogistic Logic with Complements*. In Johan van Benthem, Amitabha Gupta & Eric Pacuit, editors: *Games, Norms and Reasons*, Springer, pp. 179–197, doi:10.1007/978-94-007-0714-6_11.

[8] Larry Moss (2015): *Natural Logic*. In Chris Fox & Shalom Lappin, editors: *Handbook of Contemporary Semantic Theory*, 2 edition, Wiley-Blackwell Publishing, pp. 646–681, doi:10.1002/9781118882139.ch18.

[9] Anantha Padmanabha, R. Ramanujam & Yanjing Wang (2018): *Bundled Fragments of First-Order Modal Logic: (Un)Decidability*. In Sumit Ganguly & Paritosh K. Pandya, editors: *Proceedings of FSTTCS 2018*, *LIPIcs* 122, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 43:1–43:20, doi:10.4230/LIPIcs.FSTTCS.2018.43.

[10] Clarence Lewis Protin (2022): *A Logic for Aristotle's Modal Syllogistic*. *History and Philosophy of Logic* 0(0), pp. 1–22, doi:10.1080/01445340.2022.2107382.

[11] Adriane Rini (2011): *Aristotle's Modal Proofs*. Springer, doi:10.1007/978-94-007-0050-5.

[12] Sara L. Uckelman & Spencer Johnston (2010): *A Simple Semantics for Aristotelian Apodeictic Syllogistics.* In Valentin Goranko Lev Beklemishev & Valentin Shehtman, editors: *Advances in Modal Logic*, 8, World Scientific Publishing, pp. 454–469.

[13] Yanjing Wang (2017): *A New Modal Framework for Epistemic Logic*. In Jérôme Lang, editor: *Proceedings of TARK 2017*, *EPTCS* 251, pp. 515–534, doi:10.4204/EPTCS.251.38.

[14] Yanjing Wang (2018): *Beyond Knowing That: A New Generation of Epistemic Logics*, pp. 499–533. Springer International Publishing, Cham, doi:10.1007/978-3-319-62864-6_21.

A Proof Sketch of Theorem 8

PROOF [Sketch] Note that for $\mathbb{S}_{EAS}$, since there are L_{EAS}-formulas that cannot be negated syntactically, and we cannot equate $\Sigma \vdash_{\mathbb{S}_{EAS}} \varphi$ with that $\Sigma \cup \{\neg\varphi\}$ *is a consistent set for* $\mathbb{S}_{EAS}$. Therefore we cannot reduce strong completeness to the satisfiability of any consistent set of formulas.

We leave the full proof for the extended version of this paper and only present a sketch here. Assume that Σ is consistent, otherwise the conclusion is trivial. Separate Σ into the non-modal part Σ_0 and the modal part Σ_{K}. We consider all possible maximal consistent extension of Σ_0 in Assertoric Syllogistic and denote them by $\{\Delta_i\}_{i\in I}$. For each $\Delta_i \cup \Sigma_{\mathsf{K}}$, we construct a pointed model $\mathscr{M}_i, w_i = (W^i, R^i, D^i, \rho^i), w_i$ for it.

- $W^i = \{w_i, v_0, v_1\}$.
- R^i is the reflexive closure of $\{(w_i, v_0), (w_i, v_1)\}$.
- D^i is $\Delta_{iSome+} \sqcup \Delta'_{iSome+}$, the positive existential sentences of the form $\mathsf{Some}(\mathsf{A},\mathsf{B})$ in Δ_i and its disjoint copy.
- $\rho^i_{w_i}(X) = \{\varphi, \varphi' \mid \varphi = \mathsf{Some}(\mathsf{A},\mathsf{B}) \textit{ and } \mathsf{All}(\mathsf{A},\mathsf{X}) \in \Delta_i \textit{ or } \mathsf{All}(\mathsf{B},\mathsf{X}) \in \Delta_i\}$. Where φ' is the copy of φ.

  $\rho^i_{v_0}(X) = \{a \in D^i \mid \Delta^i \vdash \mathsf{All}(\mathsf{C},\mathsf{KX}) \textit{ for some } C \textit{ with } a \in \rho^i_w(C)\} \cup \{\varphi = \mathsf{Some}(\mathsf{B},\mathsf{X}) \in \Delta_{iSome+} \mid \Delta^i \vdash \mathsf{Some}(\mathsf{B},\mathsf{KX})\}$.

  $\rho^i_{v_1}(X) = D^i - (\{a \in D^i \mid \Delta^i \vdash \mathsf{All}(\mathsf{C},\mathsf{K}\neg\mathsf{X}) \textit{ for some } C \textit{ with } a \in \rho^i_w(C)\} \cup \{\varphi' \in \Delta'_{iSome+} \mid \varphi = \mathsf{Some}(\mathsf{B},\mathsf{B}), \Delta^i \vdash \mathsf{Some}(\mathsf{B},\mathsf{K}\neg\mathsf{X})\})$.

The idea for the model is roughly the following: In the new world v_0, an object a can have a property X only if a has property C in the real world and $\Delta^i \cup \Sigma$ thinks $\mathsf{All}(\mathsf{C},\mathsf{KX})$; or $\Delta^i \cup \Sigma$ thinks $\mathsf{Some}(\mathsf{B},\mathsf{KX})$ and a happens to be $\mathsf{Some}(\mathsf{B},\mathsf{X})$. In the new world v_1, an object a has every property A unless a has property C and $\Delta^i \cup \Sigma$ thinks $\mathsf{All}(\mathsf{C},\mathsf{K}\neg\mathsf{X})$; or $\Delta^i \cup \Sigma$ thinks $\mathsf{Some}(\mathsf{B},\mathsf{K}\neg\mathsf{X})$ and a happens to be the copy of $\mathsf{Some}(\mathsf{B},\mathsf{B})$. We need a disjoint copy of Δ_{iSome+} in the domain so that the mere fact that φ happens to have property C does not validate a universal sentence.

These models collectively describe all the possible models for Σ under logical equivalence. Therefore, it can be shown that if $\Sigma \nvdash_{\mathbb{S}_{EAS}} \varphi$, we can always find a $\Delta_i \supseteq \Sigma_0$ s.t. $\mathscr{M}_i, w \not\models \varphi$. The collection of these models are called the canonical model family of Σ. Eventually, we will be able to prove that:

1. All models in the canonical model family satisfy Σ.
2. If φ is satisfied by all models in the canonical model family of Σ, then $\Sigma \vdash_{\mathbb{S}_{EAS}} \varphi$.

1. is standard practice. To show 2, we prove the converse: if $\Sigma \nvdash_{\mathbb{S}_{EAS}} \varphi$ then there is one model in the canonical family that falsify it. As an example, we sketch the proof for case $\Sigma \nvdash_{\mathbb{S}_{EAS}} \mathsf{All}(\mathsf{A},\mathsf{KB})$, the other cases are similar. Consider

$$\Sigma' = \Sigma_0 \cup \{\varphi \mid \mathsf{K}\varphi \in \Sigma_K\} \cup \{\mathsf{Some}(\mathsf{A},\neg\mathsf{C}) \mid \mathsf{All}(\mathsf{C},\mathsf{KB}) \in \Sigma_K\}$$

It can be shown to be consistent as a set of assertoric syllogistic. The main idea is that $\Sigma_0 \cup \{\varphi \mid \mathsf{K}\varphi \in \Sigma_K\}$ is proof theoretic consequence of Σ, hence it is by assumption consistent. And if $\Sigma_0 \cup \{\varphi \mid \mathsf{K}\varphi \in \Sigma_K\}$ deduces $\mathsf{All}(\mathsf{A},\mathsf{C})$ for $\mathsf{All}(\mathsf{C},\mathsf{KB}) \in \Sigma_K$, then $\Sigma \vdash_{\mathbb{S}_{EAS}} \mathsf{All}(\mathsf{A},\mathsf{KB})$, which is a contradiction to the assumption.

Finally, Σ' has a maximal consistent extension Δ^i. It can be shown that the model $\mathscr{M}_i, w_i$ for Δ^i in the canonical model falsifies $\mathsf{All}(\mathsf{A},\mathsf{KB})$.

After establishing 1 and 2, Completeness thus follows. ■

Exploiting Asymmetry in Logic Puzzles: Using ZDDs for Symbolic Model Checking Dynamic Epistemic Logic

Daniel Miedema
Bernoulli Institute
University of Groningen
The Netherlands
daniel2Miedema@gmail.com

Malvin Gattinger
ILLC
University of Amsterdam
The Netherlands
malvin@w4eg.eu

Binary decision diagrams (BDDs) are widely used to mitigate the state-explosion problem in model checking. A variation of BDDs are Zero-suppressed Decision Diagrams (ZDDs) which omit variables that must be false, instead of omitting variables that do not matter.

We use ZDDs to symbolically encode Kripke models used in Dynamic Epistemic Logic, a framework to reason about knowledge and information dynamics in multi-agent systems. We compare the memory usage of different ZDD variants for three well-known examples from the literature: the Muddy Children, the Sum and Product puzzle and the Dining Cryptographers. Our implementation is based on the existing model checker SMCDEL and the CUDD library.

Our results show that replacing BDDs with the right variant of ZDDs can significantly reduce memory usage. This suggests that ZDDs are a useful tool for model checking multi-agent systems.

1 Introduction

There are several formal frameworks for reasoning about knowledge in multi-agent systems, and many are implemented in the form of epistemic model checkers. Here we are concerned with the *data structures* used in automated epistemic reasoning. This is a non-issue in theoretical work, where Kripke models are an elegant mathematical tools. But they are not very efficient: models where agents know little tend to be the largest. More efficient representations are often based on Binary Decision Diagrams (BDDs), which use the idea that a representation of a function not depending on p can simply ignore that variable p. This fits nicely to the models encountered in epistemic scenarios, such as the famous example of the Muddy Children: If child 2 does not observe whether it is muddy, i.e. whether p_2 is true or false, then we can save memory by omitting p_2 in the encoding of the knowledge of child 2. However, which variables matter may change, and in many examples the claim that “many variables do not matter” only holds in the initial model. This motivates us to look at Zero-suppressed Decision Diagrams (ZDDs) which use an asymmetric reduction rule to omit variables that *must* be *false*, instead of the symmetric reduction rule targeting variables that *do not matter*.

Our informal research question is thus: Is it more memory efficient to have a default assumption that “anything we do not mention does not matter” or, for example “anything we do not mention must be false”? Obviously, the answer will depend on many aspects. Here we make the question precise for the case of Dynamic Epistemic Logic, and consider three well-known examples from the literature.

The article is structured as follows. We discuss related work in the rest of this section, then we provide the relevant background in Sections 2 and 3. Section 4 describes our experiment design and the formal models used. We present our results in Section 5 and conclude in Section 6.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 407–420, doi:10.4204/EPTCS.379.32

Related work Model checking aims to verify properties of formally specified systems. Standard model checking methods search through a whole state transition graph and thus suffer from the state explosion problem: the number of states grows exponentially with the number of components or agents. To tackle this problem *symbolic* methods were developed [4]. These reduce the amount of resources needed, by reasoning about sets instead of individual states. Starting with SMV from [16], most approaches use Binary Decision Diagrams (BDDs) [2] to encode Boolean functions. Zero-suppressed Decision Diagrams (ZDDs) are an adaption of BDDs, introduced by Minato [18]. ZDDs naturally fit combinatorial problems and many comparisons between BDDs and ZDDs have been done. For both an elegant introduction into the topic of BDDs and many more references we refer to [13]. Symbolic model checking using ZDDs has not been studied much, partly due to underdeveloped construction methods [19].

Most existing symbolic model checkers use temporal logics such as LTL or CTL. Yet problems come in many forms and for examples typically described using epistemic operators (e.g. in multi-agent systems), Dynamic Epistemic Logic (DEL) is an established framework [8]. Also DEL model checking can be done symbolically [1], by encoding Kripke models as so-called knowledge structures. This lead to its implementation, SMCDEL, which is extended in this work. Another encoding, sometimes also called "symbolic models", is based on mental programs [6]. In concrete applications such as "Hintikka's World" these also get encoded as BDDs [5]. To our knowledge no previous work used ZDDs or other BDD variants for DEL model checking, with the exception of [12] where Algebraic Decision Diagrams (ADDs) are used for probabilistic DEL.

Here our main research questions is: Can ZDDs be more compact than BDDs when encoding the Kripke models for classical logic puzzles? We answer this question by adding ZDD functionality to SMCDEL and then comparing the sizes for three well-known examples from the literature.

2 Theory: Decision Diagrams

Symbolic model checkers, including SMCDEL, rely on efficient representations of Boolean functions. The most widely used data structure for this are Binary Decision Diagrams (BDDs). In this section we recall their definition and explain the difference between standard BDDs and ZDDs. How Boolean functions are then used for model checking DEL will be explained in the next section. Before we get to decision diagrams we define Boolean formulas and functions.

Definition 1. *The* Boolean formulas *over a set of variables P (also called* vocabulary*) are given by* $\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi$ *where* $p \in P$. *We define* $\bot := \neg\top$, $\varphi \vee \psi := \neg(\neg\varphi \wedge \neg\psi)$ *and* $\varphi \rightarrow \psi := \neg(\varphi \wedge \neg\psi)$.

We write $\vDash$ *for the usual Boolean semantics using assignments of type* $P \rightarrow \{0,1\}$. *When P is given we identify an assignment (also called* state*) with the set of variables it maps to* 1. *A* Boolean function *is any* $f\colon \mathscr{P}(P) \rightarrow \{0,1\}$. *For any* φ *we define the Boolean function* $f_\varphi(s) := \{\mathit{if}\ s \vDash \varphi\ \mathit{then}\ 1\ \mathit{else}\ 0\}$.

For example, if our vocabulary is $P = \{p,q,r\}$ and $s(p) = 0$, $s(q) = 1$ and $s(r) = 0$ then we identify s with $\{q\}$ and we have $s \vDash (\neg p \wedge q) \vee r$. In the following we will also just write φ for f_φ. Notably, two different formulas can correspond to the same Boolean function, but not vice versa.

Definition 2. *For any* φ, ψ, *and* p, *let* $\varphi(\frac{p}{\psi})$ *be the result of replacing every occurrence of* p *in* φ *by* ψ. *For any* $A = \{p_1,\ldots,p_n\}$, *let* $\varphi(\frac{A}{\psi}) := \psi(\frac{p_1}{\psi})(\frac{p_2}{\psi})\ldots(\frac{p_n}{\psi})$. *We use* $\forall p\varphi$ *to denote* $\varphi\left(\frac{p}{\top}\right) \wedge \varphi\left(\frac{p}{\bot}\right)$. *For any* $A = \{p_1,\ldots,p_n\}$, *let* $\forall A\varphi := \forall p_1 \forall p_2 \ldots \forall p_n \varphi$.

Decision Diagrams A decision diagram is a rooted directed acyclic graph, used to encode a Boolean function. Any terminal node (i.e. leaf) is labelled with 0 or 1, corresponding to the result of the function.

Any internal node n is labelled with a variable and has two outgoing edges to successors denoted by $\textsc{Then}(n)$ and $\textsc{Else}(n)$ — each representing a possible value for the variable. A path from the root to a leaf in a decision diagram corresponds to an evaluation of the encoded function. A decision diagram is called *ordered* if the variables are encountered in the same order on all its paths.

Example 3. *The first (left-most) decision diagram in Figure 1 is a full decision tree for $q \wedge \neg r$. To evaluate it at state $\{p, q\}$ we start at the root and then go along the solid* $\textsc{Then}$*-edge because p is true, then again along a* $\textsc{Then}$*-edge as q is true and then along the dashed* $\textsc{Else}$*-edge as r is false. We get* 1 *as a result, reflecting the fact that $\{p, q\} \vDash q \wedge \neg r$. Similarly we can use the second and third diagram.*

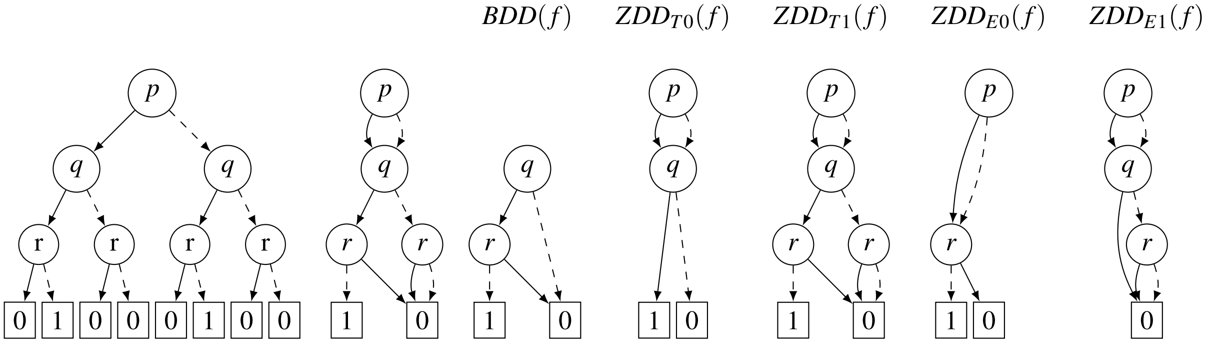


Figure 1: Seven decision diagrams for $f := q \wedge \neg r$, assuming vocabulary $\{p, q, r\}$.

Binary Decision Diagrams (BDDs) were introduced by [2] and are particularly compact decision diagrams, obtained using two reduction rules. The first rule identifies isomorphic subgraphs, i.e. we merge nodes that have the same label and the same children. In Figure 1 we get from the first to the second diagram. The second rule eliminates redundant nodes. A node is considered redundant if both its $\textsc{Then}$- and $\textsc{Else}$-edge go to the same child. In Figure 1 this gets us from the second to the third diagram.

Zero-suppressed Decision Diagrams (ZDDs) were introduced by [18] and use a different second rule than BDDs. While in BDDs a node n is eliminated when $\textsc{Then}(n) = \textsc{Else}(n)$, in ZDDs a node is eliminated when $\textsc{Then}(n) = 0$. In Figure 1 this rule gets us from the second to the fourth diagram called $ZDD_{T0}(f)$. The idea is to not ignore the variables that “do not matter” (as p in $q \wedge \neg r$), but to remove the nodes of variables that must be false (as r in $q \wedge \neg r$). To evaluate $ZDD_{T0}(f)$ at state $\{p, q\}$ we again start at the root and twice follow a solid edge because p and q are true, but then we notice that the solid edge goes from q to 1, without asking for the remaining variable r. When evaluating a ZDD_{T0} such a transition demands that the variable we “jump over” must be false — hence the name “zero-suppressed”. Indeed r is false in our state, so we do reach 1. If r would have been true, the result would have been 0.

Generalizing Elimination Rules The elimination rule “remove nodes that have a $\textsc{Then}$-edge leading to 0” can be modified in two obvious ways: instead of $\textsc{Then}$- we could consider $\textsc{Else}$-edges, and instead of 0 we could consider 1. This leads us to three additional elimination rules.

Definition 4. *We denote five different node elimination rules as follows. A node n with pairs of children $(\textsc{Then}(n), \textsc{Else}(n))$ is eliminated if it matches the left side of the rule, and any edges leading to n are diverted to the successor s on the right side of the rule.*

$$\begin{array}{lll} EQ: \ (s,s) \Rightarrow s & T0: \ (0,s) \Rightarrow s & E0: \ (s,0) \Rightarrow s \\ & T1: \ (1,s) \Rightarrow s & E1: \ (s,1) \Rightarrow s \end{array}$$

Here EQ is the rule for BDDs, while $T0$ (for “Then 0”) is the traditional ZDD rule. The remaining three are variations. For example, $E0$ says that any node with an ELSE-edge to 0 is removed, and any edge that led to the removed node should be diverted to where the THEN-edge of the removed node led.

In Figure 1 the $E0$ rule gets us from the second to the sixth diagram $ZDD_{E0}(f)$. Note that we used the rule twice: After deleting an r node the q node has an ELSE-branch to 0, so it is also eliminated. All diagrams encode the same function f, but when evaluating them we must interpret “jumps” differently.

A crucial feature of BDDs and ZDDs is that they are *canonical* representations: given a fixed variable order there is a unique BDD and a unique ZDD for each variant. It also becomes clear that for different Boolean functions a different kind of diagram can be more or less compact.

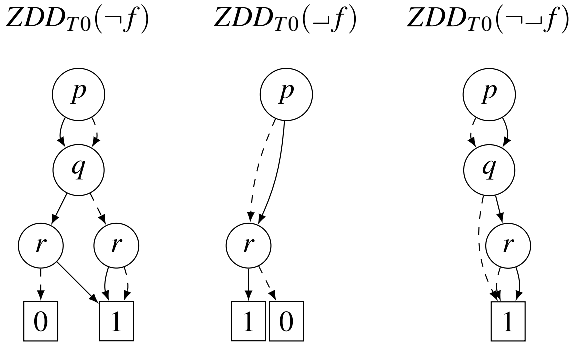
Definition 5. *For any Boolean function f, recall that $\neg f$ denotes its complement. Let $\lrcorner f$ denote the result of complementing all atomic propositions inside f. (For example, $\lrcorner(q \wedge \neg r) = \neg q \wedge r$.) For any decision diagram d, let* $\mathsf{flipLeaf}(d)$ *be the result of changing the labels of all leaves from 0 to 1 and vice versa; and let* $\mathsf{flipEdge}(d)$ *be the result of changing the labels of all edges from* THEN *to* ELSE *and vice versa.*

There is a correspondence between $\neg$ and $\mathsf{flipLeaf}$, and between $\lrcorner$ and $\mathsf{flipEdge}$. Moreover, we can use these operations to relate the four different variants of ZDDs as follows.

Fact 6. *For any Boolean function f we have:*

$$\begin{array}{rcl} DD_{T1}(f) & = & \mathsf{flipLeaf}\, DD_{T0}(\neg f) \\ DD_{E0}(f) & = & \mathsf{flipEdge}\, DD_{T0}(\lrcorner f) \\ DD_{E1}(f) & = & \mathsf{flipEdge}\, \mathsf{flipLeaf}\, DD_{T0}(\neg \lrcorner f) \end{array}$$

Example 7. *We illustrate Fact 6 using our running example $f := q \wedge \neg r$ with vocabulary $\{p,q,r\}$. Figure 2 shows the $T0$ decision diagrams mentioned in Fact 6. We see that for example $DD_{T1}(f)$ shown in Figure 1 is the same graph as $DD_{T0}(\neg f)$ with only the labels of the leaf nodes exchanged. Similarly, $DD_{E1}(f)$ in Figure 1 is the same graph as $DD_{T0}(\neg\lrcorner f)$ with flipped edges and leaves.*



$ZDD_{T0}(\neg f)$ $ZDD_{T0}(\lrcorner f)$ $ZDD_{T0}(\neg\lrcorner f)$

Figure 2: ZDDs with the same shape as the variants for $f := p \wedge \neg q$.

Fact 6 is crucial for our implementation, because the CUDD library we use does not support $T1$, $E0$ and $E1$ explicitly. Hence instead we always work with $T0$ diagrams of the negated or flipped functions.

3 Theory: Symbolic Model Checking DEL

Kripke Models We recap the standard syntax and semantics of Public Announcement Logic (PAL), the most basic version of Dynamic Epistemic Logic (DEL).

Definition 8. *Fix a vocabulary V and a finite set of agents I. The DEL language $\mathcal{L}(V)$ is given by $\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid [\varphi]\varphi$ where $p \in V$, $i \in I$.*

As usual, $K_i\varphi$ is read as *"agent i knows that φ"*. The formula $[\psi]\varphi$ says that after a *public announcement* of ψ, φ holds. The standard semantics for $\mathcal{L}(V)$ on Kripke models are as follows.

Definition 9. *A* Kripke model *for a set of agents $I = \{1,\dots,n\}$ is a tuple $\mathcal{M} = (W, \pi, \mathcal{K}_1, \dots, \mathcal{K}_n)$, where W is a set of* worlds, *π associates with each world a state $\pi(w)$, and $\mathcal{K}_1, \dots, \mathcal{K}_n$ are equivalence relations on W. A* pointed Kripke model *is a pair $(\mathcal{M}, w)$ consisting of a model and a world $w \in W$.*

Definition 10. *Semantics for $\mathcal{L}(V)$ on pointed Kripke models are given inductively as follows.*

- $(\mathcal{M}, w) \vDash p$ *iff* $\pi^M(w)(p) = \top$.
- $(\mathcal{M}, w) \vDash \neg\varphi$ *iff not* $(\mathcal{M}, w) \vDash \varphi$
- $(\mathcal{M}, w) \vDash \varphi \wedge \psi$ *iff* $(\mathcal{M}, w) \vDash \varphi$ *and* $(\mathcal{M}, w) \vDash \psi$
- $(\mathcal{M}, w) \vDash K_i\varphi$ *iff for all* $w' \in W$, *if* $w\mathcal{K}_i^M w'$, *then* $(\mathcal{M}, w') \vDash \varphi$.
- $(\mathcal{M}, w) \vDash [\psi]\varphi$ *iff* $(\mathcal{M}, w) \vDash \psi$ *implies* $(\mathcal{M}^\psi, w) \vDash \varphi$ *where* $\mathcal{M}^\psi$ *is a new model based on the set* $W^{\mathcal{M}^\psi} := \{w \in W^{\mathcal{M}} \mid (\mathcal{M}, w) \vDash \psi\}$ *and appropriate restrictions of* $\mathcal{K}_i$ *and* π *to* $W^{\mathcal{M}^\psi}$.

More expressive versions of DEL also include common knowledge and complex epistemic or ontic actions such as private communication, interception, spying and factual change. Moreover, DEL can work both with S5 models and with arbitrary Kripke models. All of this is compatible with the symbolic semantics we recall in the next section, but for our purposes in this article the restricted language above is sufficient, and we only consider S5 models.

Knowledge Structures While the semantics described above is standard, it has the disadvantage that models are represented explicitly, i.e. the number of worlds also determines the amount of memory needed to represent a model. To combat this well-known state-explosion problem we can replace Kripke models with symbolic knowledge structures. Their main advantage is that knowledge and results of announcements can be computed via purely Boolean operations, as shown in [1].

Definition 11. *Suppose we have n agents. A* knowledge structure *is a tuple $\mathcal{F} = (V, \theta, O_1, \dots, O_n)$ where V is a finite set of atomic variables, θ is a Boolean formula over V and for each agent i, $O_i \subseteq V$. The set V is the* vocabulary *and the formula θ is the* state law *of $\mathcal{F}$. The O_i are called* observational variables. *An assignment over V that satisfies θ is a* state of *$\mathcal{F}$. A* scene *is a pair $(\mathcal{F}, s)$ where s is a state of $\mathcal{F}$.*

Example 12. *Consider the knowledge structure $\mathcal{F} := (V = \{p,q\}, \theta = p \to q, O_1 = \{p\}, O_2 = \{q\})$. The states of $\mathcal{F}$ are the three assignments $\varnothing$, $\{q\}$ and $\{p,q\}$. Moreover, $\mathcal{F}$ has two agents who each observe one of the propositions: agent* 1 *knows whether p is true and agent* 2 *knows whether q is true.*

We now give semantics for $\mathcal{L}(V)$ on knowledge structures.

Definition 13. *Semantics for $\mathcal{L}(V)$ on scenes are defined as follows.*

- $(\mathcal{F}, s) \vDash p$ *iff* $s \vDash p$.
- $(\mathcal{F}, s) \vDash \neg\varphi$ *iff not* $(\mathcal{F}, s) \vDash \varphi$
- $(\mathcal{F}, s) \vDash \varphi \wedge \psi$ *iff* $(\mathcal{F}, s) \vDash \varphi$ *and* $(\mathcal{F}, s) \vDash \psi$
- $(\mathcal{F}, s) \vDash K_i\varphi$ *iff for all* t *of* $\mathcal{F}$, *if* $s \cap O_i = t \cap O_i$, *then* $(\mathcal{F}, t) \vDash \varphi$.
- $(\mathcal{F}, s) \vDash [\psi]\varphi$ *iff* $(\mathcal{F}, s) \vDash \psi$ *implies* $(\mathcal{F}^\psi, s) \vDash \varphi$ *where* $\mathcal{F}^\psi := (V, \theta \wedge \|\psi\|_{\mathcal{F}}, O_1, \dots, O_n)$.

where $\| \cdot \|_{\mathscr{F}}$ is defined in parallel in the following definition.

Definition 14. *For any knowledge structure $\mathscr{F} = (V, \theta, O_1, \dots, O_n)$ and any formula φ we define its* local Boolean translation *$\|\varphi\|_{\mathscr{F}}$ as follows.*

$$\begin{array}{lcllcl}
\|p\|_{\mathscr{F}} & := & p & \|K_i\psi\|_{\mathscr{F}} & := & \forall(V \setminus O_i)(\theta \to \|\psi\|_{\mathscr{F}}) \\
\|\neg\psi\|_{\mathscr{F}} & := & \neg\|\psi\|_{\mathscr{F}} & \|[\psi]\xi\|_{\mathscr{F}} & := & \|\psi\|_{\mathscr{F}} \to \|\xi\|_{\mathscr{F}^\psi} \\
\|\psi_1 \wedge \psi_2\|_{\mathscr{F}} & := & \|\psi_1\|_{\mathscr{F}} \wedge \|\psi_2\|_{\mathscr{F}} & & &
\end{array}$$

where the case for $K_i\psi$ quantifies over the variables not observed by agent i, using Boolean quantification as defined in Definition 2 above.

A main result from [1] based on [21] is that for any finite Kripke model there is an equivalent knowledge structure and vice versa. This means we can see knowledge structures as just another, hopefully more memory-efficient, data structure to store a Kripke model. An additional twist is that we usually store the state law θ not as a formula but only the corresponding Boolean function — which can be represented using a decision diagram as discussed in Section 2.

4 Methods: Logic Puzzles as Benchmarks

Our leading question is whether ZDDs provide a more compact encoding than BDDs for models encountered in epistemic model checking. To answer it we will work with three logic puzzles from the literature. All examples start with an initial model which we encode as a knowledge structure with the state law as a decision diagram. Then we make updates in the form of public announcements, changing the state law. We record the size of the decision diagrams for each update step.

As a basis for our implementation and experiments we use *SMCDEL*, the symbolic model checker for DEL from [1]. SMCDEL normally uses the BDD library CacBDD [15] which does not support ZDDs. Hence we also use the library CUDD [20] which does support ZDDs. However, also CUDD does not support the generalized elimination rules from Definition 4. Therefore we use Fact 6 to simulate the $T1$, $E0$ and $E1$ variants. Our new code — now merged into SMCDEL — provides easy ways to create and update knowledge structures where the state law is represented using any of the four ZDD variants.

An additional detail is that CUDD always uses so-called complement edges to optimize BDDs, but not for ZDDs. To compare the sizes of ZDDs to BDDs without complement edges we still use CacBDD. Altogether in our data set we thus record the sizes of six decision diagrams for each state law: the EQ rule with and without complement edges (called BDD and BDDc) and the four ZDD variants from Definition 4. We stress that by size of a diagram we mean the node count and not memory in bytes, because the former is independent of what libraries are used, whereas the latter depends on additional optimisations.

It now remains to choose examples. We picked three well-known logic puzzles from the literature with different kinds of state laws, such that we also expect the advantage of ZDDs to vary between them.

Muddy Children The Muddy Children are probably the best-known example in epistemic reasoning, hence we skip the explanation here and refer to the literature starting with [14]. A formalisation of the puzzle can be found in [8, Section 4.10] and the symbolic encoding in [1, Section 4].

Dining Cryptographers This problem and the protocol to solve it was first presented by [7]:

> "Three cryptographers gather around a table for dinner. The waiter informs them that the meal has been paid for by someone, who could be one of the cryptographers or the National Security Agency (NSA). The cryptographers respect each other's right to make an anonymous payment, but want to find out whether the NSA paid."

The solution uses random coin flips under the table, each observed by two neighbouring cryptographers but not visible to the third one. A formalisation and solution using Kripke models can be found in [11]. To encode the problem in a knowledge structure we let p_0 mean that the NSA paid, p_i for $i \in \{1,2,3\}$ that i paid. Moreover, p_k for $k \in \{4,5,6\}$ represents a coin. The initial scenario is then $(V = \{p_0,\dots,p_6\}, \theta = \otimes_1\{p_0,p_1,p_2,p_3\}, O_1 = \{p_1,p_4,p_5\}, O_2 = \{p_2,p_4,p_6\}, O_3 = \{p_3,p_5,p_6\})$ where the state law θ says that exactly one cryptographer or the NSA must have paid. In the solution then each cryptographer announces the XOR ($\otimes$) of all bits they observe, with the exception that the payer should invert their publicly announced bit. Formally, we get a sequence of three public announcements $[?!(\otimes p_1,p_4,p_5)][?!(\otimes p_2,p_4,p_6)][?!(\otimes p_3,p_5,p_6)]$ where $[?!\psi]\varphi := [!\psi]\varphi \wedge [\neg!\psi]\varphi$ abbreviates announcing whether. The protocol can be generalised to any odd number n instead of three participants.

Sum and Product The following puzzle was originally introduced in 1969 by H. Freudenthal. The translation is from [9] where the puzzle is also formalised in DEL:

> A says to S and P: I have chosen two integers x,y such that $1 < x < y$ and $x+y \leq 100$. In a moment, I will inform S only of $s = x+y$, and P only of $p = xy$. These announcements remain private. You are required to determine the pair (x,y). He acts as said. The following conversation now takes place: P says: "I do not know it." — S says: "I knew you didn't." — P says: "I now know it." — S says: "I now also know it." — Determine the pair (x, y).

Solving the puzzle using explicit model checking is discussed in [10]. To represent the four variables and their values in propositional logic we need a binary encoding, using $\lceil \log_2 N \rceil$ propositions for each variable that take values up to N. For example, to represent $x \leq 100$ we use $p_1,\dots,p_7$ and encode the statement $x = 5$ as $p_1 \wedge p_2 \wedge p_3 \wedge p_4 \wedge \neg p_5 \wedge p_6 \wedge \neg p_7$, corresponding to the bit-string 0000101 for 5.

The initial state law for Sum and Product is a big disjunction over all possible pairs of x and y with the given restrictions, and the observational variables ensure that agents S and P know the values of s and p respectively. For a detailed definition of the knowledge structure, see [1, Section 5].

The announcements in the dialogue are formalised as follows, combining the first two into one: First S says $K_S \neg \bigvee_{i+j\leq 100} K_P(x = i \wedge y = j)$, then P says $\bigvee_{i+j\leq 100} K_P(x = i \wedge y = j)$ and finally S says $\bigvee_{i+j\leq 100} K_S(x = i \wedge y = j)$. Solutions to the puzzle are states where these three formulas can be truthfully announced after each other. A common variation on the problem is to change the upper bound for $x+y$. We use this to turn obtain a scalable benchmark, starting with 65 to ensure there exists at least one answer.

It is well known that ZDDs perform better on sparse sets [3]. In our case, sparsity is the number of states in the model divided by the total number of possible states for the given vocabulary. Our three examples vary a lot in their sparsity: Muddy Children's sparsity is 0.5 on average (going from 0.875 to 0.125, for 3 agents), Dining Cryptographers is fairly sparse from start to finish (0.25 to 0.0625, for 3 agents), and Sum and Product is extremely sparse (e.g. starting with $< 1.369 \cdot 10^{-7}$ for $x+y \leq 100$).

5 Results

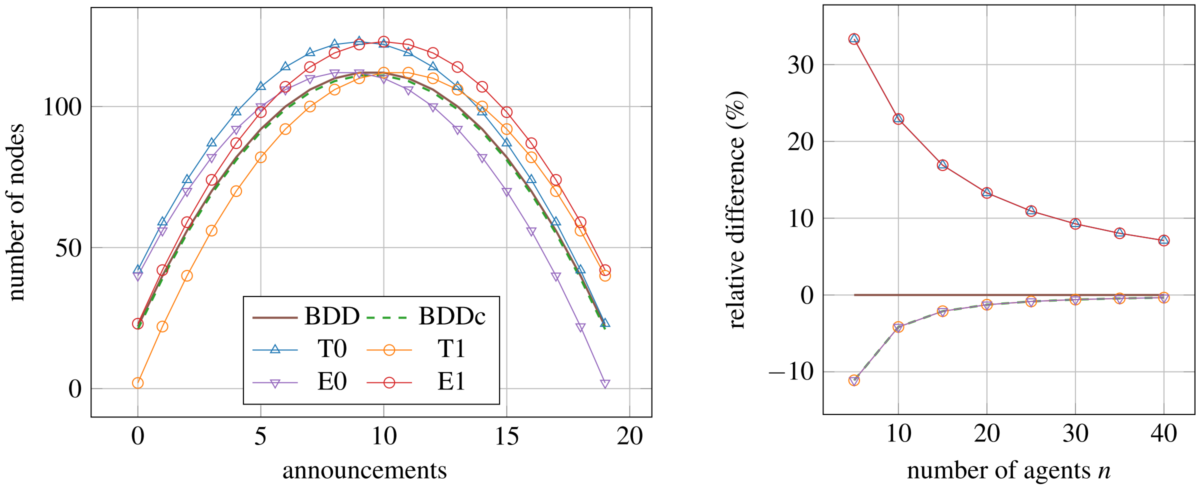
For each example we present a selection of results we deem most interesting, showing differences between BDD and ZDD sizes. The full data set for two examples can be found in the appendix where we also include instructions how all of the results can be reproduced.

Muddy children We vary the number of children n from 5 to 40, in steps of 5. We can also vary the number of muddy children $m \leq n$, but mostly report results here where $m = n$. Given any number of children, we record the size of the decision diagrams of the state law after the kth announcement, where k ranges from 0 (no announcements made yet) to $m-1$ (after which all children know their own state).

As an example, let us fix $n = m = 20$. Figure 3a shows the size of the decision diagrams after each announcement. The lines all follow a similar curve, with the largest relative differences in the initial and final states. Initially the most compact variant is T1 whereas at the end E0 is the most compact. This matches the asymmetry in the Muddy Children story: at the start the state law is $p_1 \vee \ldots \vee p_n$, hence all THEN edges lead to 1 and $T1$ removes all nodes. In contrast, at the end the state law is $p_1 \wedge \ldots \wedge p_n$ which means that all ELSE edges lead to 0 and thus $E0$ eliminates all nodes.

Hence at different stages different variants are more compact. But we want a representation that is compact throughout the whole process. We thus consider the average size over all announcements, varying n from 5 to 40. Figure 3b shows the relative size differences, with standard BDDs as 100%. The $T0$/$E1$ and the BDDc/$E0$/$T1$ lines overlap. We see that $T1$ and $E0$ are more compact for small models, but not better than BDDs with complement edges and this advantage shrinks with a larger number of agents.

We also computed sizes for $m < n$, i.e. not all children being muddy. In this case the sizes for each update step stay the same but there are fewer update steps because the last truthful announcement is in round $m-1$. As expected this is in favour of the $T1$ variant.

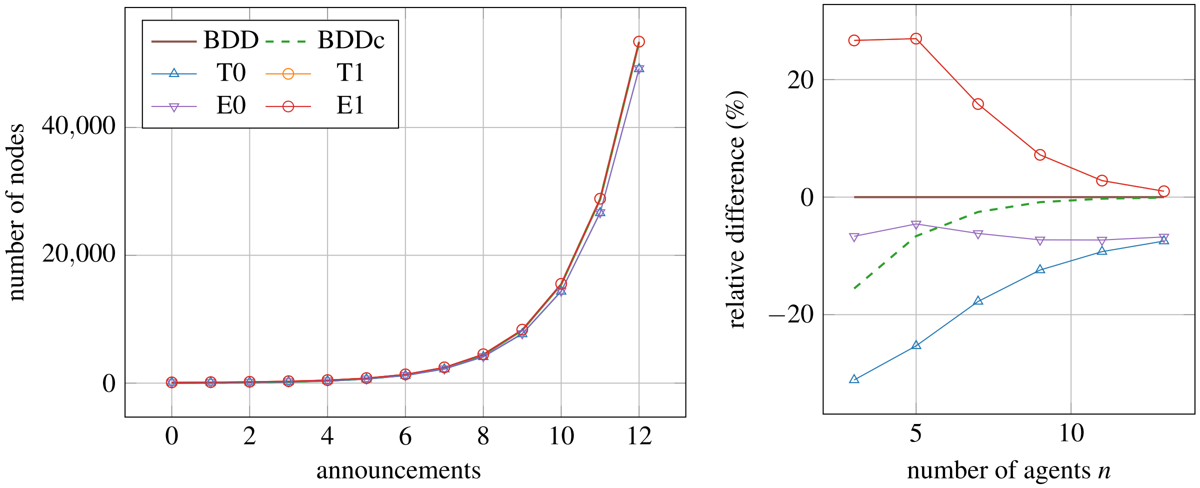


(a) Absolute sizes per announcement, for $n = 20$.

(b) Relative average sizes.

Figure 3: Results for Muddy Children.

Dining cryptographers For 13 agents we show the sizes after each announcement in Figure 4a. It becomes clear that there is little difference between the variants, which can be explained by the sparsity of the model throughout the whole story. Still, the $T0$/$E0$ variants slightly outperform the BDD(c) and the $T1$/$E1$ variants. This makes sense as most variables saying that agent i paid will be false. For lower numbers of agents the difference is larger, as visible in Figure 4b where we vary the number of agents from 3 to 13. Note that $T1$ and $E1$ overlap here, and $T0$ provides the best advantage.

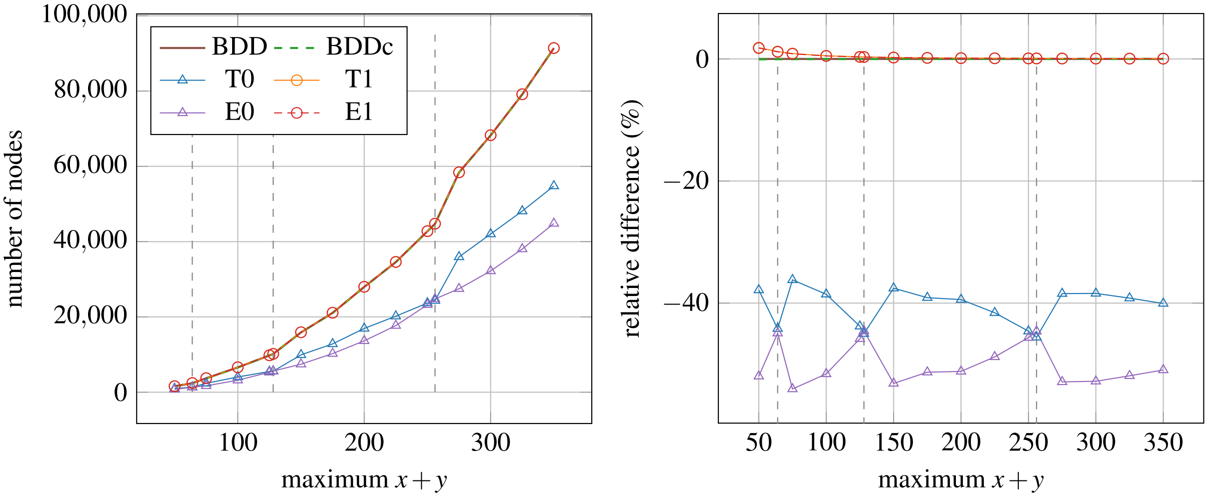


(a) Absolute sizes per announcement, for $n = 13$.

(b) Relative average sizes.

Figure 4: Results for Dining Cryptographers.

Sum and Product In this last example we can vary the upper bound of $x+y$ from 50 to 350, but not the number of agents and announcements. Figure 5a shows the sizes averaged over all four stages. We note that the BDD(c), $T1$ and $E1$ lines all overlap (with insignificant differences), and that T0 and E0 perform the best here. In contrast to the first two examples, this advantage does not disappear for larger instances of the puzzle, as can be seen in Figure 5b where we show the relative differences. Interestingly, we see that $T0$ and $E0$ meet up and diverge again wherever the bound for $x+y$ is a power of 2 (i.e. 64, 128 or 256) which we mark by vertical dashed lines. This is due to the bit-wise encoding where just above powers of two an additional bit is needed, but it must be false for almost all values.



(a) Average sizes per maximum $x+y$.

(b) Relative sizes per maximum $x+y$.

Figure 5: Results for Sum and Product.

6 Conclusion

In all experiments we find a ZDD elimination rule that can reduce the number of nodes compared to BDDs, with the exception that in the Muddy Children example complement edges provide the same advantage. This leads us to conclude that ZDDs are a promising tool for DEL model checking. Specifically, if domain knowledge about the particular model allows one to predict which ZDD variant will be more compact, ZDDs can outcompete BDDs.

The BDD elimination rule treats true and false atomic propositions symmetrically, whereas ZDD rules are asymmetric. This means their success depends on asymmetry in the model.

When translating an example from natural language to a formal models we usually try to avoid redundant variables, which already reduces the number of BDD-eliminable nodes. This is likely the reason why using ZDDs provides an advantage or, for examples with a sparsity around 0.5 like the Muddy Children, at least the same performance as BDDs with complement edges.

Specifically for logic puzzles, usually all variables are needed, and models become asymmetric and sparse as information is revealed and possible answers are ruled out. Our results confirm that sparsity and the kind of asymmetry prevalent in the model can predict which ZDD variant is most beneficial.

In this article we only considered S5. SMCDEL also provides modules for K and in further experiments we compared the sizes of ZDDs and BDDs of the state law of *belief* structures. As an example we used the famous Sally-Anne false belief task. The results were similar to those here and can be found in [17].

Future work An obvious limitation is that we only compared memory and not computation time. The size of a decision diagram correlates with the computation time needed to build it. But the step-wise construction techniques in SMCDEL are slower for ZDDs than for BDDs. For example, to compute the Sum and Product result we rather convert each state law BDD to ZDDs instead of computing ZDDs directly. Before a meaningful comparison of computation time can be done, the construction methods for ZDDs need to be further optimized.

We found some indicators which elimination rule is most compact in which case, but a more general approach to formalise domain knowledge and use it to make a correct prediction would be a powerful tool.

Acknowledgements This work is based on the master thesis [17] by the first author, written at the University of Groningen and co-supervised by Rineke Verbrugge and the second author.

We thank the TARK reviewers for their careful reading and helpful comments on this article.

References

[1] Johan van Benthem, Jan van Eijck, Malvin Gattinger & Kaile Su (2018): *Symbolic Model Checking for Dynamic Epistemic Logic — S5 and Beyond.* Logic and Computation 28(2), pp. 367—402, doi:10.1093/logcom/exx038.

[2] Randal E Bryant (1986): *Graph-based algorithms for boolean function manipulation.* IEEE Transactions on Computers 100(8), pp. 677–691, doi:10.1109/TC.1986.1676819.

[3] Randal E. Bryant (2018): *Binary Decision Diagrams.* In Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith & Roderick Bloem, editors: Handbook of Model Checking, Springer, pp. 191–217, doi:10.1007/978-3-319-10575-8_7.

[4] Jerry R Burch, Edmund M Clarke, Kenneth L McMillan, David L Dill & Lain-Jinn Hwang (1992): *Symbolic model checking:* 10^{20} *states and beyond.* Information and computation 98(2), pp. 142–170, doi:10.1016/0890-5401(92)90017-a.

[5] Tristan Charrier, Sébastien Gamblin, Alexandre Niveau & François Schwarzentruber (2019): *Hintikka's World: Scalable Higher-order Knowledge*. In: *IJCAI 2019*, pp. 6494–6496, doi:10.24963/ijcai.2019/934.

[6] Tristan Charrier, Sophie Pinchinat & François Schwarzentruber (2019): *Symbolic model checking of public announcement protocols*. *Logic and Computation* 29(8), pp. 1211–1249, doi:10.1093/logcom/exz023.

[7] David Chaum (1988): *The dining cryptographers problem: Unconditional sender and recipient untraceability*. *Journal of cryptology* 1(1), pp. 65–75, doi:10.1007/BF00206326.

[8] Hans van Ditmarsch, Wiebe van Der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. Springer, doi:10.1007/978-1-4020-5839-4.

[9] Hans van Ditmarsch, Jan van Eijck & Rineke Verbrugge (2009): *Publieke werken—freudenthal's som-en-productraadsel*. *Nieuw Archief voor Wiskunde* 10(2), pp. 126–131. Available at `https://www.nieuwarchief.nl/serie5/pdf/naw5-2009-10-2-126.pdf`.

[10] Hans van Ditmarsch, Ji Ruan & Rineke Verbrugge (2007): *Sum and Product in Dynamic Epistemic Logic*. *Logic and Computation* 18(4), pp. 563–588, doi:10.1093/logcom/exm081.

[11] Jan van Eijck & Simona Orzan (2007): *Epistemic verification of anonymity*. *Electronic Notes in Theoretical Computer Science* 168, pp. 159–174, doi:10.1016/j.entcs.2006.08.026.

[12] Sébastien Gamblin, Alexandre Niveau & Maroua Bouzid (2022): *A Symbolic Representation for Probabilistic Dynamic Epistemic Logic*. In: *AAMAS 2022*, pp. 445–453. Available at `https://dl.acm.org/doi/abs/10.5555/3535850.3535901`.

[13] Donald E. Knuth (2011): *The Art of Computer Programming, volume 4A: Combinatorial Algorithms, Part 1*. Addison-Wesley.

[14] John E Littlewood (1953): *A Mathematician's Miscellany*. Methuen and Company Limited.

[15] Guanfeng Lv, Kaile Su & Yanyan Xu (2013): *CacBDD: A BDD package with dynamic cache management*. In: *Computer Aided Verification*, Springer, pp. 229–234, doi:10.1007/978-3-642-39799-8_15.

[16] Kenneth L McMillan (1993): *Symbolic model checking*. Springer, doi:10.1007/978-1-4615-3190-6.

[17] Daniel Miedema (2022): *Zero-suppression Decision Diagrams versus Binary Decision Diagrams on Dynamic Epistemic Logic Model Checking*. Master's thesis, University of Groningen. Available at `https://fse.studenttheses.ub.rug.nl/27287/`.

[18] Shin-ichi Minato (1993): *Zero-suppressed BDDs for set manipulation in combinatorial problems*. In: *Proceedings of the 30th international Design Automation Conference*, pp. 272–277, doi:10.1145/157485.164890.

[19] Shin-ichi Minato (2001): *Zero-suppressed BDDs and their applications*. *International Journal on Software Tools for Technology Transfer* 3(2), pp. 156–170, doi:10.1007/s100090100038.

[20] Fabio Somenzi (2012): *CUDD: CU decision diagram package*. Available at `http://vlsi.colorado.edu/~fabio/CUDD/`. Version 2.5.0.

[21] K. Su, A. Sattar, G. Lv & Y. Zhang (2009): *Variable Forgetting in Reasoning about Knowledge*. *Journal of Artificial Intelligence Research* 35, pp. 677–716, doi:10.1613/jair.2750.

Appendix

The ZDD encoding of knowledge structures has been integrated into SMCDEL itself. All our results can be reproduced using the Haskell Tool *Stack* from `https://haskellstack.org` as follows.

```
git clone https://github.com/jrclogic/SMCDEL
cd SMCDEL
git checkout zdd-experiments
stack bench --no-run-benchmarks # build but do not run yet
stack bench smcdel:bench:sizes-muddychildren
stack bench smcdel:bench:sizes-diningcryptographers
stack bench smcdel:bench:sizes-sumandproduct
```

The last three commands will create `.dat` files containing the results. On a system with a 4.8 GHz CPU the last three commands above take approximately 10 seconds, one minute and three hours.

We include the results for Dining Crytographers and Sum and Product here, but omit the (several pages long) results for the Muddy Children.

Results for Dining Cryptographers

```
# Note: round -1 indicates the average.
n          m          round      BDD        BDDc       T0         T1         E0         E1
3          1          0          9          7          9          13         11         13
3          1          1          15         12         10         19         14         19
3          1          2          21         19         12         25         17         25
3          1          -1         11.25      9.5        7.75       14.25      10.5       14.25
5          1          0          13         11         18         26         22         26
5          1          1          25         20         21         38         29         38
5          1          2          41         37         29         54         40         54
5          1          3          64         61         44         77         57         77
5          1          4          98         96         68         111        82         111
5          1          -1         60.25      56.25      45.0       76.5       57.5       76.5
7          1          0          17         15         31         43         37         43
7          1          1          35         28         36         61         48         61
7          1          2          61         55         50         87         67         87
7          1          3          102        97         79         128        100        128
7          1          4          170        166        133        196        157        196
7          1          5          285        282        228        311        254        311
7          1          6          479        477        388        505        415        505
7          1          -1         287.25     280.0      236.25     332.75     269.5      332.75
9          1          0          21         19         48         64         56         64
9          1          1          45         36         55         88         71         88
9          1          2          81         73         75         124        98         124
9          1          3          140        133        118        183        147        183
9          1          4          242        236        202        285        236        285
9          1          5          423        418        359        466        397        466
9          1          6          747        743        645        790        686        790
9          1          7          1326       1323       1156       1369       1199       1369
9          1          8          2352       2350       2052       2395       2096       2395
9          1          -1         1344.25    1332.75    1177.5     1441.0     1246.5     1441.0
11         1          0          25         23         69         89         79         89
11         1          1          55         44         78         119        98         119
11         1          2          101        91         104        165        133        165
11         1          3          178        169        161        242        198        242
11         1          4          314        306        275        378        319        378
11         1          5          561        554        494        625        544        625
11         1          6          1015       1009       906        1079       961        1079
11         1          7          1852       1847       1671       1916       1730       1916
11         1          8          3392       3388       3077       3456       3139       3456
11         1          9          6211       6208       5636       6275       5700       6275
11         1          10         11333      11331      10244      11397      10309      11397
11         1          -1         6259.25    6242.5     5678.75    6435.25    5802.5     6435.25
13         1          0          29         27         94         118        106        118
13         1          1          65         52         105        154        129        154
13         1          2          121        109        137        210        172        210
13         1          3          216        205        208        305        253        305
13         1          4          386        376        352        475        406        475
13         1          5          699        690        633        788        695        788
13         1          6          1283       1275       1171       1372       1240       1372
13         1          7          2378       2371       2190       2467       2265       2467
13         1          8          4432       4426       4106       4521       4186       4521
13         1          9          8277       8272       7687       8366       7771       8366
13         1          10         15449      15445      14341      15538      14428      15538
```

```
13        1         11        28764     28761     26628     28853     26717     28853
13        1         12        53342     53340     49156     53431     49246     53431
13        1         -1        28860.25  28837.25  26702.0   29149.5   26903.5   29149.5
```

Results for Sum and Product

```
# Note: round -1 indicates the average.
n         round     BDD       BDDc      T0        T1        E0        E1
50        0         5032      5030      3082      5060      2505      5060
50        1         697       696       458       726       288       724
50        2         547       546       361       576       221       574
50        3         1         1         0         31        0         31
50        -1        1569.25   1568.25   975.25    1598.25   753.5     1597.25
64        0         7916      7914      4247      7944      4602      7944
64        1         930       929       614       959       381       957
64        2         760       759       501       789       308       787
64        3         1         1         0         31        0         31
64        -1        2401.75   2400.75   1340.5    2430.75   1322.75   2429.75
75        0         12534     12532     7847      12566     5986      12566
75        1         1274      1273      900       1307      456       1305
75        2         939       938       658       972       335       970
75        3         35        34        27        68        12        66
75        -1        3695.5    3694.25   2358.0    3728.25   1697.25   3726.75
100       0         22438     22436     13514     22471     11279     22471
100       1         2289      2288      1567      2323      870       2321
100       2         1594      1593      1087      1628      596       1626
100       3         36        35        28        70        12        68
100       -1        6589.25   6588.0    4049.0    6623.0    3189.25   6621.5
125       0         33826     33824     18476     33859     19074     33859
125       1         3149      3148      2095      3183      1262      3181
125       2         2101      2100      1383      2135      835       2133
125       3         36        35        28        70        12        68
125       -1        9778.0    9776.75   5495.5    9811.75   5295.75   9810.25
128       0         35315     35313     18823     35348     20401     35348
128       1         3149      3148      2095      3183      1262      3181
128       2         2101      2100      1383      2135      835       2133
128       3         36        35        28        70        12        68
128       -1        10150.25  10149.0   5582.25   10184.0   5627.5    10182.5
150       0         55028     55026     33874     55065     26559     55065
150       1         5147      5146      3526      5185      1938      5183
150       2         3354      3353      2261      3392      1267      3390
150       3         40        39        32        78        12        76
150       -1        15892.25  15891.0   9923.25   15930.0   7444.0    15928.5
175       0         73233     73231     43763     73270     36869     73270
175       1         6753      6752      4635      6791      2549      6789
175       2         4265      4264      2893      4303      1599      4301
175       3         40        39        32        78        12        76
175       -1        21072.75  21071.5   12830.75  21110.5   10257.25  21109.0
200       0         98044     98042     58134     98082     49615     98082
200       1         8498      8497      5929      8537      3096      8535
200       2         5275      5274      3666      5314      1877      5312
200       3         41        40        33        80        12        78
200       -1        27964.5   27963.25  16940.5   28003.25  13650.0   28001.75
225       0         121863    121861    69555     121901    64632     121901
225       1         10103     10102     6910      10142     3827      10140
225       2         6284      6283      4289      6323      2317      6321
225       3         41        40        33        80        12        78
225       -1        34572.75  34571.5   20196.75  34611.5   17697.0   34610.0
250       0         148149    148147    80231     148187    83173     148187
250       1         14131     14130     8991      14170     6071      14168
250       2         8664      8663      5470      8703      3659      8701
250       3         41        40        33        80        12        78
250       -1        42746.25  42745.0   23681.25  42785.0   23228.75  42783.5
256       0         154815    154813    81895     154853    88925     154853
```

```
256       1         14992     14991     9616      15031     6371      15029
256       2         9084      9083      5787      9123      3786      9121
256       3         41        40        33        80        12        78
256       -1        44733.0   44731.75  24332.75  44771.75  24773.5   44770.25
275       0         203227    203225    123011    203269    98715     203269
275       1         18794     18793     12876     18837     7046      18835
275       2         11435     11434     7808      11478     4189      11476
275       3         45        44        37        88        12        86
275       -1        58375.25  58374.0   35933.0   58418.0   27490.5   58416.5
300       0         238864    238862    144850    238906    116069    238906
300       1         21339     21338     14568     21382     8066      21380
300       2         12671     12670     8634      12714     4662      12712
300       3         45        44        37        88        12        86
300       -1        68229.75  68228.5   42022.25  68272.5   32202.25  68271.0
325       0         277256    277254    165770    277298    137410    277298
325       1         24822     24821     16902     24865     9451      24863
325       2         14341     14340     9749      14384     5305      14382
325       3         45        44        37        88        12        86
325       -1        79116.0   79114.75  48114.5   79158.75  38044.5   79157.25
350       0         318340    318338    187632    318382    160813    318382
350       1         29838     29837     19932     29881     11776     29879
350       2         17313     17312     11500     17356     6686      17354
350       3         45        44        37        88        12        86
350       -1        91384.0   91382.75  54775.25  91426.75  44821.75  91425.25
```

A Theory of Bounded Inductive Rationality

Caspar Oesterheld
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
oesterheld@cmu.edu

Abram Demski
Machine Intelligence Research Institute
Berkeley, California, USA

Vincent Conitzer
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA

The dominant theories of rational choice assume logical omniscience. That is, they assume that when facing a decision problem, an agent can perform all relevant computations and determine the truth value of all relevant logical/mathematical claims. This assumption is unrealistic when, for example, we offer bets on remote digits of π or when an agent faces a computationally intractable planning problem. Furthermore, the assumption of logical omniscience creates contradictions in cases where the environment can contain descriptions of the agent itself. Importantly, strategic interactions as studied in game theory are decision problems in which a rational agent is predicted by its environment (the other players). In this paper, we develop a theory of rational decision making that does not assume logical omniscience. We consider agents who repeatedly face decision problems (including ones like betting on digits of π or games against other agents). The main contribution of this paper is to provide a sensible theory of rationality for such agents. Roughly, we require that a boundedly rational inductive agent tests each efficiently computable hypothesis infinitely often and follows those hypotheses that keep their promises of high rewards. We then prove that agents that are rational in this sense have other desirable properties. For example, they learn to value random and pseudo-random lotteries at their expected reward. Finally, we consider strategic interactions between different agents and prove a folk theorem for what strategies bounded rational inductive agents can converge to.

1 Introduction

The dominant theories of rational decision making – in particular Bayesian theories – assume logical omniscience, i.e., that rational agents can determine the truth value of any relevant logical statement. In some types of decision problems, this prevents one from deriving any recommendation from these theories, which is unsatisfactory (Sect. 3). For one, there are problems in which computing an optimal choice is simply computationally intractable. For example, many planning problems are intractable. Second, the assumption of logical omniscience creates contradictions (resembling classic paradoxes of self reference, such as the liar's paradox) if the environment is allowed to contain references to the agent itself. These issues arise most naturally when multiple rational agents interact and reason about one another.

This paper develops a novel theory of boundedly rational inductive agents (BRIAs) that does not assume logical omniscience and yields sensible recommendations in problems such as the ones described above. Rather than describing how an agent should deal with an individual decision, the theory considers how an agent learns to choose on a sequence of different decision problems. We describe the setting in more detail in Sect. 2.

The core of our theory is a normative rationality criterion for such learning agents. Roughly, the criterion requires that a boundedly rational inductive agent test each efficiently computable hypothesis (or more generally each hypothesis in some class) infinitely often and follows hypotheses that keep their promises of high rewards. We describe the criterion in detail in Sect. 4. Importantly, the criterion can be satisfied by computationally bounded agents, as we show in Sect. 5.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 421–440, doi:10.4204/EPTCS.379.33

We demonstrate the appeal of our criterion by showing that it implies desirable and general behavioral patterns. In Sect. 6, we show that on sequences of decision problems in which one available option guarantees a payoff of at least l, BRIAs learn to obtain a reward of at least l. Thus, in particular, they avoid Dutch books (in the limit). We further show that similarly on sequences of decision problems in which one available option pays off truly or algorithmically randomly with mean μ, BRIAs learn to obtain a reward of at least μ. Finally, we consider decision problems in which one BRIA plays a strategic game against another BRIA. We show that BRIAs can converge to any individually rational correlated strategy profile. BRIAs are thus a promising model for studying ideas such as superrationality (i.e., cooperation in the one-shot Prisoner's Dilemma) [11] (cf. Sect. 8). Related work is discussed in Sect. 8. Throughout this paper, we describe the key ideas for our proofs in the main text. Detailed proofs are given in Appendix A.

2 Setting

Informally, we consider an agent who makes decisions in discrete time steps. At each time step she faces some set of available options to choose from. She selects one of options and receives a reward. She then faces a new decision problem, and so on.

Formally, let $\mathscr{T}$ be some language describing available *options*. A *decision problem* $\mathrm{DP} \in \mathrm{Fin}(\mathscr{T})$ is a finite set of options. A *decision problem sequence* is a sequence of decision problems $\mathrm{DP}_1, \mathrm{DP}_2, \ldots$ An agent for $\overline{\mathrm{DP}}$ is a sequence $\bar{c}$ of $c_t \in \mathrm{DP}_t$. The rewards are numbers $r_1, r_2, r_3, \ldots \in [0,1]$. Note that in contrast to the literature on multi-armed bandit problems (Sect. 8) counterfactual rewards are not defined.

It is generally helpful to imagine that (similar to multi-armed bandit problems) at each time t the agent first sees DP_t; then chooses c_t from DP_t according to some algorithm that looks at the available options in DP_t and takes past experiences into account; then the environment calculates some reward as a function of c_t; the agent observes the reward and learns from it. The sequence of decision problems DP_t may in turn be calculated depending on the agent's choices.

We focus on learning myopically optimal behavior. That is, we want our agent to learn to choose whatever gives the highest reward for the present decision problem, regardless of what consequences that has for future decision problems.

3 Computational constraints and paradoxes of self-reference

In this paper, we develop a normative theory of rational learning in this setting. The standard theory for rational decision making under uncertainty is Bayesian decision theory (BDT) ([24, 12]; for contemporary overviews, see [19, 29]). The main ideas of this paper are motivated by a specific shortcoming of BDT: the assumption that the agent who is subject to BDT's recommendations is logically omniscient and in particular not limited by any computational constraints. We develop a theory that gives recommendations to computationally bounded agents. In the following, we give two kinds of examples to illustrate the role of logical omniscience in BDT and motivate our search for an alternative theory.

Mere intractability The first problem is that in most realistic choice problems, it is intractable to follow BDT. Bayesian updating and Bayes-optimal decision making are only feasible if the environment is small or highly structured. Even if the agent had a perfectly accurate world model, determining the optimal choice may require solving computationally hard problems, such as the traveling salesman problem, planning in 2-player competitive games, etc. Optimal choice may also rely on whether particular mathematical claims are true, e.g., when assessing the safety of particular cryptographic methods. In all these problems, BDT requires the agent to perfectly solve the problem at hand. However, we would like a theory of rational choice that makes recommendations to realistic, bounded agents who can only solve such problems approximately.

Consider a decision problem $\mathrm{DP} = \{a_1, a_2\}$, where the agent knows that option a_1 pays off the value of the 10^{100}-th digit of the binary representation of π. Option a_2 pays off 0.6 with certainty. In our formalism, r equals the 10^{100}-th digit of the binary representation of π if $c = a_1$ and $r = 0.6$ if $c = 0.6$. All that Bayesian decision theory has to say about this problem is that one should calculate the 10^{100}-th digit of π; if it is 1, choose a_1; otherwise choose a_2. Unfortunately, calculating the 10^{100}-th digit of π is likely intractable.[1] Hence, Bayesian decision theory does not have any recommendations for this problem for realistic reasoners. At the same time, we have the strong normative intuition that – if digits of π indeed cannot be predicted better than random under computational limitations – it is rational to take a_2. We would like our theory to make sense of that intuition.

We close with a note on what we can expect from a theory about rational decision making under computational bounds. A naïve hope might be that such a theory could tell us how to optimally use some amount of compute (say, 10 hours on a particular computer system) to approximately solve any given problem (cf. our discussion in Sect. 8 of Russell et al.'s [21, 23, 22] work on bounded optimality); or that it might tell us *in practice* at what odds to bet on, say, Goldbach's conjecture with our colleagues. In this paper, we do not provide such a theory and such a theory cannot exist. We must settle for a more modest goal. Since our agents face decision problems repeatedly, our rationality requirement will be that the agent *learns* to approximately solve these problems optimally in the limit. For example, if digits of π are pseudo-random in the relevant sense, then a rational agent must converge to betting 50-50 on remote binary digits of π. But it need not bet 50-50 "out-of-the-box".

Paradoxes of self-reference, strategic interactions, and counterfactuals A second problem with BDT and logical omniscience more generally is that it creates inconsistencies if the values of different available options depend on what the agent chooses. As an example, consider the following decision problem, which we will call the Simplified Adversarial Offer (SAO) (after a decision problem introduced by [18]). Imagine that an artificial agent chooses between two available alternatives a_0 and a_1, where a_0 is known to pay off $1/2$ with certainty, and a_1 is known to pay off 1 if the agent's program run on this decision problem chooses a_0, and 0 otherwise. Now assume that the agent chooses deterministically and optimally given a logically omniscient belief system. Then the agent knows the value of each of the options. This also means that it knows whether it will select a_0 or a_1. But given this knowledge, the agent selects a different option than what the belief system predicts. This is a contradiction. Hence, there exists no agent that complies with standard BDT in this problem. Compare the example of Oesterheld and Conitzer [18] and Spencer [28]; also see Demski and Garrabrant ([7], Sect. 2.1) for a discussion of another, subtler issue that arises from logical omniscience and introspection.

We are particularly interested in problems in which such failure modes apply. SAO is an extreme and unrealistic example, selected to be simple and illustrative. However, strategic interactions between different rational agents share the ingredients of this problem: Agent 1 is thinking about what agent 2 is choosing, thereby creating a kind of reference to agent 2 in agent 2's environment. We might even imagine that two AI players know each others' exact source code (cf. [20], Sect. 10.4; [31]; [32]; [3]; [6]; [17]). Further, it may be in agent 2's interest to prove wrong whatever agent 1 believes about agent 2. For a closely related discussion of issues of bounded rationality and the foundations of game theory,

[1] Remote digits of π are a canonical example in the literature on bounded rationality and logical uncertainty (see [25], for an early usage). To the knowledge of the authors it is unknown whether the n-th digit of π can be guessed better than random in less than $O(n)$ time. For a general, statistical discussion of the randomness of digits of π, see Marsaglia [14].

see Binmore [4] and references therein (cf. [20], Ch. 10; [7]).

4 The rationality criterion

4.1 Preliminary definitions

An *estimating agent* $\bar{\alpha}$ is a sequence of choices from the available options $\alpha_t^c \in \mathrm{DP}_t$ and *estimates* $\alpha_t^e \in [0,1]$. Our rationality criterion uses estimating agents. For brevity, we will say *agent* instead of estimating agent throughout the rest of this paper. For example, let $\mathrm{SAO}_{\alpha,t}$ be the Simplified Adversarial Offer for the agent at time t as described in Sect. 3. Then we might like an agent who learns to choose $\alpha_t^c = a_0$ (which pays $1/2$ with certainty) and estimate $\alpha_t^e = 1/2$.

A *hypothesis* h has the same type signature as an estimating agent. When talking about hypotheses, we will often refer to the values of h_t^e as promises and to the values of h_t^c as recommendations.

Our rationality criterion will be relative to a particular set of hypotheses $\mathbb{H}$. In principle, $\mathbb{H}$ could be any set of hypotheses, e.g., all computable ones, all three-layer neural nets, all 8MB computer programs, etc. Generally, $\mathbb{H}$ should contain any hypothesis (i.e., any hypothesis about how the agent should act) that the agent is willing to consider, similar to the support of the prior in Bayesian theories of learning, or the set of experts in the literature on multi-armed bandits with expert advice. Following Garrabrant et al. [10], we will often let $\mathbb{H}$ be the set of functions computable in $O(g(t))$ time, where g is a non-decreasing function. We will call these hypotheses *efficiently computable (e.c.)*. Note that not all time complexity classes can be written as $O(g(t))$. For example, the set of functions computable in polynomial time cannot be written in such a way. This simplified set is used to keep notation simple. Our results generalize to more general computational complexity classes.

4.2 No overestimation

We now describe the first part of our rationality requirement, which is that the estimates should not be systematically above what the agent actually obtains. The criterion itself is straightforward, but its significance will only become clear in the context of the hypothesis coverage criterion of the next section.

Definition 1. *For $T \in \mathbb{N}$, we call $\mathscr{L}_T(\bar{\alpha},\bar{r}) := \sum_{t=1}^{T} \alpha_t^e - r_t$ the cumulative overestimation of an agent $\bar{\alpha}$ on $\bar{r}$.*

Definition 2. *We say that an agent $\bar{\alpha}$ for $\overline{\mathrm{DP}},\bar{r}$ does not overestimate (on average in the limit) if $\mathscr{L}_T(\bar{\alpha},\bar{r})/T \le 0$ as $T \to \infty$.*

In other words, for all $\varepsilon > 0$, there should be a time t such that for all $T > t$, $\mathscr{L}_T(\bar{\alpha},\bar{r})/T \le \varepsilon$. Note that the per-round overestimation of boundedly rational inductive agents as defined below will usually but need not always converge to 0; it can be negative in the limit.

4.3 Covering hypotheses

We come to our second requirement, which specifies how the agent $\bar{\alpha}$ relates to the hypotheses in $\mathbb{H}$.

Definition 3. *We say that $\bar{h}$ outpromises $\bar{\alpha}$ or that $\bar{\alpha}$ rejects $\bar{h}$ at time t if $h_t^e > \alpha_t^e$.*

We distinguish two kinds of hypotheses: First, there are hypotheses that promise higher rewards than $\bar{\alpha}^e$ in only finitely many rounds. For example, this will be the case for hypotheses that $\bar{\alpha}$ trusts and takes into account when choosing and estimating. Also, this could include hypotheses who recommend an inferior option with an accurate estimate, e.g., hypotheses that recommend "$1/3$" and promise $1/3$ in

$\{$"$1/3$","$2/3$"$\}$. For all of these hypotheses, we do not require anything of $\bar{\alpha}$. In particular, $\bar{\alpha}$ need not test these hypotheses. Second, some hypotheses do infinitely often outpromise $\bar{\alpha}^e$. For these cases, we will require our boundedly rational inductive agents to have some reason to reject these hypotheses. To be able to provide such a reason, $\bar{\alpha}$ needs to test these hypotheses infinitely often. Testing a hypothesis requires choosing the hypothesis' recommended action.

Definition 4. *We call a set $M \subseteq \mathbb{N}$ a test set of $\bar{\alpha}$ for $\bar{h}$ if for all $t \in M$, $\alpha_t^c = h_t^c$.*

For $\bar{\alpha}$ to infinitely often reject $\bar{h}$, these tests must then show that $\bar{h}$ is not to be trusted (in those rounds in which they promise a reward that exceeds $\bar{\alpha}^e$). That is, on these tests, the rewards must be significantly lower than what the hypothesis promises. We thus introduce another key concept.

Definition 5. *Let $\bar{h}$ be a hypothesis and $M \subseteq \mathbb{N}$ be a test set of $\bar{\alpha}$ for $\bar{h}$. We call $l_T(\bar{\alpha},\bar{r},M,\bar{h}) := \sum_{t\in M_{\leq T}} r_t - h_t^e$ the (empirical) record of h (on M).*

Here, $M_{\leq T} := \{t \in M \mid t \leq T\}$ is defined to be the set of elements of M that are at most T.

We now have all the pieces together to state the coverage criterion, which specifies how we want our agents to relate to the hypotheses under consideration.

Definition 6. *Let $\bar{\alpha}$ be an agent, $\bar{h}$ be a hypothesis, and let B be the set of times t at which $\bar{\alpha}$ rejects $\bar{h}$. We say that $\bar{\alpha}$ covers $\bar{h}$ with test set M if either B is finite or the sequence $\left(l_T(\bar{\alpha},\bar{r},M,\bar{h})\right)_{T\in B}$ goes to negative infinity.*

4.4 The boundedly rational inductive agent criterion

We now state the BRIA criterion, the main contribution of this paper.

Definition 7. *Let $\bar{\alpha}$ be an agent for $\overline{\mathrm{DP}},\bar{r}$. Let $\mathbb{H} = \{h_1,h_2,...\}$ be a set of hypotheses. We say $\bar{\alpha}$ is a boundedly rational inductive agent (BRIA) for $\overline{\mathrm{DP}},\bar{r}$ covering $\mathbb{H}$ with test sets $M_1,M_2,...$ if $\bar{\alpha}$ does not overestimate and for all i, $\bar{\alpha}$ covers h_i with test set M_i.*

In the following, whenever $\bar{\alpha}$ is a BRIA, we will imagine that the test sets are given as a part of $\bar{\alpha}$. For example, if we say that $\bar{\alpha}$ is computable in, say, time polynomial in t, then we will take this to mean that $\bar{\alpha}$ together with a list at time t of tested hypotheses can be computed in polynomial time.

4.5 Examples

Betting on digits of π Consider the decision problem sequence with $\mathrm{DP}_t = \{a_t^\pi, x_t\}$ for all t, where a_t^π pays off the 2^t-th binary digit of π – i.e., r_t is the 2^t-th digit of π if $\alpha_t^c = a_t^\pi$ – and $x_t \in [0,1]$ pays off x_t. As usual we assume that the 2^t-th binary digits of π are pseudorandom (in a way we will make precise in Sect. 6) uniformly distributed (as they seem to be, cf. footnote 1). We would then expect boundedly rational agents to (learn to) choose a_t^π when $x_t < 1/2$ and choose x_t when $x_t > 1/2$.

We now consider an agent $\bar{\alpha}$ for this decision problem sequence. We will step-by-step impose the components of the BRIA criterion on $\bar{\alpha}$ to demonstrate their meaning and (joint) function in this example. We start by imposing the no overestimation criterion on $\bar{\alpha}$ without any assumptions about hypothesis coverage – what can we say about $\bar{\alpha}$ if we assume that does not overestimate? As noted earlier, the no overestimation criterion alone is weak and in particular does not constrain choice at all. For instance, $\bar{\alpha}$ might always choose $\alpha_t^c = a_t^\pi$ and alternate estimates of 0 and 1; or it might always choose x_t and estimate x_{t-1}.

We now impose instances of the hypothesis coverage criterion. We start with the hypothesis h_x which always recommends choosing x_t and promises a reward of x_t. Note that for all we know about the decision problem sequence this hypothesis does not give particularly good recommendations. However, in the context of our theory, h_x is useful because it always holds its promises. In particular, h_x's empirical record on any test set is 0. Hence, if α is to cover h_x, then α can only reject h_x finitely many times. By definition, this means that $\alpha_t^e \geq x_t$ for all but finitely many $t \in \mathbb{N}$. With the no overestimation criterion, it follows that α on average obtains utilities at least equal to x_t. But α's choices may still not match our bounded ideal. For example, α may always choose x_t.

Next, consider for $\varepsilon > 0$, the hypothesis h_π^ε that always recommends a_t^π and estimates $1/2 - \varepsilon$. Whether h_π^ε holds its promises is a more complicated question. But let us assume that $\bar{\alpha}$ covers h_π^ε with some test set M, and let us further assume that whether $t \in M$ is uncorrelated with the 2^t-th binary digit of π, for instance, because predicting the 2^t-th binary digit of π better than random cannot be done using the agent's computational capabilities. Then h_π^ε's empirical record on M will go to ∞, assuming that M is infinite – after all, following h_π^ε's recommendations yields a reward of $1/2$ on average, exceeding its promises of $1/2 - \varepsilon$. With the assumption that $\bar{\alpha}$ covers h_π^ε, it follows that for all but finitely many t, $\alpha_t^e \geq 1/2 - \varepsilon$. Now imagine that α not only covers one particular h_π^ε, but that there exist arbitrarily small positive ε such that α covers the hypothesis h_π^ε. Then it follows that in the limit as $t \to \infty$, $\alpha_t^e \geq 1/2$.

The above three conditions – no overestimation, coverage of h_x and coverage of h_π^ε for arbitrarily small ε – jointly imply that $\bar{\alpha}$ exhibits the desired behavior. Specifically, we have shown that $\bar{\alpha}$ must estimate at least $\max\{1/2, x_t\}$ in the limit. By the no overestimation criterion, $\bar{\alpha}$ also has to actually obtain at least $\max\{1/2, x_t\}$ on average. And if $\bar{\alpha}$ cannot guess the 2^t-th digits of π better than random, then the only way to achieve $\max\{1/2, x_t\}$ on average is to follow with limit frequency 1 the policy of choosing a_t^π when $x_t < 1/2$ and x_t when $x_t > 1/2$.

Adversarial offers Let α be an agent who faces a sequence of instances of SAO. In particular at time t, the agent faces $\mathrm{SAO}_{\alpha,t} = \{a_0, a_1\}$, where a_0 pays off $1/2$ with certainty. Intuitively, a_1 is evaluated to 1 if on the present problem α chooses a_0 and to 0 otherwise. Note, however, that the former fact is never relevant to computing r_t. So effectively $r_t = 1/2$ if $\alpha_t^c = a_0$ and $r_t = 0$ otherwise.

Assume that α does not overestimate and that it covers the hypothesis h which estimates $1/2$ and recommends a_0 in every round. Hypothesis h will always have an empirical record of 0 on any test set M since it holds its promises exactly. Hence, if α is to cover h, it can reject h only finitely many times. Thus, $\alpha_t^e \geq 1/2$ in all but finitely many rounds. To satisfy the no overestimation criterion, α must therefore obtain rewards of at least $1/2$ on average in the limit. Since a_1 pays off 0 whenever it is taken by α, it must be $\alpha_t^c = a_0$ with limit frequency 1.

5 Computing boundedly rational inductive agents

As described in Sect. 3, the goal of this paper is to formulate a rationality requirement that is not self-contradictory and that can be satisfied by computationally bounded agents. Therefore, we must show that one can actually construct BRIAs for given $\mathbb{H}$ and that under some assumptions about $\mathbb{H}$, such BRIAs are computable (within some asymptotic bounds).

Theorem 1. *Let $\mathbb{H}$ be a computably enumerable set consisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists an algorithm that computes a BRIA covering $\mathbb{H}$ (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$) for any $\overline{\mathrm{DP}}, \bar{r}$.*

We here give a sketch of our construction. For each decision problem, we run a first-price sealed-bid auction among the hypotheses. The highest-bidding hypothesis determines the agent's choice and estimate and is tested in this round. For each hypothesis, we maintain a wealth variable that tracks the hypothesis' empirical record. A hypothesis' bid is bound by its wealth. Thus, when a hypothesis outpromises the agent, this implies that the hypothesis' wealth is low. Upon winning an auction, the hypothesis pays its promise and gains the reward obtained after following the hypothesis' recommendation. We further distribute at each time t allowance to the hypotheses. The overall allowance per round is finite and goes to zero. The cumulative allowance for each hypothesis goes to ∞ over time. Thus, if a hypothesis is rejected infinitely often, then this requires the hypothesis to have spent all allowance and thus for its record among those rejection rounds to go to $-\infty$. Moreover, the cumulative overestimation is bound by overall allowance distributed and thus per-round overestimation goes to 0.

The next result shows that the BRIAs given by Theorem 1 are optimal in terms of complexity.

Theorem 2. *Let α be a BRIA for $\overline{\mathrm{DP}}, \bar{r}, \mathbb{H}$. Assume that there are infinitely many t such that $|\mathrm{DP}_t| \geq 2$ and $\alpha_t^e < 1$. If $\mathbb{H}$ is the set of $(O(g(t))$-)computable hypotheses, then α is not computable (in $O(g(t))$).*

6 Lower bounds on average rewards

Options with payoff guarantees Throughout this section, we will show that BRIAs satisfy many desiderata that one might have for rational decision makers. We start with a simple result which shows that if at each time t one of the options can be efficiently shown to have a value of at least L_t, then a BRIA will come to obtain at least L_t on average.

Theorem 3. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}}, \bar{r}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha_t^c = a_t \implies r_t \geq L_t$ for some e.c. sequence $\bar{L}$. We require also that the a_t are efficiently identifiable from the sets DP_t. Then in the limit as $T \to \infty$ it holds that $\sum_{t=1}^{T} r_t/T \geq \sum_{t=1}^{T} L_t/T$.*

The proof idea is simple. Consider the hypothesis that estimates L_t and recommends a_t if $t \in S$ and promises 0 otherwise. This hypothesis always keeps its promises. Hence, to cover this hypothesis, α can be outpromised by this hypothesis only finitely many times.

We can interpret Theorem 3 as providing an immunity to money extraction schemes, a widely discussed rationality condition. If a BRIA can leave with a certain payoff of L_t, it will on average leave with at least L_t. For example, in SAO of Sect. 3, a BRIA walks away with at least $1/2$, which in turn means that it chooses $a_0 =$ "$1/2$" with frequency 1.

Options with algorithmically random payoffs Theorem 8 only tells us something about *truly* random variables. But a key goal of our theory is to also be able to assign expected rewards to *algorithmically* random sequences, i.e., sequences that are deterministic, but relevantly unpredictable under computational constraints. We first offer a formal notion of algorithmic randomness.

Definition 8. *We say a sequence $\bar{y}$ is $(O(h(t))$ boundedly) van Mises–Wald–Church (vMWC) random with means $\bar{\mu}$ if for every infinite set $S \subseteq \mathbb{N}$ that is decidable (in $O(h(t))$ time) from available information, we have that* $\lim_{T\to\infty} \sum_{t \in S_{\leq T}} y_t - \mu_t = 0$.

Thus, we call a sequence random if there is no $(O(g(t))$-)computable way of selecting in advance members of the sequence whose average differs from the means $\bar{\mu}$. Definition 8 generalizes the standard definition of (unbounded) vMWC randomness (e.g. [8], Definition 7.4.1) to non-binary values with means $\bar{\mu}$ other than $1/2$ and computational constraints with outside input (e.g., from $\overline{\mathrm{DP}}$, which could contain options containing information such as, "by the way, the trillionth digit of π is 2"). The notion of vMWC randomness is generally considered quite weak (e.g. [8], Sect. 6.2).

Theorem 4. *Let $\bar{\mu}$ be an e.c. sequence on $[0,1]$. Let α be an $O(h(t))$-computable BRIA for decision problem sequence $\overline{\mathrm{DP}}$ with rewards $\bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the payoffs r_t in rounds with $\alpha_t^c = a_t$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^{T} r_t/T \geq \sum_{t=1}^{T} \mu_t/T$.*

We show an analogous result for Schnorr bounded randomness [26, 1, 33, 30] in Appendix E. Analogous results for truly random options follow from results for algorithmically random $\bar{r}$ and the fact that a sequence of truly random, independent numbers is algorithmically random almost surely. We give a direct proof in Appendix B.

7 Boundedly rational inductive agents as a foundation for game theory

We first recap basic game-theoretic concepts. A *(two-player) game* consists of two finite sets of *(pure) strategies* A_1,A_2, one set for each player, and two payoff functions $u_1,u_2\colon A_1 \times A_2 \to [0,1]$. A correlated strategy profile is a distribution $\mathbf{c} \in \Delta(A_1 \times A_2)$ over $A_1 \times A_2$. We can naturally extend utility functions to correlated strategy profiles as follows: $u_i(\mathbf{c}) = \sum_{\mathbf{a}\in A_1\times A_2} c_{\mathbf{a}} u_i(\mathbf{a})$. We call a correlated strategy profile $\mathbf{c}$ *strictly individually rational* if each player's payoff in $\mathbf{c}$ is greater than their pure strategy maximin payoff, i.e., $u_i(\mathbf{c}) > \max_{a_i\in A_i} \min_{a_{-i}\in A_{-i}} u_i(a_i, a_{-i})$.

Now imagine that two BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ learn to play a game against each other. That is, we consider BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ for $\bar{\mathrm{D}}\mathrm{P}^{\bar{\alpha}_1}, \bar{\mathrm{D}}\mathrm{P}^{\bar{\alpha}_2}$ respectively, where $\bar{\mathrm{D}}\mathrm{P}^{\bar{\alpha}_i} = A_i$ for $i = 1,2$ and $r_{i,t} = u_i(\alpha^c_{1,t}, \alpha^c_{1,t})$.

Theorem 5 (Folk theorem)**.** *Let Γ be a game. Let $\mathbb{H}_1, \mathbb{H}_2$ be any sets of hypotheses. Let $\mathbf{c} \in \Delta(A_1 \times A_2)$ be strictly individually rational. Then there exists $\mathbf{c}'$ arbitrarily close to $\mathbf{c}$ and BRIAs $\bar{\alpha}_1, \bar{\alpha}_2$ covering $\mathbb{H}_1, \mathbb{H}_2$ for decision problem sequences $\overline{\mathrm{DP}}^{\alpha_1}, \overline{\mathrm{DP}}^{\alpha_2}$ with rewards $\bar{r}_1, \bar{r}_2$ based on Γ as defined above s.t. the empirical distribution of (α^c_1, α^c_2) converges to $\mathbf{c}'$, i.e., for all $\mathbf{a} \in A_1 \times A_2$, $1/T \sum_{t=1}^{T} \mathbb{1}[(\alpha^c_1, \alpha^c_2) = \mathbf{a}] \to c'_{\mathbf{a}}$ as $T \to \infty$. Conversely, if α_1, α_2 are BRIAs for sets of hypotheses $\mathbb{H}_1$ and $\mathbb{H}_2$ that contain at least the constant-time deterministic hypotheses, $\sum_{t=1}^{T} u_i(\alpha^c_{1,t}, \alpha^c_{2,t})/T \geq \max_{a_i} \min_{a_{-i}} u_i(a_i, a_{-i})$ as $T \to \infty$. That is, in the limit each player receives at least their maximin utility.*

Theorem 5 is compelling, because it means BRIAs can learn to cooperate in one-shot games where rational agents would otherwise fail to cooperate (e.g., contrast fictitious play, or regret learning, both of which necessarily converge to defecting in the Prisoner's Dilemma). Note that our BRIA criterion is myopic, i.e., aimed at maximizing reward in the *current* round. Thus, even though the BRIAs in the above setting play repeatedly, the above result is unrelated to the folk theorems for repeated games.

8 Related work

Multi-armed bandit problems Our setting resembles a multi-armed bandit problem with expert advice (where $\mathbb{H}$ is the set of "experts"). The main difference is that we only define r_t, the reward actually obtained by the agent. The literature on multi-armed bandit problems assumes that the problem also defines the (counterfactual) rewards of untaken options and defines rationality in terms of these rewards. As discussed in Sect. 3, one of our motivations is to do away with these counterfactuals.

Within the multi-armed bandit literature, the most closely related strand of work is the literature on adversarial multi-armed bandit problems with expert advice ([2]; [13]). Like this paper, this literature addresses this problem of bounded rationality by formulating rationality relative to a set of hypotheses (the eponymous experts). However, its rationality criterion is very different from ours: they require regret minimization and in particular that cumulative regret is sublinear, a condition sometimes called Hannan-consistency. As the Simplified Adversarial Offer shows, Hannan-consistency is not achievable in our setting. However, it does become achievable if we assume that the agent has access to a source of random noise that is independent from $\overline{\mathrm{DP}}$ (see, e.g, the Exp4 algorithm of [2], Sect. 7). Importantly, the rationality criterion itself ignores the ability to randomize, i.e., it does not prescribe that the use of randomization be optimal in any sense.

We find it implausible to *require* rational agents to randomize to minimize regret; most importantly, regret minimization can require minimizing the rewards one actually obtains – see Appendix C.

Decision theory of Newcomb-like problems Problems in which the environment explicitly predicts the agent have been discussed as Newcomb-like problems by (philosophical) decision theorists ([16]). Most of this literature has focused on discussing relatively simple cases (similar to SAO). In these cases, BRIAs generally side with what has been called evidential decision theory. For example, by Theorem 3, BRIAs learn to one-box in Newcomb's problem. Of course, BRIAs differ structurally from how a decision theorist would usually conceive of an evidential decision theory-based agent. E.g., BRIAs are not based on expected utility maximization (though they implement it when feasible; see Appendix B). We also note that the decision theory literature has, to our knowledge, not produced any formal account of how to assign the required conditional probabilities in Newcomb-like problems.

Bounded rationality The motivations of the present work as per Sect. 3, especially Sect. 3, coincide with some of the motivations for the study of bounded rationality. However, other motivations have been given for the study of bounded rationality as well (see, e.g., [27], Sect. 2). More importantly, since much of bounded rationality is geared towards explaining or prescribing *human* (as opposed to AI) behavior, the characterization and analysis of "computational capacities" often differ from ours (e.g. [5]). For instance, for most humans dividing 1 by 17 is a challenge, while such calculation are trivial for computers. A few authors have also explicitly connected the general motivations of bounded rationality with paradoxes of self reference and game theory as discussed in Sect. 3 ([4], [20, Ch. 10]). Anyway, the literature on bounded rationality is vast and diverse. Much of it is so different from the present work that a comparison hardly makes sense. Below we discuss a few approaches in this literature that somewhat resemble ours. In particular, like the present paper (and Hannan consistency) they specify rationality relative to a given set of hypotheses (that in turn is defined by computational constraints).

Russell et al.'s bounded optimality Like our approach and the other approaches discussed in this related work section, Russell et al. define *bounded optimality* as a criterion relative to a set of (computationally bounded) hypotheses called *agent programs* ([21], Sect. 1.4; [23]; [22]). Roughly, an agent program is boundedly optimal if it is the optimal program from some set of bounded programs. The main difference between our and Russell et al.'s approach is that we address the problems of Sect. 3 by developing a theory of learning to make such decisions, while Russell et al. address them by moving the decision problem one level up, from the agent to the design of the agent (cf. [7], Sect. 2.2 for a discussion of this move). As one consequence, we can design general BRIAs, while it is in general hard to design boundedly optimal agents. Of course, the feasibility of designing BRIAs comes at the cost of our agents only behaving reasonably in the limit. Moreover, the designer of boundedly optimal agents as per Russell et al. may become a subject of the paradoxes of Sect. 3 in problematic ways.

Garrabrant inductors The present is in part inspired by the work of Garrabrant et al. [10], who address the problem of assigning probabilities under computational constraints and possibilities of self-reference. As an alternative to the present theory of BRIAs, one could also try to develop a theory of boundedly rational choice by maximizing expected utility using the Garrabrant inductor's probability distributions. Unfortunately, this approach fails for reasons related to the challenge of making counterfactual claims, as pointed out by Garrabrant [9]. As in the case of Hannan consistency, we can address this problem using randomization over actions. However, like Garrabrant (ibid.), we do not find it satisfactory to *require* randomization (cf. again Appendix C). We conjecture that Garrabrant inductors with (pseudo-)randomization could be used to construct BRIAs.

9 Conclusion

We developed BRIA theory as a theory of bounded inductive rationality. We gave results that show the normative appeal of BRIAs. Furthermore, we demonstrated the theory's utility by using it to justify Nash equilibrium play. At the same time, the ideas presented lead to various further research questions, some of which we have noted above. We here give three more that we find particularly interesting. Can we modify the BRIA requirement so that it implies coherence properties à la Garrabrant et al. [10]? Do the frequencies with which BRIAs play the given pure strategies of a game converge to mixed Nash and correlated equilibria? Can BRIA theory be used to build better real-world systems?

Acknowledgments

We thank Emery Cooper, Daniel Demski, Sam Eisenstat, Daniel Kokotajlo, Alexander Oldenziel, Nisan Stiennon, Johannes Treutlein and attendants of OptLearnMAS 2021 for helpful discussions.

References

[1] K. Ambos-Spies, S.A. Terwijn & Z. Xizhong (1997): *Resource bounded randomness and weakly complete problems*. *TCS* 172(1–2), pp. 195–207, doi:10.1016/S0304-3975(95)00260-X.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund & R.E. Schapire (2002): *The nonstochastic multiarmed bandit problem*. *SIAM journal on computing* 32(1), pp. 48–77, doi:10.1137/S0097539701398375.

[3] M. Barasz, P. Christiano, B. Fallenstein, M. Herreshoff, P. LaVictoire & E. Yudkowsky (2014): *Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic*. Available at https://arxiv.org/abs/1401.5577.

[4] K. Binmore (1987): *Modeling Rational Players: Part I*. *Economics & Philosophy* 3(2), pp. 179–214, doi:10.1017/S0266267100002893.

[5] J. Conlisk (1996): *Why bounded rationality?* *Journal of Economic Literature* 34(2), pp. 669–700.

[6] A. Critch (2016): *Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents*. Available at https://arxiv.org/abs/1602.04184.

[7] A. Demski & S. Garrabrant (2020): *Embedded Agency*. Available at arxiv.org/pdf/1902.09469.pdf.

[8] R.G. Downey & D.R. Hirschfeldt (2010): *Algorithmic randomness and complexity*. Springer, doi:10.1007/978-0-387-68441-3.

[9] S. Garrabrant (2017): *Two Major Obstacles for Logical Inductor Decision Theory*. Available at alignmentforum.org/posts/5bd75cc58225bf06703753d4/.

[10] S. Garrabrant, T. Benson-Tilsen, A. Critch, N. Soares & J. Taylor (2017): *A Formal Approach to the Problem of Logical Non-Omniscience*. In J. Lang, editor: *Proceedings of the 16th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2017)*, 251, EPTCS, pp. 221–235, doi:10.4204/EPTCS.251.16. Available at https://arxiv.org/abs/1707.08747.

[11] D. Hofstadter (1983): *Dilemmas for Superrational Thinkers*. *Scientific American* 248(6).

[12] R.C. Jeffrey (1965): *The Logic of Decision*. McGraw-Hill, New York.

[13] T. Lattimore & C. Szepesvari (2020): *Bandit Algorithms*. To be published by Cambridge University Press, doi:10.1017/9781108571401. Available at tor-lattimore.com/downloads/book/book.pdf.

[14] G. Marsaglia (2005): *On the Randomness of Pi and Other Decimal Expansions*. *InterStat*.

[15] J. v. Neumann (1928): *Zur Theorie der Gesellschaftsspiele*. *Mathematische Annalen* 100(1), pp. 295–320, doi:10.1007/BF01448847.

[16] R. Nozick (1969): *Newcomb's Problem and Two Principles of Choice*. In N.R. et al., editor: *Essays in Honor of Carl G. Hempel*, Springer, pp. 114–146, doi:10.1007/978-94-017-1466-2_7.

[17] C. Oesterheld (2019): *Robust Program Equilibrium*. *Theory and Decision* 86(1), pp. 143–159, doi:10.1007/s11238-018-9679-3.

[18] C. Oesterheld & V. Conitzer (2021): *Extracting Money from Causal Decision Theorists*. *Phil Quarterly* 71(4), doi:10.1093/pq/pqaa086.

[19] M. Peterson (2009): *An Introduction to Decision Theory*. Cambridge University Press, doi:10.1017/CBO9780511800917.

[20] A. Rubinstein (1998): *Modeling Bounded Rationality*. Zeuthen Lecture Book Series, The MIT Press, doi:10.7551/mitpress/4702.001.0001.

[21] S. Russell & E. Wefald (1991): *Do the Right Thing – Studies in Limited Rationality*. The MIT Press.

[22] S.J. Russell & D. Subramanian (1995): *Provably Bounded-Optimal Agents*. *JAIR* 2, pp. 575–609, doi:10.1613/jair.133.

[23] S.J. Russell, D. Subramanian & R. Parr (1993): *Provably bounded optimal agents*. In: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*, pp. 338–344.

[24] L.J. Savage (1954): *The Foundations of Statistics*. John Wiley and Sons, New York.

[25] L.J. Savage (1967): *Difficulties in the Theory of Personal Probability*. *Philosophy of Science* 34(4), pp. 305–310, doi:10.1086/288168.

[26] C.P. Schnorr (1971): *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. *Lecture Notes in Mathematics* 218, Springer, doi:10.1007/BFb0112468.

[27] R. Selten (1990): *Bounded Rationality*. *Journal of Institutional and Theoretical Economics* 146(4).

[28] J. Spencer (2021): *An argument against causal decision theory*. *Analysis* 81(1), pp. 52–61, doi:10.1093/analys/anaa037.

[29] K. Steele & H.O. Stefánsson (2016): *Decision Theory*. In E.N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, winter 2016 edition.

[30] D.M. Stull (2020): *Resource bounded randomness and its applications*. In: *Algorithmic Randomness – Progress and Prospects*, Lecture Notes in Logic, Cambridge University Press, pp. 301–348, doi:10.1017/9781108781718.010.

[31] M. Tennenholtz (2004): *Program equilibrium*. *Games and Economic Behavior* 49(2), pp. 363–373, doi:10.1016/j.geb.2004.02.002.

[32] W. van der Hoek, C. Witteveen & M. Wooldridge (2013): *Program equilibrium – a program reasoning approach*. *International Journal of Game Theory* 42, pp. 639–671, doi:10.1007/s00182-011-0314-6.

[33] Y. Wang (2000): *Resource bounded randomness and computational complexity*. *TCS* 237(1–2), pp. 33–55, doi:10.1016/S0304-3975(98)00119-4.

A Proofs

A.1 An easy lemma about test sets

We start with a simple lemma which we will use to simplify a few of our proofs. Roughly, the lemma shows that to cover a hypothesis h, it never helps to test h in rounds in which $h_t = 0$, i.e., in rounds in which h doesn't make any promises.

Lemma 6. *Let $\bar{h}$ be a hypothesis and $N \subseteq \mathbb{N}$ s.t. $t \in N$ implies $h_t^e = 0$. Then if $\bar{\alpha}$ covers $\bar{h}$ with test set M, $\bar{\alpha}$ covers $\bar{h}$ with test set $M - N$.*

Proof. For all T, we have that

$$\begin{aligned} l_T(\bar{\alpha},\bar{r},M,\bar{h}) = \sum_{t\in M_{\leq T}} r_t - h^e_t &= \sum_{t\in M_{\leq T}-N} r_t - h^e_t + \sum_{t\in M_{\leq T}\cap N} r_t - h^e_t \\ &= \sum_{t\in M_{\leq T}-N} r_t - h^e_t + \sum_{t\in M_{\leq T}\cap N} r_t \\ &\geq \sum_{t\in M_{\leq T}-N} r_t - h^e_t \\ &= l_t(\bar{\alpha},\bar{r},M-N,\bar{h}). \end{aligned}$$

Thus, if $l_T(\bar{\alpha},\bar{r},M,\bar{h}) \to -\infty$ as $T \to -\infty$, it must also be $l_T(\bar{\alpha},\bar{r},M-N,\bar{h}) \to -\infty$ as $T \to -\infty$. □

A.2 Proof of Theorem 1

Theorem 1. *Let $\mathbb{H}$ be a computably enumerable set consisting of ($O(g(t))$-)computable hypotheses. (Let $g \in \Omega(\log)$.) Then there exists an algorithm that computes a BRIA covering $\mathbb{H}$ (in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$) for any $\overline{\mathrm{DP}},\bar{r}$.*

Proof. Our proof is divided into four parts. First, we give the generic construction for a BRIA (1). Then we show that this is indeed a BRIA by proving that it satisfies the no overestimation criterion (2), as well as the coverage criterion (3). Finally, we show that under the assumptions stated in the theorem, this BRIA is computable in the claimed time complexity (4).

1. The construction

First, we need an *allowance function* $A : \mathbb{N}\times\mathbb{N} \to \mathbb{R}_{\geq 0}$, which for each time n, specifies a positive amount $A(n,i)$ given to hypothesis h_i's wealth at time n. The allowance function must satisfy the following requirements:

- Each hypothesis must get infinite overall allowance, i.e., $\sum_{n=1}^{\infty} A(n,i) = \infty$ for all hypotheses h_i.
- The overall allowance distributed per round n must go to zero, i.e.,

$$\sum_{n=1}^{N} \frac{1}{N} \sum_{i=1}^{\infty} A(n,i) \underset{N\to\infty}{\to} 0. \tag{1}$$

  In particular, the allowance distributed in any particular round must be finite.

An example of such a function is $A(n,i) = n^{-1}i^{-2}$.

We can finally give the algorithm itself. Initialize the wealth variables as (for example) $w_0(i) \leftarrow 0$ for each hypothesis $h_i \in \mathbb{H}$.

At time t, we run a (first-price sealed-bid[2]) auction for the present decision problem among all hypotheses. That is, we determine a winning hypothesis

$$i^*_t \in \arg\max_{i\in\mathbb{N}} \min(h^e_{i,t}, w_t(i)) \tag{2}$$

with arbitrary tie breaking. Intuitively, each hypothesis h_i bids $h^e_{i,t}$, except that it is constrained by its wealth $w_t(i)$. The idea is that if h_i has performed poorly relative to its promises, then α should not trust

[2] This format is mainly chosen for its simplicity. We could just as well use a second-price (or third-price, etc.) auction. We could use even different formats to get somewhat different BRIA-like properties. For instance, with combinatorial auctions, one could achieve cross-decision optimization.

h_i's promise for the present problem. Let $e_t^* \in [0,1]$ be the maximum (wealth-bounded) bid itself. We then define our agent at time t as $\alpha_t := (h^c_{i_t^*,t}, e_t^*)$.

We update the wealth variables as follows. For all hypotheses $i \neq i_t^*$, we merely give allowance, i.e., $w_{t+1}(i) \leftarrow w_t(i) + A(t,i)$. For the winning hypothesis i_t^*, we update wealth according to $w_{t+1}(i_t^*) \leftarrow w_t(i_t^*) + A(t,i_t^*) + r_t - e_t^*$. That is, the highest-bidding hypothesis receives the allowance and the reward obtained after following its recommendation (r_t), but pays its (wealth-bounded) bid (e_t^*).

2. No overestimation We will show that the cumulative overestimation is bounded by the sum of the allowance.

For each T, let B_T^+ be the set of hypotheses whose wealth $w_t(i)$ is positive for at least one time $t \in \{0,...,T\}$. Note that all highest-bidding hypotheses in rounds $1....,T$ are in B_T^+ for all j. We can then write the overall wealth of the hypotheses in B_T^+ at time T as

$$\sum_{i\in B_T^+} w_T(i) = \sum_{i\in B_T^+} \sum_{n=1}^{T} A(n,i) + \sum_{t=1}^{T} r_t - \alpha_t^e.$$

That is, the overall wealth at time T is the allowance distributed at times $1,...,T$ plus the money earned/lost by the highest-bidding hypotheses.

Now notice that by the construction above, if a wealth variable $w_t(i)$ is non-negative once, it remains non-negative for all future t. Thus, for all $i \in B_T^+$, $w_T(i) \geq 0$. Second, the last term is the negated cumulative overestimation of $\bar{\alpha}$. Thus, re-arranging these terms and dividing by T gives us the following upper bound on the per-round overestimation:

$$\frac{1}{T}\mathscr{L}_T(\alpha,\bar{r}) = \frac{1}{T}\left(\sum_{i\in B_T^+}\sum_{n=1}^{T} A(n,i) - \sum_{i\in B_T^+} w_T(i)\right) \leq \frac{1}{T}\sum_{i\in B_T^+}\sum_{n=1}^{T} A(n,i) \leq \sum_{i=1}^{\infty}\frac{1}{T}\sum_{n=1}^{T} A(n,i),$$

which goes to zero as $T \to \infty$ by our requirement on the function A (line 1).

3. Hypothesis coverage Given a hypothesis h_i that strictly outpromises $\bar{\alpha}$ infinitely often, we use as a test M_i, the set of times t at which h_i is the winning hypothesis (i.e., the set of times t s.t. $i = i_t^*$). We have to show that M_i is infinite, is a valid test set (as per Definition 4), and that it satisfies the justified rejection requirement in the hypothesis coverage criterion.

A) We show that M_i is infinite. That is, we need to show that infinitely often h_i is the highest-bidding hypothesis in the auction that computes $\bar{\alpha}$. Assume for contradiction that M_i is finite. We will show that at some point h_i's bidding in the construction of $\bar{\alpha}$ will not be constrained anymore by h's wealth. We will then find a contradiction with the assumption that h_i strictly outpromises α infinitely often.

Consider that for $T' > T$, it is $w_{T'}(i) = w_T(i) + \sum_{t=T+1}^{T'} A(t,i)$. That is, from time T to any time T', hypothesis i's wealth only changes by h_i receiving allowance, because i is (by assumption) not the winning hypothesis i_t^* in any round $t \geq T$. Because we required $\sum_{n=1}^{\infty} A(n,i) = \infty$, we can select a time $T* \geq T$ such that $w_{T*}(i) \geq 1$. Note that again it is also for all $t > T*$ the case that $w_t(i) \geq 1$.

We now see that if $t \geq T*$ the wealth constraints is not restrictive. That is, for all such t it is $\min(h^e_{i,t}, w_t(i)) = h^e_{i,t}$. But it is infinitely often $h^e_{i,t} > \alpha_t^e$. This contradicts the fact that by construction, α_t is equal to the highest wealth-restricted hypothesis.

B) The fact that M_i is a valid test set follows immediately from the construction – α always chooses the recommendation of the highest-bidding hypothesis.

C) We come to the justification part of the coverage criterion. Let B_i be the set of rounds in which $\bar{h}_i$ strictly outpromises $\bar{\alpha}$.

At each time $t \in B_i$, by construction $w_T(i,j) < h^e_{i,t}(\mathrm{DP}_T)$. We have that $h^e_{i,t}(\mathrm{DP}_T) \leq 1$ and

$$w_T(i) = \sum_{n=1}^{T} A(n,i) + \sum_{t \in M_i : t<T} r_t - h^e_{i,t}.$$

Hence, from the fact that $w_T(i) < h^e_{i,t}(\mathrm{DP}_T)$ for all $T \in B_i$, it follows that for all $T \in B_i$, it is

$$\sum_{t \in M_i : t<T} h^e_{i,t} - r_t > \sum_{n=1}^{T} A(n,i),$$

which goes to infinity as $T \to \infty$, as required.

4. Computability and computational complexity It is left to show that if $\mathbb{H}$ can be computably enumerated and consist only of $(O(g(t))$-)computable hypotheses, then we can implement the above-described BRIA for $\mathbb{H}$, $\overline{\mathrm{DP}}$,$\bar{r}$ in an algorithm (that runs in $O(g(t)q(t))$, for arbitrarily slow-growing, $O(g(t))$-computable q with $q(t) \to \infty$).

The main challenge is that the construction as described above performs at any time t, operations for all (potentially infinitely many) hypotheses. The crucial idea is that for an appropriate choice of A, we only need to keep track of a finite set of hypotheses, when calculating $\bar{\alpha}$ in the first T time steps. Each hypothesis starts with an initial wealth of 0. Then a hypothesis i can only become relevant at the first time t at which $A(t,i) > 0$. At any time t, we call such hypotheses *active*. Before that time, we do not need to compute $\bar{h}_i$ and do not need to update its wealth. By choosing a function A s.t. (in addition to the above conditions) $A(t,\cdot)$ has finite, e.c. support at each time t, we can keep the set of active hypotheses finite at any given time. (An example of such a function is $A(n,i) = n^{-1}i^{-2}$ for $i < n$ and $A(n,i) = 0$ otherwise.) We have thus shown that it is enough to keep track at any given time of only a finite number of hypotheses.

At any time, we therefore only need to keep track of a finite number of wealth variables, only need to compute the recommendations and promises of a finite set of hypotheses, and only need to compute a minimum of a finite set in line 2.

Computability is therefore proven. We proceed to show the claim about computational complexity. At any time t, let $C_{\max}(t)$ be the largest constant by which the computational complexity of hypotheses at time t are bounded relative to $g(t)$. Further, let $h_b(t)$ be the set of active hypotheses. Then the computational cost from simulating all active hypotheses at time t is at most $h_b(t)C_{\max}(t)g(t)$. All of $C_{\max}(t)$ and $h_b(t)$ must go to ∞ as $t \to \infty$. However, this can happen arbitrarily slowly, up to the limits of fast $(O(g(t)))$ computation. Hence, if we let $q(t) = h_b(t)C_{\max}(t)g(t)$, we can let q grow arbitrarily slowly (again, up to the limits of fast computation).

Finally, we have to verify that all other calculations can be done in $O(q(t)g(t))$: To determine the winning hypothesis given everyone's promises, we have to calculate the maximum of $h_b(t) \in O(q(t))$ numbers, which can be done in $O(q(t))$ time. We also need to conduct the wealth variable updates themselves, which accounts for $O(h_b(t))$ additions. Again, this is in $O(g(t)q(t))$. And so on. □

A.3 Proof of Theorem 2

Theorem 2. *Let α be a BRIA for $\overline{\mathrm{DP}}, \bar{r}, \mathbb{H}$. Assume that there are infinitely many t such that $|\mathrm{DP}_t| \geq 2$ and $\alpha^e_t < 1$. If $\mathbb{H}$ is the set of $(O(g(t))$-)computable hypotheses, then α is not computable (in $O(g(t))$).*

This is shown by a simple diagonalization argument. If a BRIA α were computable (in $O(g(t))$), then consider the hypothesis who in rounds in which $|\mathrm{DP}_t| \geq 2$ and $\alpha^e_t < 1$, promises 1 and recommends an option other than α^c_t; and promises 0 otherwise. This hypothesis strictly outpromises α infinitely often, is computable (in $O(g(t))$) but is never tested .

A.4 Proof of Theorem 3

Theorem 3. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}}, \bar{r}$ and the set of e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. for all $t \in \mathbb{N}$, it holds that $a_t \in \mathrm{DP}_t$ and $\alpha^c_t = a_t \implies r_t \geq L_t$ for some e.c. sequence $\bar{L}$. We require also that the a_t are efficiently identifiable from the sets DP_t. Then in the limit as $T \to \infty$ it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T L_t/T$.*

Proof. We will show that if the assumptions are satisfied, then for all but finitely many t, we have that $\alpha^e_t \geq L_t$. From this and the fact that $\bar{\alpha}$ doesn't overestimate, it then follows that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T L_t/T$.

We prove this new claim by proving a contrapositive. In particular, we assume that $\alpha^e_t < L_t$ for infinitely many t and will then show that $\bar{\alpha}$ is not a BRIA (using the other assumptions of the theorem).

Consider hypothesis $\bar{h}_i$ such that $h_{i,t} = (a_t, L_t)$. Because $\bar{L}$ is e.c. and the $\bar{a}$ are efficiently identifiable, $\bar{h}$ is e.c. We now show that $\bar{h}_i$ is not covered by $\bar{\alpha}$, which shows that $\bar{\alpha}$ is not a BRIA. By assumption, $\bar{h}_i$ strictly outpromises $\bar{\alpha}$ infinitely often. It is left to show that there is no M_i as specified in the hypothesis coverage criterion, i.e. no M_i on which $\bar{h}_i$ consistently underperforms its promises.

If $t \in M_i$, then $\alpha^c_t = h^c_{i,t} = a_t$ and therefore $r_t \geq L_t$. It follows that for all T,

$$l_T(\bar{\alpha}, \bar{r}, M_i, \bar{h}_i) = \sum_{t \in M_i : t < T} \underbrace{r_t}_{\geq L_t} - \underbrace{h^e_{i,t}}_{=L_t} \geq 0.$$

Thus, $\bar{\alpha}$ violates the coverage criterion for $\bar{h}_i$. □

A.5 Proof of Theorem 4

Definition 8. *We say a sequence $\bar{y}$ is ($O(h(t))$ boundedly) van Mises–Wald–Church (vMWC) random with means $\bar{\mu}$ if for every infinite set $S \subseteq \mathbb{N}$ that is decidable (in $O(h(t))$ time) from available information, we have that $\lim_{T\to\infty} \sum_{t \in S_{\leq T}} y_t - \mu_t = 0$.*

Theorem 4. *Let $\bar{\mu}$ be an e.c. sequence on $[0,1]$. Let α be an $O(h(t))$-computable BRIA for decision problem sequence $\overline{\mathrm{DP}}$ with rewards $\bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the payoffs r_t in rounds with $\alpha^c_t = a_t$ are $O(h(t))$-boundedly vMWC random with means $\bar{\mu}$. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T r_t/T \geq \sum_{t=1}^T \mu_t/T$.*

Proof. We prove the theorem by proving that for all $\varepsilon > 0$, $\alpha^e_t \geq \mu_t - \varepsilon$ for all but finitely many t. As usual, we prove this by proving the following contrapositive: assuming this is not the case, $\bar{\alpha}$ is not a BRIA. To prove this, consider hypothesis $\bar{h}_{a,\varepsilon}$ that at each time t promises $\max(\mu_t - \varepsilon, 0)$ and recommends a_t. Since $\bar{h}_{a,\varepsilon}$ infinitely often outpromises $\bar{\alpha}$, it must tested infinitely often. Let the test set be some infinite set $M \subseteq \mathbb{N}$. By Lemma 6, we can assume WLOG that for all $t \in M$, $h^e_{a,\varepsilon} = \mu_t - \varepsilon$.

Now notice that M is by assumption computable in $O(h(t))$ given the information available at time t. Now

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{r}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon) \underset{\text{w.p. } 1}{\rightarrow} \varepsilon \text{ as } T \to \infty,$$

where the final step is by the fact that among rounds where $\alpha_t^c = a_t$, $\bar{r}$ is vMWC random with means $\bar{\mu}$. Hence, $\bar{h}_{a,\varepsilon}$'s record $l_T(\alpha,\bar{r},M_i,\bar{h}_i)$ must be positive in all but finitely many rounds. Thus, $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\varepsilon}$ violate the coverage criterion. □

A.6 Proof of Theorem 5

Lemma 7 (Minimax Theorem [15]). *Let (A_1,A_2,u_1,u_2) be any game. Then*

$$\max_{\sigma_i\in\Delta(A_i)}\min_{a_{-i}\in A_{-i}} u_i(\sigma_i,\sigma_{-i}) = \min_{\sigma_{-i}\in\Delta(A_{-i})}\max_{a_i\in A_i} u_i(\sigma_i,\sigma_{-i}).$$

Theorem 5 (Folk theorem). *Let Γ be a game. Let $\mathbb{H}_1,\mathbb{H}_2$ be any sets of hypotheses. Let $\mathbf{c}\in\Delta(A_1\times A_2)$ be strictly individually rational. Then there exists $\mathbf{c}'$ arbitrarily close to $\mathbf{c}$ and BRIAs $\bar{\alpha}_1,\bar{\alpha}_2$ covering $\mathbb{H}_1,\mathbb{H}_2$ for decision problem sequences $\overline{\mathrm{DP}}^{\alpha_1},\overline{\mathrm{DP}}^{\alpha_2}$ with rewards $\bar{r}_1,\bar{r}_2$ based on Γ as defined above s.t. the empirical distribution of (α_1^c,α_2^c) converges to $\mathbf{c}'$, i.e., for all $\mathbf{a}\in A_1\times A_2$, $^1/_T\sum_{t=1}^T \mathbb{1}[(\alpha_1^c,\alpha_2^c)=\mathbf{a}]\to c'_{\mathbf{a}}$ as $T\to\infty$. Conversely, if α_1,α_2 are BRIAs for sets of hypotheses $\mathbb{H}_1$ and $\mathbb{H}_2$ that contain at least the constant-time deterministic hypotheses, $\sum_{t=1}^T u_i(\alpha_{1,t}^c,\alpha_{2,t}^c)/T \geq \max_{a_i}\min_{a_{-i}} u_i(a_i,a_{-i})$ as $T\to\infty$. That is, in the limit each player receives at least their maximin utility.*

Proof. The latter part ("Conversely,...") follows directly from Theorem 3. It is left to prove the existence claim.

We construct the BRIAs as follows. First we fix positive probabilities $p_{\mathbf{c}}\in(0,1)$ and $(p_{a_i})_{a_i\in A_i}$ for $i=1,2$ (WLOG assume A_1 and A_2 are disjoint) s.t. $p_{\mathbf{c}}+\sum_{i=1}^2\sum_{a_i\in A_i}p_{a_i}=1$. Further let v_i be some number that is strictly greater than Player i's maximin value but strictly smaller than $p_c u_i(\mathbf{c})$. By the assumption that $\mathbf{c}$ is strictly individually rational, such a number exists if we make p_c large enough. Then let $\alpha_{i,t}^e = v_i$ for all t. Then in each step the BRIAs jointly randomize[3] independently from all bidders in $\mathbb{H}_1,\mathbb{H}_2$ as follows:

- With probability $p_{\mathbf{c}}$ both players play according to $\mathbf{c}$ by jointly implementing $\mathbf{c}$, e.g., by deterministically cycling through the different strategies in the appropriate numbers.Further, $\alpha_{i,t}^e = v_i$. No hypotheses are tested.
- With probability p_{a_i}, Player i plays a_i and Player $-i$ plays from $\arg\min_{a_{-i}\in A_{-i}} u_i(a_i,a_{-i})$. Player $-i$ estimates v_{-i} and does not test any hypothesis. Player i estimates v_i and tests every hypothesis that estimates more than v_i.

We now prove that $\bar{\alpha}_1,\bar{\alpha}_2$ thus constructed are BRIAs.

No overestimation:

$$\mathscr{L}_T(\bar{\alpha}_i,\bar{r}_i)/T = \sum_{t=1}^T(\alpha_{i,t}^e - r_{i,t})/T = \sum_{t=1}^T(v_i - r_{i,t})/T \leq v_i - u_i(\mathbf{c}) \text{ as } T\to\infty.$$

By construction, $v_i - u_i(\mathbf{c})\leq 0$.

Coverage: Let $\bar{h}_i$ be a hypothesis that outbids $\bar{\alpha}_i$ infinitely often. Then in particular $\bar{h}_i$ outbids infinitely often in rounds in which $\bar{h}_i$ recommends some a_i and $\alpha_{i,t}^c = a_i$. Thus, $\bar{h}_i$ has an infinite test set M on which the hypothesis' empirical record is

$$l_T(\bar{\alpha}_i,\bar{r}_i,M,\bar{h}_i) = \sum_{t\in M_{\leq T}} r_t - h_{i,t}^e = \sum_{t\in M_{\leq T}}\min_{a_{-i}} u_i(h_{i,t}^c,a_{-i}) - h_{i,t}^e \leq \sum_{t\in M_{\leq T}}\max_{a_i}\min_{a_{-i}} u_i(a_i,a_{-i}) - v_i \to -\infty$$

as $T\to\infty$. Thus, $\bar{h}_i$ is covered. □

[3] We here use true randomization for simplicity. The same can be achieved using algorithmic randomness.

B Options with random payoffs

The following result shows, roughly, that when choosing between different lotteries whose expected utilities are efficiently computable, BRIAs converge to choosing the lottery with the highest expected utility. When other, non-lottery options are available, BRIAs converge to performing at least as well as the best lottery option.

Theorem 8. *Let $\bar{\alpha}$ be a BRIA for $\overline{\mathrm{DP}},\bar{r}$. Let $\bar{a}$ be a sequence of terms in $\mathcal{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values of r_t if $\alpha_t^c = a_t$ are drawn independently from distributions with e.c. means $\bar{\mu}$. Let the a_t be efficiently identifiable from DP_t. Then almost surely in the limit as $T \to \infty$, it holds that $\sum_{t=1}^{T} r_r/T \geq \sum_{t=1}^{T} \mu_t/T$.*

The proof idea similar to the proof idea for Theorem 3. It works by considering hypotheses that recommend a_t and promise $\mu_t - \varepsilon$ and noting that the empirical record of such hypotheses goes to $-\infty$ with probability 0.

Proof. We need only show that with probability 1 for all $\varepsilon > 0$ it holds that for all but finitely many times t that $\alpha_t^e \geq \mu_t - \varepsilon$. From this and the no overestimation property of $\bar{\alpha}$, the conclusion of the theorem follow.

Again we prove the following contrapositive: If there is some $\varepsilon > 0$ s.t. with some positive probability $p > 0$ we infinitely often have that $\alpha_t^e < \mu_t - \varepsilon$, then $\bar{\alpha}$ is with positive probability not a BRIA.

Consider the hypothesis $\bar{h}_{a,\varepsilon}$ that at each time t promises $\max(\mu_t - \varepsilon, 0)$ and recommends a_t. Since with probability p, $\bar{h}_{a,\varepsilon}$ infinitely often outpromises $\bar{\alpha}$, it must in these cases (and therefore with probability (at least) p) be tested infinitely often. (If not, we $\bar{\alpha}$ would in these cases not be a BRIA and we would be done.) In these cases (i.e., when $\bar{h}_{a,\varepsilon}$ is tested infinitely often), let the test set be some infinite set $M \subseteq \mathbb{N}$. (Note that M may depend on $\bar{r}$ and inherit its stochasticity. This will not matter for the following, though.) For simplicity, let M be the empty set if $\bar{h}_{a,\varepsilon}$ does not outpromise α infinitely often. By Lemma 6, we can assume WLOG that for all $t \in M$, $h_{a,\varepsilon}^e = \mu_t - \varepsilon$. Now notice that

$$\frac{1}{|M_{i,\leq T}|} l_T(\alpha, \bar{r}, M_i, \bar{h}_i) = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - h_{a,\varepsilon,t}^e = \frac{1}{|M_{i,\leq T}|} \sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon).$$

Conditioning on the (probability p) event that h infinitely often outbids and therefore that M is infinite, it must then with probability 1 be the case that $\sum_{t \in M_{i,\leq T}} r_t - (\mu_t - \varepsilon) \underset{\text{w.p. } 1}{\rightarrow} \varepsilon$ as $T \to \infty$ by the law of large numbers. We have thus shown that with positive probability (p) $\bar{h}_{a,\varepsilon}$ outpromises $\bar{\alpha}$ infinitely often while $\bar{h}_{a,\varepsilon}$'s record $l_T(\alpha, \bar{r}, M_i, \bar{h}_i)$ is positive in all but finitely many rounds. Thus, in this positive-probability event $\bar{\alpha}$'s infinitely many rejections of $\bar{h}_{a,\varepsilon}$ violates the coverage criterion. □

C More on randomization and regret

In the literature on multi-armed bandit problems, authors usually consider the goal of regret minimization. A natural rationality requirement is for per-round average regret to go to 0. This is sometimes called Hannan consistency. For any given agent c, the Simplified Adversarial Offer SAO_c of Sect. 3 is a problem on which regret is necessarily high. However, if we assume that the agent at time t can randomize in a way that is independent of how the rewards are assigned by D_t, it can actually be ensured that per-round regret (relative to any particular hypothesis) goes to 0 (see Sect. 8).

Arguably the assumption that the agent can independently randomize is almost always satisfied for artificial agents in practice. For instance, if an agent wanted to randomize independently, then for an adversary to predict the program's choices, it would not only need to know the program's source code. It would also require (exact) knowledge of the machine state (as used by pseudo-random number generators); as well as the exact content of any stochastic input such as video streams and hardware/true random number generators. Independent randomization might not be realistic for humans (to whom randomization requires some effort), but none of these theories under discussion (the present one, regret minimization, full Bayesian updating, etc.) are directly applicable to humans, anyway.

Nevertheless, we are conceptually bothered by the assumption of independent randomization. It seems desirable for a theory of choice to make as few assumptions as possible about the given decision problems. Moreover, we can imagine situations in which independent randomization is unavailable to a given agent. It seems odd for a theory of learning to be contingent on the fact that such situations are (currently) rare or practically insignificant. A detailed discussion of this philosophical concern is beyond the scope of this paper.

In the rest of this section, we discuss the goal of regret minimization under the assumption that algorithms *can* randomize independently of $\bar{D}$. The problems discussed in this section all involve references to the agent's choice.

We consider a version of Newcomb's problem (introduced by [16]; see Sect. 8 for further discussion and references). In particular, we consider for any chooser c the decision problem $\mathrm{NP}_c = \{a_1, a_2\}$ which is resolved as follows. First, we let $D(a_1) = 1/4 + P(c = a_1)/2$. So the value of a_1 is proportional to the probability that c chooses a_1. And second, we let $D(a_2) = D(a_1) + P(c = a_1)/4$.

If we let $p = P(c = a_1)$, then the expected reward of c in this decision problem is $1/4 + p/2 + (1-p)p/4$. It is easy to see that this is strictly increasing in p and therefore maximized if $c = a_1$ deterministically. The regret, on the other hand, of c is $p^2/4$, which is also strictly increasing in p on $[0,1]$ and therefore minimized if $c = a_2$ deterministically. Similarly, the competitive ratio is given by $\frac{1/4+3p/4}{1/4+p/2+(1-p)p/4}$, which is also strictly increasing in p on $[0,1]$ and therefore also minimized if $c = a_2$ deterministically. Regret and competitive ratio minimization as rationality criteria would therefore require choosing the policy that minimizes the actual reward obtained in this scenario, only to minimize the value of actions not taken.

As noted in Sect. 8, it is a controversial among decision theorists what the rational choice in Newcomb's problem is. However, from the perspective of this paper in this particular version of the problem, it seems undesirable to require reward minimization. Also, it is easy to construct other (perhaps more convincing) cases. For example, if a high reward can be obtained by taking some action with a small probability, then regret minimizers take that action with high probability in a positive-frequency fraction of the rounds. Or consider a version of Newcomb's problem in which $D(a_1)$ is defined as before, but $D(a_2) = D(a_1)$. On such problems, Hannan-consistency is trivially satisfied by any learner, even though taking a_1 with probability 1 is clearly optimal.

D Why an even simpler theory fails and estimates are necessary

A simple mechanism of learning to choose is the *law of effect* (LoE):

> Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things

> being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.

This notion is implicit in many reinforcement learning algorithms. In (human) psychology it is also known as operant conditioning.

In situations like ours, where situations generally do not repeat exactly, for the law of effect to be meaningful, we have to applied on a meta level to general hypotheses or policies for making choices. So let a policy be a function that maps observations to actions. Then we could phrase this meta LoE as: if following a particular policy is accompanied with high rewards, then an agent will follow this policy more often in the future.

The BRIA criterion can be seen as abiding by this meta LoE, as the BRIA criterion requires testing different hypotheses and following the ones that have experimentally proven themselves. Its main conceptual innovation relative to the meta LoE is the bidding system, i.e., having the agent as well as hypotheses give estimates for how much utility will be achieved by making a particular choice, and using these estimates for testing and evaluation. A natural question then is: Are these conceptual additions to meta LoE necessary to obtain the kind of results we obtain? We here show why the answer is yes.

The biggest problem is quite simple to understand: if we don't restrict the testing regimen for policies, then biased testing can justify clearly suboptimal behavior. As an illustrative example, imagine that for all t, $\mathrm{DP}_t \in \mathrm{Fin}([0,1])$ where $r_t = \alpha_t^c$. That is, at each time the agent is offered to choose from some set of numbers between 0 and 1 and then obtains as a reward the chosen number. The agent tests two policies: The first simply chooses the maximum number. The second chooses, e.g., the worst option that is greater than $1/2$ if there is one, and the best option otherwise.

Of course, in this situation one would like the agent to learn at some point to follow the max policy. BRIAs indeed learn this policy (when accompanying the two tested policies with appropriate estimates) (cf. Theorem 3). But now imagine that the agent tests the max hypothesis primarily in rounds where all values are at most $1/2$ and the other hypothesis primarily in rounds in which there are options greater than $1/2$. Then the max hypothesis could empirically be associated with lower rewards than the max hypothesis, simply because it is tested in rounds in which the maximum achievable reward is lower.

To avoid this issue we would have to require that the set of decision problems on which hypothesis A is tested is in all relevant aspects the same as the set of decision problems on which hypothesis B is tested. Unfortunately, we do not know what the "relevant aspects" are. For instance, in the above problem it may be sufficient to test the max hypothesis on even time steps and the other hypothesis on odd time steps. However, there may also be problems where rewards depend on whether the problem is faced in an even or in an odd time step. More generally, it is easy to show that for each deterministic procedure of deciding which hypothesis to test, there is a decision process $\bar{\mathrm{DP}}, \bar{r}$ in which which this testing procedure introduces a relevant bias. In particular, the positive results we have proven in Theorems 3, 4 and 8 seem out of reach. We conclude that a direct deterministic implementation of meta LoE (without the use of estimates) is insufficient for constructing a criterion of rational choice.

Besides the estimates-based approach to this problem that we have developed in this paper, a different (perhaps more obvious) approach to this problem is to test *randomly*. For this, we assume that we have a randomization device available to us that is independent of $\bar{\mathrm{DP}}, \bar{r}$. If we then, for example, randomize uniformly between testing two hypotheses, testing is unbiased in the sense that for any potentially property of decision problems, as the number of tests goes to infinity, both hypotheses will be tested on the same fraction of problems with and without that property. This is essentially the idea behind randomized controlled trials. We have discussed this idea in Appendix C.

E Schnorr bounded algorithmic randomness

Definition 9. *A martingale is a function $d\colon \mathbb{B}^* \to [0,\infty)$ s.t. for all $w \in \mathbb{B}^*$ we have that $d(w) = {}^1\!/\!_2 d(w0) + {}^1\!/\!_2 d(w1)$. Let $w \in \mathbb{B}^\infty$ be an infinite sequence. We say that d succeeds on w if $\limsup_{n\to\infty} d(w_1...w_n) = \infty$.*

Definition 10. *We call $w \in \mathbb{B}^\omega$ ($O(g(t))$-boundedly) Schnorr random if there is no martingale d such that d succeeds on w and d can be computed (in $O(g(t))$) given everything revealed by time t.*

Theorem 9. *Let α be an ($O(h(t))$-computable) BRIA for $\overline{\mathrm{DP}}, \bar{r}$ covering all e.c. hypotheses. Let $\bar{a}$ be a sequence of terms in $\mathscr{T}$ s.t. $a_t \in \mathrm{DP}_t$ for all $t \in \mathbb{N}$ and the values r_t in the rounds t with α_t^c are ($O(h(t))$-boundedly) Schnorr random. Then in the limit as $T \to \infty$, it holds that $\sum_{t=1}^T r_t/T \geq {}^1\!/\!_2$.*

Proof. We conduct a proof by proving the following contrapositive: if the conlusion of the theorem does not hold, then $(r_t)_{t:\alpha_t^c = a_t}$ is not Schnorr random. Assume that there is $\varepsilon > 0$ s.t. $\sum_{t=1}^T r_t/T < {}^1\!/\!_2 - \varepsilon$ for infinitely many T. Then by the no overestimation criterion, there must also be an $\varepsilon > 0$ s.t. $\sum_{t=1}^T \alpha_t^e/T < {}^1\!/\!_2 - \varepsilon$ for infinitely many T. Consider the hypothesis $h_{a,\varepsilon}$ that always estimates ${}^1\!/\!_2 - \varepsilon$ and recommends a_t. Now let M_ε be $\bar{\alpha}$'s test for $h_{a,\varepsilon}$. From the fact that $\bar{\alpha}$ rejects $h_{a,\varepsilon}$ infinitely often, it follows that there are infinitely many $T \in \mathbb{N}$ such that $\sum_{t\in M_{\leq T}} r_t - ({}^1\!/\!_2 - \varepsilon) < 0$.

From this fact, we will now define an ($O(h(t))$-computable) martingale d that succeeds on the sequence $(r_t)_{t:\alpha_t^c = a_t}$. First, define $d() = 1$. Whenever T is not in M, define $d((r_t)_{t<T:\alpha_t^c = a_t}0) = d((r_t)_{t<T:\alpha_t^c = a_t}) = d((r_t)_{t<T:\alpha_t^c = a_t}1)$. That is, when $T \notin M$, don't bet on r_T. If $T \in M$, then bet some small, constant fraction δ of the current money that the next bit will be 0. That is, $d((r_t)_{t<T:\alpha_t^c = a_t}0) = (1+\delta)d((r_t)_{t<T:\alpha_t^c = a_t})$ and $d((r_t)_{t<T:\alpha_t^c = a_t}1) = (1-\delta)d((r_t)_{t<T:\alpha_t^c = a_t})$. Clearly, d thus defined is a martingale that is computable based on $\bar{\alpha}, M$.

Now we now know that there are infinitely many T s.t. $d((r_t)_{t<T:\alpha_t^c = a_t}) \geq (1+\delta)^{T+\varepsilon T}(1-\delta)^T$. It is left to show that for small enough δ, $(1+\delta)^{T+\varepsilon T}(1-\delta)^T \to \infty$ as $T \to \infty$.

First notice that

$$(1+\delta)^{T+\varepsilon T}(1-\delta)^T = ((1+\delta)(1-\delta))^T(1+\delta)^{T\varepsilon} = (1-\delta^2)^T(1+\delta)^{T\varepsilon} = \left((1-\delta^2)(1+\delta)^\varepsilon\right)^T.$$

So we need only show that for small enough but positive δ, $(1-\delta^2)(1+\delta)^\varepsilon > 1$. The most mechanic way to do this is to take the derivative at $\delta = 0$ (where the left-hand side is equal to 1) and showing that it is positive. The derivative is $\frac{d}{d\delta}(1-\delta^2)(1+\delta)^\varepsilon = (1+\delta)^\varepsilon(\varepsilon - \delta(\varepsilon+2))$. Inserting $\delta = 0$ yields ε, which is positive. □

Cognitive Bias and Belief Revision

Panagiotis Papadamos
Technical University of Denmark
panagiotispapadamos@gmail.com

Nina Gierasimczuk
Technical University of Denmark
nigi@dtu.dk

In this paper we formalise three types of cognitive bias within the framework of belief revision: confirmation bias, framing bias, and anchoring bias. We interpret them generally, as restrictions on the process of iterated revision, and we apply them to three well-known belief revision methods: conditioning, lexicographic revision, and minimal revision. We investigate the reliability of biased belief revision methods in truth-tracking. We also run computer simulations to assess the performance of biased belief revision in random scenarios.

Keywords— belief revision, truth-tracking, cognitive bias, confirmation bias, framing bias, anchoring bias, computer simulations, learning theory

1 Introduction

Cognitive bias is a systematic human thought pattern connected with the distortion of received information, that usually leads to deviation from rationality (for a recent analysis see [18]). Such biases are specific not only to human intelligence, they can be also ascribed to artificial agents, algorithms and programs. For instance, confirmation bias can be seen as stubbornness against new information which contradicts the previously adopted view. In some cases such confirmation bias can be implemented into a system purposefully. Take as an example an authentication algorithm and a malicious user who is trying to break into an email account. Say that the algorithm, before it locks the access, allows only three attempts to enter the correct password. Hence, the algorithm (temporarily) insists that the user who tries to connect is the real holder of the credentials, despite the input being inconsistent with that hypothesis. The algorithm will not revise its 'belief' about the user's identity, until it receives the evidence to the contrary a specific number of times. Another unorthodox example of a biased artificial agent concerns anchoring bias, where an agent makes a decision based on a recent, selected piece of information, possibly ignoring other data. In the context of artificial agents, such situations may occur justifiably when resources (like time or memory) are limited. As an example consider two computers, A and B, connected within a network. Computer A attempts to communicate with computer B, but for some reason, computer A does not receive B's response within a specified time range and, as a result, erroneously considers B dead. This inability to communicate leads computer A to change its 'belief' about B's liveness, and, subsequently, to make decisions based on this distortion.

In this paper we study some dynamic aspects of three types of cognitive bias: confirmation bias, framing bias, and anchoring bias. We will apply them to three well-known belief revision methods: conditioning, lexicographic, and minimal revision [19, 17, 5, 4]. We first recall the background of the model of truth-tracking by belief revision from [7, 1, 2] (related to earlier work in [13, 14], see also [8]), which borrows from computational learning theory, and identifiability in the limit in particular [9, 11]. We proceed by investigating the effect of bias on truth-tracking properties of various belief revision policies. Finally, we

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 441–454, doi:10.4204/EPTCS.379.34

present our computer simulation in which we empirically compare the performance of biased and regular belief revision in different scenarios. We close with several directions of further work.

1.1 Background: truth-tracking and belief revision

We will now introduce basic notions, following the framework of truth-tracking by belief revision proposed in [2]. Our agents' uncertainty space will be represented by a so-called *epistemic space*, $\mathbb{S} = (S, \mathscr{O})$, where S is a non-empty, at most countable set of worlds (or states), and $\mathscr{O} \subseteq \mathscr{P}(S)$ is a set of possible observations. We will call any subset p of S a *proposition*, and we will say that a proposition p is *true in* $s \in S$ if $s \in p$.

Data streams and sequences describe the information an agent receives over time. A *data stream* is an infinite sequence of observations $\vec{O} = (O_0, O_1, \ldots)$, where $O_i \in \mathscr{O}$, for $i \in \mathbb{N}$. A *data sequence* is a finite initial segment of a data stream; we will write $\vec{O}[n]$ for the initial segment of $\vec{O}$ of length n, i.e., $\vec{O}_0, \vec{O}_1, \ldots, \vec{O}_{n-1}$. Given a (finite or infinite) data sequence σ, σ_n is the n-th element of in σ; $set(\sigma)$ is the set of elements enumerated in σ; $\#O(\sigma)$ is the frequency of observation O in σ; let τ be a finite data sequence, then $\tau \cdot \sigma$ is the concatenation of τ and σ. A special type of data streams are *sound and complete* streams. A data stream $\vec{O}$ is *sound with respect to a state* $s \in S$ if and only if every element in $\vec{O}$ is true in the world s, formally $s \in \vec{O}_n$, for all $n \in \mathbb{N}$. A data stream $\vec{O}$ is *complete with respect to a state* $s \in S$ if and only if every proposition true in s is in $\vec{O}$, formally if $s \in O$ then there is an $n \in \mathbb{N}$, such that $O = \vec{O}_n$. Sound and complete streams form the most accommodating conditions for learning.

Definition 1.1. *Given an epistemic space* $\mathbb{S} = (S, \mathscr{O})$ *and a data sequence* σ, *a* learning method L *(also referred to it as a* learner*), is a function that takes as an input the epistemic space* $\mathbb{S}$ *and the sequence* σ, *and returns a subset of* S, $L(\mathbb{S}, \sigma) \subseteq S$, *called a* conjecture.

The goal of learning is to identify the *actual world*, which is a special designated element of the epistemic space. Given the epistemic space of an agent and the incoming information, which is (to some degree) trusted, the agent learns facts about the actual world step by step in order to achieve its goal, identifying the actual world.

Definition 1.2. *Let* $\mathbb{S} = (S, \mathscr{O})$ *be an epistemic space,* $s \in \mathbb{S}$ *is identified in the limit by* L *on* $\vec{O}$, *iff there is a* k, *such that for all* $n \geq k$, $L(\mathbb{S}, \vec{O}[n]) = \{s\}$; $s \in \mathbb{S}$ *is identified in the limit by* L *iff* s *is identified in the limit by* L *on every sound and complete data stream for* s; S *is identified in the limit by* L *if all* $s \in S$ *are identified in the limit on by* L; *Finally,* $\mathbb{S}$ *is* identifiable in the limit *iff there exists an* L *that identifies it in the limit.*

To be able to talk about beliefs of our agents (and whether or not they align with the actual world), we add to the epistemic space a plausibility relation. Given an epistemic space $\mathbb{S} = (S, \mathscr{O})$, a *prior plausibility* assignment $\preceq \subseteq S \times S$ is a total preorder. Such $\mathbb{S}^{\preceq} = (S, \mathscr{O}, \preceq)$ will be called a plausibility space (generated from $\mathbb{S}$, for simplicity of our notation we will often refer to such space with $\mathbb{B}$). The prior plausibility assignment is not fixed—it may be different for different agents, and serves as starting points of their individual belief revision processes. Plausibility models allow defining beliefs of agents. For any proposition p, we will say that the agent believes p in $\mathbb{S}^{\preceq}$ if p is true in all worlds in $min_{\preceq}(S)$.

Plausibility spaces, and hence also beliefs, change during the belief revision process. We will focus on three popular belief revision methods that can drive such a learning: conditioning, lexicographic, and minimal belief revision.

Definition 1.3 (Revision method)**.** *A* one-step revision method R_1 *is a function such that for any plausibility space* $\mathbb{B} = (S, \mathscr{O}, \preceq)$ *and any observable proposition* $p \in \mathscr{O}$ *returns a new plausibility space* $R_1(\mathbb{B}, p)$*. We define three one-step revision methods:*

> Conditioning*,* $Cond_1$*, is a one-step revision method that takes as input a plausibility space* $\mathbb{B} = (S, \mathscr{O}, \preceq)$ *and a proposition* $p \in \mathscr{O}$ *and returns the restriction of* $\mathbb{B}$ *to* p*. Formally,* $Cond(\mathbb{B}, p) = (S^p, \mathscr{O}, \preceq^p)$*, where* $S^p = S \cap p$ *and* $\preceq^p = \preceq \cap (S^p \times S^p)$*.*
>
> Lexicographic revision*,* Lex_1*, is a one-step revision method that takes as input a plausibility space* $\mathbb{B} = (S, \mathscr{O}, \preceq)$ *and a proposition* $p \in \mathscr{O}$ *and returns a plausibility space* $Lex(\mathbb{B}, p) = (S, \mathscr{O}, \preceq^{'})$*, such that for all* $t, w \in S$*,* $t \preceq^{'} w$ *if and only if* $t \preceq_p w$ *or* $t \preceq_{\bar{p}} w$ *or* $(t \in p$ *and* $w \notin p)$*, where* $\preceq_p = \preceq \cap (p \times p)$*,* $\preceq_{\bar{p}} = \preceq \cap (\bar{p} \times \bar{p})$*, and* $\bar{p}$ *is the complement of* p *in* S*.*
>
> Minimal revision*,* $Mini_1$*, is a one-step revision method that takes as input a plausibility space* $\mathbb{B} = (S, \mathscr{O}, \preceq)$ *and a proposition* $p \in \mathscr{O}$ *and returns a new plausibility space* $Mini(\mathbb{B}, p) = (S, \mathscr{O}, \preceq^{'})$ *where for all* $t, w \in S$*, if* $t \in \min_p$ *and* $w \notin \min_p$*, then* $t \preceq^{'} w$*, otherwise* $t \preceq^{'} w$ *if and only if* $t \preceq w$*.*

An iterated belief revision method R *is obtained by iterating the one-step revision method* R_1*:* $R(\mathbb{B}, \lambda) = \mathbb{B}$ *if* λ *is an empty data sequence, and* $R(\mathbb{B}, \sigma \cdot p) = R_1(R(\mathbb{B}, \sigma), p)$*.*

Definition 1.4. *Let* R *be an iterated belief revision method,* $S^{\preceq}$ *a plausibility space, and* $\vec{O}$ *a stream. A* belief revision based learning method *is defined in the following way:* $L_R^{\preceq}(\mathbb{S}, \vec{O}[n]) = min_{\preceq} R(\mathbb{S}^{\preceq}, \vec{O}[n])$*.*

We will say that the revision method R *identifies* $\mathbb{S}$ *in the limit iff there is a* $\preceq$ *such that* $L_R^{\preceq}$ *identifies* $\mathbb{S}$ *in the limit. A revision method* R *is* universal *on a class* $\mathbb{C}$ *of epistemic spaces if it can identify in the limit every epistemic space* $\mathbb{S} \in \mathbb{C}$ *that is identifiable in the limit.*

Theorem 1.1 ([2])**.** *The belief revision methods Cond and Lex are universal, while Mini is not.*

Learning methods can be compared with respect to their power. We will say that a learner L' is *at least as powerful as* learner L, $L \sqsubseteq L'$, if every epistemic space $\mathbb{S}$ that is identified in the limit by L', is identified in the limit by L. We will say that L' is *strictly more powerful than* learner L, if $L \sqsubseteq L'$ and it's not the case that $L' \sqsubseteq L$. Analogously, using definition 1.4, we will apply the same terms to belief revision methods.

In the remainder of this paper we will discuss several ways of introducing cognitive bias into this picture of iterated belief revision and long-term truth-tracking, together with computer simulation results that paint a more quantitative picture of the analytical results.

2 Simulating belief revision

Throughout this work we also present the results of computer simulations we run to see how various (biased) methods compare with respect to their truth-tracking ability. To this end we implemented artificial belief revision agents (for the biased and unbiased scenarios), which try to identify a selected actual world on the basis of sound and complete streams. We use the object-oriented programming language Python. The code can be found in the repository of the project [15], and the structure of the code can be seen in Figure 1.

The simulation included both custom and random tests. Custom tests were created to check the correctness of the implemented functions, while random tests were created to investigate the reliability and the performance of the (biased) belief revision methods. In the implementation all plausibility spaces

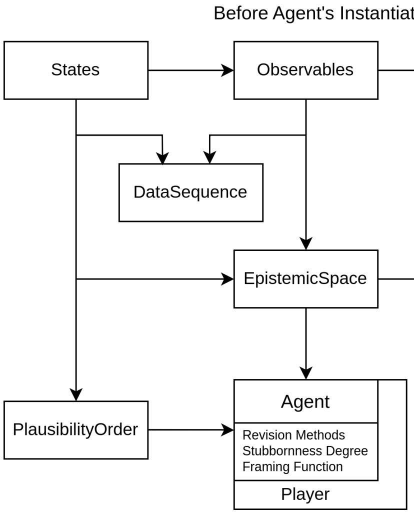


Figure 1: Communication of classes in the implementation

are finite. This choice is governed by the practicality of the implementation. We ran several series of tests. Each series of tests consisted of 200 tests, while the plausibility spaces consisted of ≈ 5 possible states and ≈ 12 observables, and the incoming data sequence was longer than the number of observables ($\approx 2-4$ more observables). These numbers were hard-coded to ensure computational feasibility of the experiment. The plausibility spaces we created for the automatic tests were completely random and so could turn out to be unidentifiable. This is the reason why there were identification failures for the universal revision methods, even for unbiased cases. After we randomly generated an epistemic space, one of the states (let us call it s) was randomly designated to be the actual world, and a sound and complete data stream σ for s was generated. A plausibility preorder over the epistemic was then randomly generated (generating a plausibility space). We then called on each of the (biased) revision methods and made them attempt to identify s from σ. As we will also see in the later comparisons, overall the frequencies of successful identification by unbiased (regular) belief revision methods were very high across experiments: for conditioning between 94% and 98%, for lexicographic revision between 97% and 99%, and for minimal revision between 77% and 82%.

3 Cognitive bias and belief revision

We will propose abstract accounts of three types of cognitive bias: confirmation bias, framing bias, and anchoring bias. For each we will describe how an agent revises its belief. We will see how the bias affects truth-tracking, both theoretically, through a learning-theoretic analysis of (non-)universality, and practically, in computer simulations.

3.1 Confirmation Bias

Hahn and Harries [10] characterized confirmation bias as a list of four ‘cognitions’, namely: hypothesis-determined information seeking, failure to pursue falsification strategy in the context of conditional reasoning, stubbornness to change of belief once formed, and overconfidence or illusion of validity of our belief. The first cognition will not concern us, as we don’t focus on agents that actively seek information, but rather we focus on how passive agents perceive incoming information.

To analyse selective bias, given a space $\mathbb{S} = (S, \mathscr{O})$, we could designate a subset of $\mathscr{O}$ to be the set of propositions that are 'important' to the agent. We would then allow that they are given a special, privileged treatment during the revision process. We choose to express this level of importance more generally with a numerical assignment, which we call the *stubbornness function*.

Definition 3.1. *Given an epistemic space* $\mathbb{S} = (S, \mathscr{O})$*, the* stubbornness function *is* $D : \mathscr{P}(S) \rightarrow \mathbb{N}$.

The stubbornness function describes the level of an agent's bias towards a proposition, intuitively the ones with stubbornness degree higher than 1 can be considered important to the agent. The higher the stubbornness degree, the more biased the agent is towards the proposition, so the more difficult it is to change its belief in that proposition—there should be strong evidence against it. For an unbiased agent the value of the function D for every proposition is 1. An unbiased agent will revise its beliefs instantly after it receives information inconsistent with its beliefs. An agent that is biased towards a proposition p and believes p, should receive information '$\neg p$' $D(p)$-many times in order to react by revising its belief with $\neg p$. The agent struggles with falsifying its belief, maintains the illusion of its belief's validity, by resisting change.

For each one-step revision method R_1 given in Definition 1.3, we will provide a confirmation-biased version or iterated revision R_{CB}. R_{CB} will take a plausibility space and a sequence of data and output a new plausibility space. Intuitively, it will attempt to execute the unbiased version of the revision method, but this will only succeed if the stubbornness degree allows it, i.e., if the data contradicting the proposition is repeated enough times.

Definition 3.2 (Confirmation-biased revision methods)**.** *Let* $\mathbb{B} = (S, \mathscr{O}, \preceq)$ *be a plausibility space and let* D *be a stubbornness function,* $\sigma \in \mathscr{O}^*$ *be a data sequence*[1]*,* $p \in \mathscr{O}$ *be an observable and* R_1 *is a one-step revision method. A confirmation-bias belief revision method* R_{CB} *is defined in the following way:*

$$R_{CB}(\mathbb{B}, \lambda) = \mathbb{B},$$

$$R_{CB}(\mathbb{B}, \sigma \cdot p) = \begin{cases} R_1(R_{CB}(\mathbb{B}, \sigma), p) & \textit{if} \quad \#p(\sigma) \geq D(\overline{p}), \\ R_{CB}(\mathbb{B}, \sigma) & \textit{otherwise}. \end{cases}$$

where λ *is an empty sequence,* $\#p(\sigma)$ *stands for the number of occurrences of* p *in* σ*, and* $\overline{p}$ *the complement of* p *in* S*.*

We obtain the confirmation-biased conditioning, lexicographic and minimal revision $Cond_{CB}$*,* Lex_{CB}*,* $Mini_{CB}$ *by substituting* R_1 *in the preceding definition by* $Cond_1, Lex_1$*, and* $Mini_1$*, respectively.*

Truth-tracking under confirmation bias An agent under confirmation bias updates its belief with respect to the stubbornness degree. Below we see that it is the crucial factor that breaks the universality of the belief revision methods.

Proposition 3.1. *Cond, Lex and, Mini are strictly more powerful than* $Cond_{CB}$*,* Lex_{CB}*, and* $Mini_{CB}$*, respectively.*

[1] Let Σ be a set, then Σ^* is a set of all finite sequences of elements from Σ.

Proof. We will give an example of an epistemic space $\mathbb{S} = (S, \mathscr{O})$ that is identified by *Cond*, but is not identified by $Cond_{CB}$. Let $\mathbb{S} = (S, \mathscr{O})$, where $S = \{w, t, s, r\}$, $\mathscr{O} = \{p, q, \bar{p}, \bar{q}\}$ and $p = \{w,t\}$,$\bar{p} = \{s, r\}$, $q = \{w, s\}$, and $\bar{q} = \{t, r\}$. Clearly, this space is identifiable by regular conditioning method *Cond*: take the plausibility order that takes all worlds to be equally plausible. Then, whichever world $s \in S$ is designated as the actual one, a sound a complete data stream for s will, in finite time, enumerate enough information to for the *Cond* method to delete all the other worlds, and so the actual world remains as the only one, and so also the minimal (most plausible) one.

To see that $Cond_{CB}$ will not be able to identify this space, let us assume that for all $x \in \mathscr{P}(S)$, $D(x) = 2$. We need to show that for any plausibility preorder on S there is a world $s \in S$, and a sound and complete stream $\vec{O}$ for s, such that $Cond_{CB}$ fails to identify s on $\vec{O}$. Take a preorder $\preceq$ on S, there are two cases, either (a) there is a unique minimal element s, or (b) there is none. For (a), take a $t \in S$, such that $s \preceq t$. There is a sound and complete stream $\vec{O}$ for t, that enumerates each observable true in t exactly once. While reading that sequence, $Cond_{CB}$ will not apply a single update, and so on a sound and complete sequence for t it will converge to s, which means it fails to identify t. For (b), a similar argument holds—for all among the minimal equiplausible worlds there will be a sound and complete sequence that enumerates every piece of data exactly once. On such a stream the update of $Cond_{CB}$ will not fire at all, and so there will be always more than one candidate for the actual world, so $Cond_{CB}$ will not converge to the singleton of the actual world.

It remains to be argued that *Cond* can identify in the limit everything that $Cond_{CB}$ can. Take an epistemic space $\mathbb{S} = (S, \mathscr{O})$, and assume that an $s \in S$ is identified in the limit by $Cond_{CB}$ on a stream $\vec{O}$ (that is sound and complete for s). That means that there is a $k \in \mathbb{N}$, such that for all $n \geq k$, $L^{\preceq}_{Cond_{CB}}(\mathbb{S}, \vec{O}[n]) = \{s\}$. So, for all $t \in S$ such that $t \neq s$, $\vec{O}[n]$ includes $O \in \mathscr{O}$, such that $t \notin O$. Hence, $L^{\preceq}_{Cond}(\mathbb{S}, \vec{O}[k]) = \{s\}$, and, since *Cond* only removes worlds, and $\vec{O}$ never enumerates anything false in s, $L^{\preceq}_{Cond}(\mathbb{S}, \vec{O}[n]) = \{s\}$, for all $n \geq k$.

A similar argument works for the Lex_{CB} and $Mini_{CB}$ method. □

Putting together Theorem 1.1 and Proposition 3.1 we get the following corollary.

Corollary 3.1. *$Cond_{CB}$ and Lex_{CB} are not universal.*

Clearly, confirmation bias can be detrimental to truth-tracking. The negative effect of stubbornness in revision can be uniformly overcome by the use of so-called fat streams, i.e., sound and complete streams that enumerate every information infinitely many times (which is possible as long as the set $\mathscr{O}$ is at most countable). Fat streams were introduced and studied before in computational learning theory in the context of memory-limited learners (see, e.g., [6]).

Simulation results We ran a comparative simulation study of confirmation-biased revision and the regular unbiased revision, following the method described in Section 2. The stubbornness values were randomly generated for all observables in the epistemic space as integers from 1 to 5. Figure 2 shows the respective frequencies of truth-tracking success.

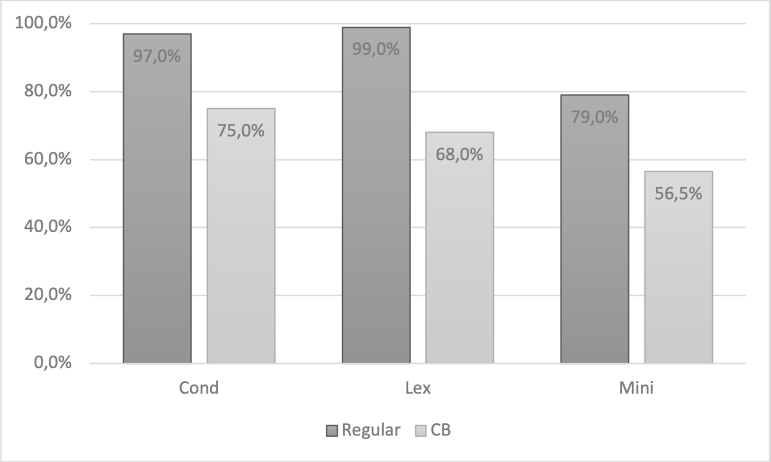


Figure 2: Confirmation-biased belief revision methods against unbiased belief revision methods

3.2 Framing Bias

Framing bias, also known as framing effect [12] refers to the fact that the way information is perceived (framed) by an agent can affect decision-making. We will introduce the framing function, FR which, broadly speaking, gives a range of interpretation for an observation, i.e., the incoming information can be 're-framed' into another information, within the range allowed by FR.

Definition 3.3. *Given an epistemic space $\mathbb{S} = (S, \mathscr{O})$, the* framing function *is $FR : \mathscr{O} \to \mathscr{P}(S)$.*

Note that the above definition is very general—we do not assume that the agent takes into account their observational apparatus, and so we allow for the observation to be interpreted as any proposition. While confirmation bias pertained to frequency of information in a stream, framing bias is related to its correctness and precision. We can pose a variety of constraints on framing, for instance we could require that the framed information is in some way related to the original information. In particular, in this paper we impose that, with the actual information O, the agent perceives X such that $X \subseteq O$. In this case, i.e, $FR(O) \subseteq \mathscr{P}(O)$. This particular kind of framing can be seen as overconfidence bias, since given an observation with some uncertainty range, the learner sees it as one with a narrower range, i.e., one that is more certain.

As before, we will formally model the three belief revision methods, conditioning, lexicographic revision, and minimal revision under the conditions of the bias.

Definition 3.4 (Framing-bias methods)**.** *Let $\mathbb{B} = (S, \mathscr{O}, \preceq)$ be a plausibility space, $\sigma \in \mathscr{O}^*$ a data sequence, $p \in \mathscr{O}$ an observable, FR a framing function, and and R_1 is a one-step revision method. We define a framing-biased method in the following way:*

$$R_{FR}(\mathbb{B}, \lambda) = \mathbb{B},$$

$$R_{FR}(\mathbb{B}, \sigma \cdot p) = R_1(R_{FR}(\mathbb{B}, \sigma), x), \text{ such that } x \in FR(p).$$

We obtain the framing-biased conditioning, lexicographic and minimal revision $Cond_{FR}$, Lex_{FR}, $Mini_{FR}$ by substituting R_1 in the preceding definition by $Cond_1, Lex_1$ and $Mini_1$, respectively.

Truth-tracking under framing bias As before, we will now investigate how framing bias affects truth-tracking capabilities of belief revision methods.

Definition 3.5. *Given a stream $\vec{O} = (O_0, O_1, \ldots)$ and a framing function FR, we define a framing of $\vec{O}$ as $FR(\vec{O}) = (P_0, P_1, \ldots)$, where for each $i \in \mathbb{N}$, $P_i \in FR(O_i)$. We will call $FR(\vec{O})$ static iff for every $i, j \in \mathbb{N}$, with $i \neq j$, if $O_i = O_j$ then $P_i = P_j$, otherwise $FR(\vec{O})$ is* dynamic.

The first observation is that there are limit cases in which framing will not restrict the learning power of any of the revision methods, for instance when framing is a static identity function, or in more complicated, lucky cases when sound and complete streams are framed into (possibly different) sound and complete streams. In general however, framing will result in a certain kind of blindness, some worlds can get overlooked during the revision process. In particular, given an observable O that is true at s, it might be the case that O will get mapped to a set P, such that $s \notin P$, in other words, the agent will interpret a true observation as a proposition that is false in the actual world. This would be detrimental to any revision method. Hence, we get the following propositions.

Proposition 3.2. *$Cond_{FR}$ and Lex_{FR} are not universal.*

Proposition 3.3. *$Mini$ is strictly more powerful than $Mini_{FR}$.*

The dynamic framing allows for *fair framing* of streams, where the agents observes input ‘erroneously’ for finitely many steps, after which it is presented a full sound and complete stream. This is a notion analogous to that of *fair streams* in [2], and the following is a direct consequence of the result therein of *Lex* being universal on fair streams.

Proposition 3.4. *Lex_{FR} is universal on fairly framed streams.*

Simulation results As before, we ran a comparative simulation study of confirmation-biased revision and the regular unbiased revision. As before we generate a sound and complete stream, which then gets transformed into its framed version, by applying the framing function to each observation independently. By the restrictions we impose, the framing function outputs always a random subset of the original proposition, which can be the empty set. Figure 3 shows the respective frequencies of truth-tracking success.

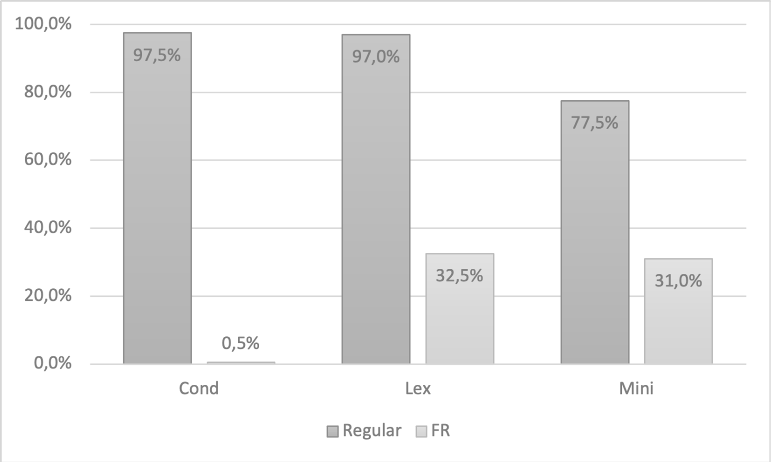


Figure 3: Framing-biased belief revision methods against unbiased belief revision methods

3.3 Anchoring Bias

Anchoring bias plays a role in decision-making influenced by the most recently received information, and it is strongly connected to lack of resources. We make everyday decisions under time pressure. These decisions are, often unconsciously, influenced by the piece of information received last before the decision point [16]. Moreover, anchoring bias in real-life scenarios can introduce a level of randomness in decision making. Consider, as an example, a student who takes part in an exam involving a multiple choice test. Due to lack of time they have to answer a question without being able to analyse it properly. While going through possible answers, the student might pick one that reminds them of something they have seen recently in their notes. As in the previous cases, we will provide a general definition of anchoring-biased methods. The mechanism will consists of two components, one is that the revision mechanism will always perform a minimal change, the other one is that in the case the revision step results in multiple minimal possible words, one of them will be chosen at random and made most plausible overall. In order to phrase this formally, we need several new notions. Given a set S, a preorder $\preceq\subseteq(S\times S)$, and $x\in S$, we define $\preceq\uparrow x:=(\preceq\cap(S\setminus\{x\}\times S\setminus\{x\}))\cup\{(x,s)\mid s\in S\setminus\{x\}\}$. Intuitively, this operation takes an order and outputs a new updated version of it, with x upgraded to be the most plausible world. Now we will define new versions of one-step revision methods, which include in their first part the unbiased one-step revision methods and in their second part the upgrade operator. Let $\mathbb{B}=(S,\mathscr{O},\preceq)$, $p\in\mathscr{O}$ and $Lex_1(\mathbb{B},p)=(S,\mathscr{O},\preceq')$, we define

$$Lex_1^+(\mathbb{B},p)=\begin{cases}(S,\mathscr{O},\preceq') & \text{if } |min_{\preceq'}S|=1;\\ (S,\mathscr{O},\preceq'\uparrow x),\ \text{with } x\in min_{\preceq'}S & \text{otherwise.}\end{cases}$$

The upgraded minimal revision, $Mini_1^+$, is defined analogously. It remains to discuss what happens when conditioning results in several minimal worlds. We propose the following interpretation. Let $\mathbb{B}=(S,\mathscr{O},\preceq)$, $p\in\mathscr{O}$ and $Cond_1(\mathbb{B},p)=(S',\mathscr{O}',\preceq')$, we define

$$Cond_1^+(\mathbb{B},p)=\begin{cases}(S',\mathscr{O}',\preceq') & \text{if } |min_{\preceq'}S'|=1;\\ (\{x\},\mathscr{O}',\emptyset),\ \text{with } x\in min_{\preceq'}S' & \text{otherwise.}\end{cases}$$

$Cond_1^+$ is a very 'impatient' method, as long as a singular minimal world is available, it just follows the usual drill, but if at any stage several worlds are most plausible, it picks one of them and throws away the rest of the space. This is very radical, but this way we avoid upgrading the order, which would go against the spirit of conditioning.

Definition 3.6 (Anchoring-biased methods)**.** *Let $\mathbb{B}=(S,\mathscr{O},\preceq)$, $\sigma\in O^*$ a data sequence, $p\in O$ an observable. We define the anchoring-biased methods R_{AB} as:*

$$R_{AB}(\mathbb{B},\lambda)=\mathbb{B},$$

$$R_{AB}(\mathbb{B},\sigma\cdot p)=R_1^+(R_{AB}(\mathbb{B},\sigma),min_{\preceq_{AB}}(S_{AB}\cap p)),$$

where $R_{AB}(\mathbb{B},\sigma)=(S_{AB},\mathscr{O}_{AB},\preceq_{AB})$. We obtain the anchoring-biased conditioning, lexicographic and minimal revision $Cond_{AB}$, Lex_{AB}, $Mini_{AB}$ by substituting R_1^+ above by $Cond_1^+$,Lex_1^+ and $Mini_1^+$, respectively.

Unbiased minimal belief revision is in itself, interestingly, a form of anchoring bias. An agent using minimal belief revision actually uses the most plausible worlds where the incoming information is true to update its belief accordingly. When it comes to lexicographic revision, the definition is slightly different, but the behavior of anchoring-biased lexicographic belief revision is the same as that of unbiased minimal revision. By imposing the extra upgrade condition we make anchoring-biased methods more 'actionable', reflecting the fact that anchoring bias often plays a role in quick decision-making. After each revision step anchoring ensures that there is a candidate for the best possible world, which is randomly selected among the minimal worlds at that stage. This is especially important if resources for performing revision are limited (in the simulation these cases will be labeled '-res'). We will see that this augmentation positively affects the biased methods, even though in general the anchoring biased belief revision methods are not universal.

Truth-tracking under anchoring bias Anchoring bias is most prominently connected to lack of resources. For example, when someone needs to make a decision under the pressure of time, anchoring bias can be used as heuristic. In this section we will show that, even though anchoring bias breaks universality, it can facilitate faster identification of the actual world.

Example 3.1. *Consider the plausibility space $\mathbb{B} = (\mathbb{S}, \preceq)$, where $\mathbb{S} = (S, \mathscr{O})$, $S = \{w, r, s, t\}$ and s the actual world. The initial plausibility order is $w \preceq t \simeq s \preceq r$, so the agent is indifferent between the worlds t and s, and the observable propositions are $p = \{w\}, q = \{r, t, w\}, \bar{p} = \{r, s, t\}$ and $\bar{q} = \{s\}$. Consider also a sound a complete data stream with respect to the actual world, $\vec{O} = (\bar{p}, \ldots, \bar{q}, \ldots)$. An agent using anchoring-biased conditioning identifies the actual world in the first piece of information with probability .5. Of course, with probability .5 the actual world is excluded and so the agent will not identify it. Assuming that the biased agent identifies the actual world, anchoring-biased conditioning is faster than conditioning by $k-1$ steps, where $\vec{O}_\kappa$ is the first occurrence of $\bar{q}$ in the data stream $\vec{O}$. Note that unbiased minimal revision will identify the world s only after receiving $\bar{q}$.*

The above example points at the following proposition.

Proposition 3.5. *$Cond_{AB}$ is not universal.*

Moreover, since Lex_{AB} is a version of $Mini$, based on Theorem 1.1, we can state the following.

Proposition 3.6. *Lex_{AB} is not universal.*

Even though anchoring-biased lexicographic belief revision is not universal, it can facilitate faster truth tracking. The argument includes cases wherein the agent is indifferent between more than one most plausible worlds. Recall that an agent which uses anchoring-biased lexicographic revision revises similarly to one that uses unbiased minimal revision, but if the set of the worlds which considers most plausible is not a singleton, it selects one of the most plausible worlds with equal probability.

Unbiased minimal revision can be seen as a form of anchoring bias, as an agent that uses minimal belief revision, minimally updates its belief to be compatible with $min_{\preceq}(p)$. The difference is in the way they select the most plausible worlds after each update. Anchoring bias minimal revision and unbiased minimal revision will be compared in simulations below, where we investigate if the randomness included in anchoring-biased minimal revision improves the performance with respect to unbiased minimal revision.

Simulation results We again ran a comparative simulation study of confirmation-biased revision and the regular unbiased revision, following the method described in Section 2. In the case there was more than one minimal state at a certain stage of the belief revision process, the anchoring method selected one of the minimal states at random to be the conjecture of the learning method. Figure 2 shows the respective frequencies of truth-tracking success.

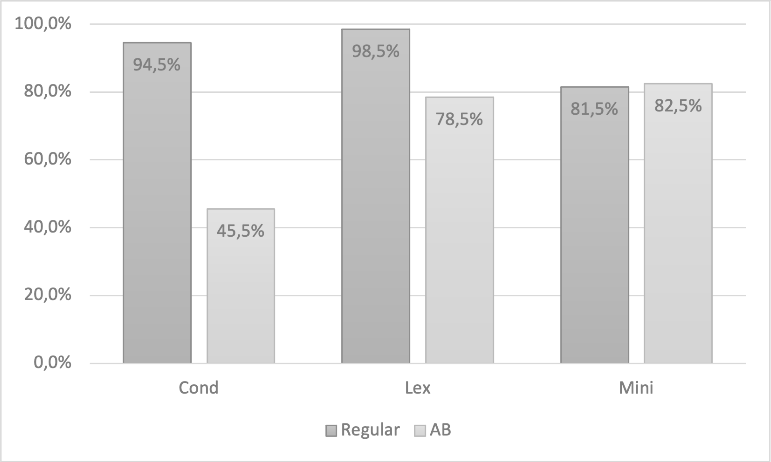


Figure 4: Anchoring-biased belief revision against unbiased belief revision

As anchoring bias often shows up in the context of limited resources, we run another experiment, wherein we included a parameter (a real number between 0 and 100) which decreases each time a revision takes place, and the process terminates when the resource is depleted. In this particular implementation, each time a revision is executed the available resource is halved and the agent stops revising when its resources fall below 1. As we can see in Figure 5 the anchoring ability to select a random world to be the candidate for the actual world improves the truth-tracking ability, especially in the case of minimal revision.

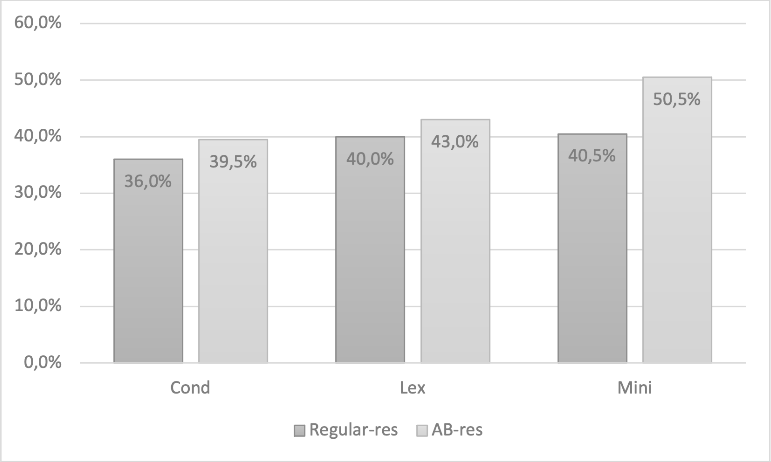


Figure 5: Anchoring-biased belief revision against unbiased belief revision - limited resources

Finally, let us summarize some general observations about the simulation. Various components of a plausibility space affect the performance of the methods, both biased and unbiased ones. Specifically, an increase in the number of states negatively affects the performance of the belief revision methods (see Figure 6), while an increase in the number of observables decreases the number of non-identifiable worlds, which in effect can make unbiased methods fail. More plots with the results can be found on the project repository [15].

We also saw that, as expected, cognitive-biased belief revision methods perform worse than the unbiased ones. An exception is the anchoring-biased minimal belief revision method. Additionally, when limited resources are implemented, anchoring-biased belief revision methods perform better than the unbiased ones. This is a significant result, as it provides a potential alternative tool for truth-tracking when the resources are limited, which is usually the case in real life scenarios.

4 Conclusions

Cognitive bias in artificial intelligence is an interesting topic with a bright future, and as such deserves to be investigated in the context of belief revision and knowledge representation. In this paper we provided ways to formalize bias in belief-revision and learning. The three kinds of bias we discussed had completely different character, and employed different components of our belief revision based learners. We have also shown that bias can be detrimental to learning understood as truth-tracking.

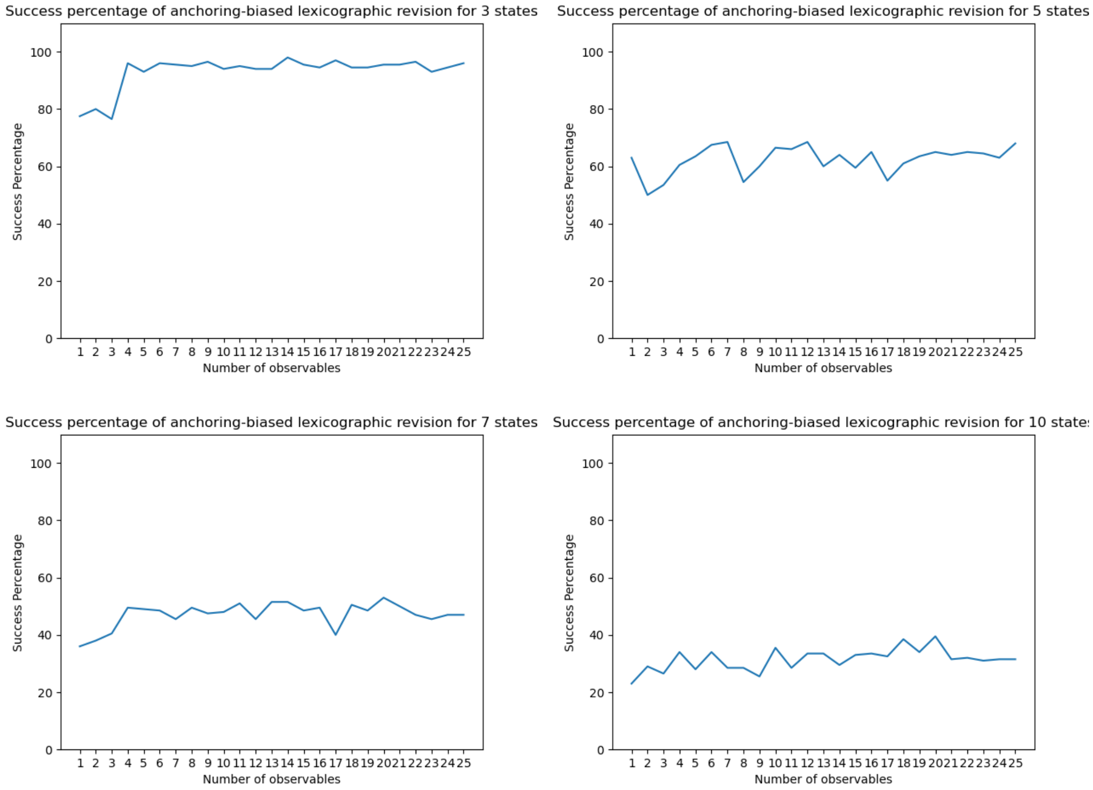
In general, biased methods are by far less reliable than the unbiased ones. While cognitive bias is generally problematic for truth-tracking, when resources are scarce it can be considered a tool or a heuristic. Anchoring-biased methods are a good example here, as the tests we conducted showed. This point can also serve as a rehabilitation of minimal revision, which in general is not a universal learning method.

When it comes to the simulation, we have found, in line with our expectations, that *Cond* and *Lex* identify the actual world in almost every test. Moreover, in general, the larger the number of observables, the higher the chances for the agent to identify the actual world. The same holds for the length of the data sequence, see Figure 6. Biased belief revision methods, are in general less successful than the unbiased ones—in particular, the information loss in framing can be fatal for truth-tracking by conditioning. On the other hand, anchoring bias can be used as a heuristic for faster identification.

In our work we model only some types of cognitive bias, the ones more applicable in artificial intelligence. Types mostly related to human emotional decision-making were intentionally excluded, but they would be a very interesting topic of future work. Moreover, although we investigated how randomness on the states, observables, and data streams affects truth-tracking, randomness of the environment itself is not a factor in this model. Assigning some bias to the elements of the tests could potentially give better insights into truth-tracking. Finally, it would be very interesting to relate our results to the existing work on resource bounded belief revision in the AGM paradigm, in particular to [20], to look for expressibility results in the context of dynamic logic of learning theory (DLLT, [3]), and, last but not least, make steps towards empirical predictions for cognitive science of bias.

References

[1] A. Baltag, N. Gierasimczuk & S. Smets (2011): *Belief revision as a truth-tracking process*. In K. Apt, editor: *TARK'11: Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge, Groningen, The Netherlands, July 12-14, 2011*, ACM, New York, NY, USA, pp. 187–190, doi:10.1145/2000378.2000400.


Success percentage of anchoring-biased lexicographic revision for 3 states

Success percentage of anchoring-biased lexicographic revision for 5 states

Success percentage of anchoring-biased lexicographic revision for 7 states

Success percentage of anchoring-biased lexicographic revision for 10 state

Figure 6: Number of observables against anchoring bias lexicographic revision's performance for different number of possible states. The remaining graphs can be accessed in our GitHub repository [15].

[2] A. Baltag, N. Gierasimczuk & S. Smets (2019): *Truth-Tracking by Belief Revision*. Studia Logica 107(5), pp. 917–947, doi:10.1007/s11225-018-9812-x.

[3] A. Baltag, N. Gierasimczuk, A. Özgün, A.L. Vargas Sandoval & S. Smets (2019): *A dynamic logic for learning theory*. Journal of Logical and Algebraic Methods in Programming 109, p. 100485, doi:10.1016/j.jlamp.2019.100485.

[4] J. van Benthem (2007): *Dynamic logic for belief revision*. Journal of Applied Non-Classical Logics 17(2), pp. 129–155, doi:10.3166/jancl.17.129-155.

[5] C. Boutilier (1993): *Revision Sequences and Nested Conditionals*. In: *IJCAI'93: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Chambery, France, pp. 519–525, doi:10.5555/1624025.1624098.

[6] L. Carlucci, J. Case, S. Jain & F. Stephan (2007): *Results on memory-limited U-shaped learning*. Information and Computation 205(10), pp. 1551–1573, doi:10.1016/j.ic.2007.04.001.

[7] N. Gierasimczuk (2010): *Knowing One's Limits. Logical Analysis of Inductive Inference*. Ph.D. thesis, Universiteit van Amsterdam, The Netherlands.

[8] N. Gierasimczuk, V.F. Hendricks & D. de Jongh (2014): *Logic and Learning*. In A. Baltag &

S. Smets, editors: *Johan van Benthem on Logic and Information Dynamics*, Springer International Publishing, Cham, pp. 267–288, doi:10.1007/978-3-319-06025-5_10.

[9] E.M. Gold (1967): *Language Identification in the Limit*. *Information and Control* 10, pp. 447–474, doi:10.1016/S0019-9958(67)91165-5.

[10] U. Hahn & A.J. Harris (2014): *Chapter Two - What Does It Mean to be Biased: Motivated Reasoning and Rationality*. In B.H. Ross, editor: *Psychology of Learning and Motivation*, 61, Academic Press, pp. 41–102, doi:10.1016/B978-0-12-800283-4.00002-2.

[11] S. Jain, D. Osherson, J.S. Royer & A. Sharma (1999): *Systems that Learn*. MIT Press, Chicago.

[12] D. Kahneman & A. Tversky (1979): *Prospect Theory: An Analysis of Decision under Risk*. *Econometrica* 47(2), pp. 263–291, doi:10.2307/1914185.

[13] K.T. Kelly (1998): *The learning power of belief revision*. In: *TARK'98: Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 111–124, doi:10.5555/645876.671884.

[14] K.T. Kelly (1999): *Iterated Belief Revision, Reliability, and Inductive Amnesia*. *Erkenntnis* 50(1), pp. 7–53, doi:10.1023/A:1005444112348.

[15] P. Papadamos & N. Gierasimczuk (2023): *Source Code for Simulations in Cognitive Bias and Belief Revision*. Available at `https://github.com/papos8/BeliefRevisionSimulation`.

[16] J. Rezaei (2021): *Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making*. *Journal of Decision Systems* 30(1), pp. 72–96, doi:10.1080/12460125.2020.1840705.

[17] H. Rott (1989): *Conditionals and theory change: Revisions, expansions, and additions*. *Synthese* 81(1), pp. 91–113, doi:10.1007/BF00869346.

[18] A. Solaki, F. Berto & S. Smets (2021): *The Logic of Fast and Slow Thinking*. *Erkenntnis* 86(3), pp. 733–762, doi:10.1007/s10670-019-00128-z.

[19] W. Spohn (1988): *Ordinal Conditional Functions: A Dynamic Theory of Epistemic States*. In W.L. Harper & B. Skyrms, editors: *Causation in Decision, Belief Change, and Statistics: Proceedings of the Irvine Conference on Probability and Causation*, Springer Netherlands, Dordrecht, pp. 105–134, doi:10.1007/978-94-009-2865-7_6.

[20] R. Wassermann (1999): *Resource Bounded Belief Revision*. *Erkenntnis* 50(2), pp. 429–446, doi:10.1023/A:1005565603303.

An Abstract Look at Awareness Models and Their Dynamics

Carlo Proietti
ILC, CNR
Genova, Italy
carlo.proietti@ilc.cnr.it

Fernando R. Velázquez-Quesada
University of Bergen
Norway
Fernando.VelaquezQuesada@uib.no

Antonio Yuste-Ginel
Universidad Complutense de Madrid,
Spain [*]
antoyust@ucm.es

This work builds upon a well-established research tradition on modal logics of awareness. One of its aims is to export tools and techniques to other areas within modal logic. To this end, we illustrate a number of significant bridges with abstract argumentation, justification logics, the epistemic logic of knowing-what and deontic logic, where basic notions and definitional concepts can be expressed in terms of the awareness operator combined with the $\Box$ modality. Furthermore, these conceptual links point to interesting properties of awareness sets beyond those standardly assumed in awareness logics – i.e. positive and negative introspection. We show that the properties we list are characterised by corresponding canonical formulas, so as to obtain a series of off-the-shelf axiomatisations for them. As a second focus, we investigate the general dynamics of this framework by means of event models. Of specific interest in this context is to know under which conditions, given a model that satisfies some property, the update with an event model keeps it within the intended class. This is known as the closure problem in general dynamic epistemic logics. As a main contribution, we prove a number of closure theorems providing sufficient conditions for the preservation of our properties. Again, these results enable us to axiomatize our dynamic logics by means of reduction axioms.

1 Introduction

Epistemic logics of awareness [20, 34] are extensions of propositional epistemic logic (EL; [26]) introduced for modelling a form of (*explicit*) knowledge that lacks closure under logical consequence (therefore avoiding the *logical omniscience* problem). The idea is that knowledge requires both lack of uncertainty (the standard $\Box$ modality) *and* awareness, with the latter a unary modality that, semantically, verifies whether the given formula belongs to a specified world-dependant *awareness* set. One can deal with specific awareness properties (e.g., awareness introspection) by specifying not only the properties of the awareness sets but also their interaction with the accessibility relations. One can also look at dynamics of awareness in the dynamic epistemic logic style (DEL; [6, 17, 9, 7]), defining model-changing actions for representing acts of awareness *elicitation* or *forgetting* [11, 38, 15, 21].

The epistemic awareness setting can also be interpreted more generally by abstracting away from this specific reading (see Section 2). At a general level, one can read the awareness entities as a set $\mathbf{O}$ of generic objects, and the corresponding awareness modality as capturing the notion of "owning some

[*] The research activity of Antonio Yuste-Ginel was partly funded by the project PID2020-117871GB-I00 whose principal investigator is Francisco Salguero Samillar. Part of the research activity that crystallised in this paper was carried out when he was employed at the IRIT (Toulouse) by a postdoc contract funded by EUICT-482020 project TAILOR GA952215.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 455–469, doi:10.4204/EPTCS.379.35

abstract object $o \in \mathbf{O}$". By doing so, one can find connections with other modal logics where abstract objects are used as additional or definitional concepts. For example, other approaches in epistemic and deontic logic [31, 28, 23, 37] can be seen as instances of a more general awareness-like framework. From this perspective, model properties connecting the "owning-the-object" operator "O" with $\Box$ constitute interesting desiderata. This paper defines a number of such properties and characterises them with formulas of the **O**-language.

A second aim of this work is to investigate the dynamics of general **O**-models. We use *event models* as in [5] for their power to encode epistemic and factual changes at an extreme level of granularity [18]. Yet, a drawback of it is the often non-trivial *closure problem*: guaranteeing that, for a given class $\mathfrak{M}$ of models, the product update of an $\mathfrak{M}$-model with an event model remains in $\mathfrak{M}$. Closure results clarify the general constraints for the executability of actions, and therefore provide safe guidance for modelling them. Some closure theorems are available for DEL, establishing sufficient conditions for the preservation of accessibility relations [3, 7]. However, this issue is relatively underexplored for properties relating accessibility relations and awareness sets, as the ones mentioned above (with the exception of [33]). As a central contribution of our work, we prove closure theorems for these properties. As an important byproduct, this serves to find direct roads to axiomatisation via reduction axioms.

The paper proceeds as follows. Section 2 introduces the general **O**-framework, illustrating some of its applications. Crucially, it also lists meaningful model properties (at both the individual and multi-agent level; Subsection 2.1), providing their syntactic characterizations as well as their complete axiomatisations as a main result. Section 3 is about the dynamics of **O**-models, semantically: we introduce event models and the closure problem, identifying sufficient conditions for the preservation of the discussed model properties. Section 4 looks at dynamics from the syntactic side, providing sound and complete axiomatisations for dynamic **O**-logics. We end with a discussion of our results in Section 5. Sketches of proofs are to be found in the Appendix.

2 Basic framework

Through this document, let Ag be a finite non-empty set of agents, At be a countable set of propositional variables, and **O** be a countable non-empty set of abstract objects. An **O**-model is just a multi-relational model together with a function that assigns, to each agent, a set of objects from **O** at each possible world.

Definition 1 (O-Model) *An* **O**-model *is a tuple* $\mathcal{M} = (\mathcal{W}, \mathcal{R}, \mathcal{O}, \mathcal{V})$ *where* $\mathcal{W} \neq \emptyset$ *is a set whose elements are called possible worlds,* $\mathcal{R} : \mathsf{Ag} \to \wp(\mathcal{W} \times \mathcal{W})$ *assigns a binary relation on* $\mathcal{W}$ *to each agent* $i \in \mathsf{Ag}$*,* $\mathcal{O} : (\mathsf{Ag} \times \mathcal{W}) \to \wp(\mathbf{O})$ *assigns a set of objects to each agent* $i \in \mathsf{Ag}$ *at each world* $w \in \mathcal{W}$*, and* $\mathcal{V} : \mathsf{At} \to \wp(\mathcal{W})$ *is an atomic valuation function. Note:* $\mathcal{R}_i$ *abbreviates* $\mathcal{R}(i)$*, and* $\mathcal{O}_i(w)$ *stands for* $\mathcal{O}(i,w)$*. The set of worlds of a given* $\mathcal{M}$ *is referred to as* $\mathcal{W}[\mathcal{M}]$*; the same convention applies* $\mathcal{R}$*,* $\mathcal{O}$ *and* $\mathcal{V}$*. We use infix notation for each* $\mathcal{R}_i$*. A* pointed **O**-model *is a tuple* $(\mathcal{M}, w)$ *with* $\mathcal{M}$ *an* **O***-model and* $w \in \mathcal{W}[\mathcal{M}]$*. Finally,* $\mathfrak{M}^{\mathbf{O}}$ *denotes the class of all* **O***-models.* ◀

The language for describing **O**-models is the following.

Definition 2 (Language $\mathcal{L}$) *Given* Ag*,* At*, and* **O** *as above, formulas* φ *of the language* $\mathcal{L}$ *are given by*

$$\varphi ::= \top \mid p \mid \mathrm{O}_i\, o \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_i \varphi$$

with $p \in \mathsf{At}$*,* $i \in \mathsf{Ag}$ *and* $o \in \mathbf{O}$*. Other Boolean constants/operators are defined as usual; likewise for the modal dual* $\Diamond_i \varphi$*, defined as* $\neg\Box_i\neg\varphi$*. Formulas of* $\mathcal{L}$ *are interpreted at pointed* **O***-models. The truth-clauses for the multi-modal fragment of* $\mathcal{L}$ *are the standard ones; for the new formulas,*

$$\mathcal{M}, w \models \mathrm{O}_i o \quad \textit{iff} \quad o \in \mathcal{O}_i(w).$$

Global truth of a formula and a set of formulas in a model is defined as usual [12], and denoted $\mathcal{M} \models \varphi$ *and* $\mathcal{M} \models \Phi$*, respectively. Likewise for the notion of validity (notation:* $\models \varphi$*).* ◀

Let us now present some particular interpretations and instantiations of **O**-models.

Models for general and atomic awareness A model for *general* awareness [20] is an **O**-model where **O** is the language $\mathcal{L}$ itself. In this context, $\mathcal{O}_i$ is called the *awareness function* and it is denoted by $\mathcal{A}_i$; notationally, the operator O_i is replaced by the awareness operator A_i. A model for *atomic awareness* [20, 22] is instead one where the awareness function $\mathcal{A}$ returns a set of atoms from At, with agent i aware of φ at a world w if and only if the set of atoms in φ is a subset of $\mathcal{A}_i(w)$. These structures correspond to **O**-models where **O** is a set of atoms At. Syntactically, $\mathcal{M}, w \models \mathrm{A}_i p$ iff $p \in \mathcal{A}_i(w)$, and then one can define inductively an additional modality $\widetilde{\mathrm{A}}_i$ that works over arbitrary formulas:

$$\begin{array}{lll} \widetilde{\mathrm{A}}_i\top := \top, & \widetilde{\mathrm{A}}_i\neg\varphi := \widetilde{\mathrm{A}}_i\varphi, & \widetilde{\mathrm{A}}_i\Box_j\varphi := \widetilde{\mathrm{A}}_i\varphi, \\ \widetilde{\mathrm{A}}_i p := \mathrm{A}_i p, & \widetilde{\mathrm{A}}_i(\varphi\wedge\psi) := \widetilde{\mathrm{A}}_i\varphi\wedge\widetilde{\mathrm{A}}_i\varphi, & \widetilde{\mathrm{A}}_i\widetilde{\mathrm{A}}_j\varphi := \widetilde{\mathrm{A}}_i\varphi. \\ \widetilde{\mathrm{A}}_i\mathrm{A}_j p := \widetilde{\mathrm{A}}_i p. & & \end{array}$$

In this way, $\mathcal{M}, w \models \widetilde{\mathrm{A}}_i\varphi$ if and only if $\mathrm{atm}(\varphi) \subseteq \mathcal{A}_i(w)$, with $\mathrm{atm}(\varphi)$ the set of atoms occurring in φ.

Models for awareness of arguments One can also conceive **O** as a set of *abstract arguments* and $\mathcal{O}$ as a function indicating the set of arguments that each agent is aware of at each world [35], so that $\mathbf{O}_i a$ means "agent i is aware of argument a" or "agent i is able to use argument a". The resulting models constitute 'epistemic' versions of the abstract models of argumentation introduced in [19]. The main idea behind *abstract argumentation* is to represent arguments as nodes of a graph, and attacks among them as arrows of the graph. There, argumentative notions such as argument acceptability are reduced to graph-theoretical notions, such as stability of a set within a graph. In the modalized (multi-agent) versions, a possible world is constituted by one such graph plus the specification of which arguments and attacks each agent is aware of. This enables us to express higher-order uncertainty about awareness of arguments [35], which is in turn crucial for modelling strategic reasoning in an argumentative environment [36] and its dynamics [32, 33]. In a similar vein, **O**-models have been applied to more structured frameworks for argumentation [13, 14], with **O** understood as a set of ASPIC$^+$ arguments [30].

Justification logics In the justification logics of [2], justifications are abstract objects which have structure and operations on them. Formally, the set of *justification terms* J is built from sets of justification constants and justification variables by means of the operations of application ('$\cdot$') and sum ('$+$'). Thanks to them, one can define the language $\mathcal{L}_J$ as the basic (multi)modal language plus expressions of the form $t{:}\varphi$ (with t a term and φ a formula), read as *"t is a justification for φ"*. Formulas of this extended language are interpreted over *justification models*, tuples $M = (\mathcal{W}, \mathcal{R}, \mathcal{E}, \mathcal{V})$ where $\mathcal{W}$, $\mathcal{R}$ and $\mathcal{V}$ are as in a **O**-models. The new component, the evidence function $\mathcal{E} : (J \times \mathcal{L}_J) \to \wp(\mathcal{W})$, provides the set of worlds $\mathcal{E}(t, \varphi)$ in which the term t is relevant/admissible evidence for the formula φ. For this to work properly, $\mathcal{E}$ should satisfy both

$$\mathcal{E}(s, \varphi \to \psi) \cap \mathcal{E}(t, \varphi) \subseteq \mathcal{E}(s{\cdot}t, \psi) \quad \text{and} \quad \mathcal{E}(s, \varphi) \cup \mathcal{E}(t, \varphi) \subseteq \mathcal{E}(s+t, \varphi).$$

Then, $(M, w) \models t{:}\varphi$ if and only if both $w \in \mathcal{E}(t, \varphi)$ and φ holds in all worlds $\mathcal{R}$-reachable from w.

A justification model can be seen as an **O**-model in which the codomain of $\mathcal{O}$ is a set of pairs of the form (justification, formula). Indeed, the evidence function can be equivalently defined as $\mathcal{E}' : \mathcal{W} \to \wp(J \times \mathcal{L}_J)$, with $(t, \varphi) \in \mathcal{E}(w)$ indicating that t is relevant/admissible for φ at w. Its constraints become

$$\{(s, \varphi \to \psi), (t, \varphi)\} \subseteq \mathcal{E}(w) \Rightarrow (s{\cdot}t, \psi) \in \mathcal{E}(w) \quad \text{and} \quad (s, \varphi) \in \mathcal{E}(w) \Rightarrow \{(s+t, \varphi), (t+s, \varphi)\} \subseteq \mathcal{E}(w)$$

and thus a justification model $M = (\mathcal{W}, \mathcal{R}, \mathcal{E}, \mathcal{V})$ can be equivalently stated as $M' = (\mathcal{W}, \mathcal{R}, \mathcal{E}', \mathcal{V})$. Finally, for the language, one can simply define $t{:}\varphi := \mathsf{O}(t, \varphi) \wedge \Box\varphi$.

Models for knowing-what Plaza's analysis of the *knowing-what-the-value-of-a-constant-is* operator (*knowing-what* for short; [31]) has played a crucial role in the emerging of a *new generation of epistemic logics* [39] that go beyond standard *knowing-that* modalities. Adding D as a denumerable set of constants (rigid designators) to the framework, a D-model extends a multi-relational model with a function $\mathcal{V}_{\mathsf{D}} : (\mathcal{W} \times \mathsf{D}) \to S$, assigning a value in S to each object in D at each world in $\mathcal{W}$. Syntactically, the language extends the standard modal language with expressions of the form $Kv_i d$ (for $i \in \mathsf{Ag}$ and $d \in \mathsf{D}$), intuitively read as "agent i knows the value of constant d". Semantically, this is the case iff d denotes the same object in all i's epistemically accessible worlds:

$$M, w \models_v Kv_i d \quad \text{iff} \quad \forall u, u' \in \mathcal{W},\ w\mathcal{R}_i u \text{ and } w\mathcal{R}_i u' \text{ imply } \mathcal{V}_{\mathsf{D}}(u, d) = \mathcal{V}_{\mathsf{D}}(u', d).$$

A D-model can be seen as an $\mathbf{O}$-model where $\mathbf{O}$ is the set of tuples $D \times S$, and with each possible world w having a single set $\mathcal{O}(w)$. [1] Moreover, these sets should contain exactly one pair (d, s) for each $d \in \mathsf{D}$. Finally, using the 'owning' operator O, the formula $Kv_i d$ is definable as $Kv_i d := \Diamond_i \mathsf{O}(d, s) \to \Box_i \mathsf{O}(d, s)$.

Deontic logic The *Kanger-Anderson reductionist approach* to deontic logic [28, 1] consists in expressing the *OB* operator 'it is obligatory that' by means of the alethic modality $\Box$ plus a new propositional constant. In Kanger's terms, the propositional constant d has the intuitive meaning 'all normative demands are satisfied' (i.e., the situation is 'ideal'). The $OB\varphi$ operator is defined as $\Box(d \to \varphi)$, and Kanger's system of deontic logic is obtained by adding, to the modal logic K, the axiom $\Diamond d$, which semantically defines *strong seriality*: for any world w there is a v s.t. $w\mathcal{R}v$ and v is ideal. From our perspective, it is natural to interpret $\mathbf{O}$ as representing the set of normative demands. Interestingly, when $\mathbf{O}$ is finite, it is easy to rewrite d as $\bigwedge_{o \in \mathbf{O}} o$ and capture its intended meaning. Indeed, the following holds:

Remark 1 *In the class of* $\mathbf{O}$*-models with* $\mathbf{O}$ *finite, the formula* $\Diamond \bigwedge_{o \in \mathbf{O}} o$ *characterizes strong seriality.* ■

While the original Kanger-Anderson's framework cannot handle contrary-to-duty obligations, further refinements, dating back to [23], allow this. The key idea is to use the $\Diamond$ operator to express betterness as a pre-order among worlds, where $\Diamond\varphi$ means that φ is the case in some world that is at least as good as the actual. As suggested by [37], it is also natural to encode betterness by syntactic means, i.e. via an ordering $\prec$ between formulas, where if $\varphi \prec \psi$ then ψ logically implies φ. Along similar lines, by regarding our objects as normative demands (desirable properties), one can define a betterness ordering as, e.g., $\bigwedge_{o \in S} o \prec \bigwedge_{o \in S'} o$ iff $S \subsetneq S'$, where $S, S' \subseteq \mathbf{O}$, and therefore $\bigwedge_{o \in \mathbf{O}} o$ is the maximal element.

Remark 2 *Under this reading, the formula* $\mathbf{O}o \to \Box\mathbf{O}o$ *characterizes the fact that* $\mathcal{R}$ *is a betterness relation: only worlds that are at least as ideal can be seen. Further,* $\neg\mathbf{O}o \to \Diamond\mathbf{O}o$ *says that all non-ideal worlds failing some normative demand have access to some world satisfying it. Together with* $\mathbf{O}o \to \Box\mathbf{O}o$ *and the axiom* $\Diamond\Diamond p \to \Diamond p$ *for transitivity, this implies strong seriality.* ■

2.1 Some useful/important properties of O-models

Depending on the particular interpretation, an $\mathbf{O}$-model may be asked to satisfy requirements connecting the $\mathcal{O}$-sets with the accessibility relations $\mathcal{R}_i$. This section lists some examples, providing their syntactic characterisations and discussing the settings in which they might be useful/important.

Individual properties We start with the simplest properties relating accessibility relations with objects: those whose formulations involve a single agent. These *individual properties* are summarised in Table 1,

[1] Alternatively, all $\mathcal{O}_i$-sets are the same at each possible world.

$(\mathcal{W},\mathcal{R},\mathcal{O},\mathcal{V})$ **is s.t.**	**iff, for every** $w,u \in \mathcal{W}$,	**Characterising schema**
$\mathcal{R}_i$ preserves $\mathcal{O}_i$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \subseteq \mathcal{O}_i(u)$	$\mathrm{O}_i\, o \to \Box_i \mathrm{O}_i\, o$
$\mathcal{R}_i$ anti-preserves $\mathcal{O}_i$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(u) \subseteq \mathcal{O}_i(w)$	$\neg \mathrm{O}_i\, o \to \Box_i \neg \mathrm{O}_i\, o$
$\mathcal{O}_i$ is invariant under $\mathcal{R}_i$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) = \mathcal{O}_i(u)$	$(\mathrm{O}_i\, o \to \Box_i \mathrm{O}_i\, o) \wedge (\neg \mathrm{O}_i\, o \to \Box_i \neg \mathrm{O}_i\, o)$
$\mathcal{R}_i$ inverts $\mathcal{O}_i$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cap \mathcal{O}_i(u) = \emptyset$	$\mathrm{O}_i\, o \to \Box_i \neg \mathrm{O}_i\, o$
$\mathcal{R}_i$ anti-inverts $\mathcal{O}_i$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cup \mathcal{O}_i(u) = \mathbf{O}$	$\neg \mathrm{O}_i\, o \to \Box_i \mathrm{O}_i\, o$

Table 1: Some individual properties.

with a model $\mathcal{M}$ satisfying an individual property (e.g., preservation of $\mathcal{O}$) iff it satisfies it for every agent $i \in \mathsf{Ag}$. *Preservation* and *anti-preservation* come from awareness logic [20, 25, 34], where they capture the idea of *awareness introspection*. Indeed, if $\mathcal{R}_i$ preserves (anti-preserves) $\mathcal{O}_i$, then agent i's awareness is positively (negatively) introspective: whenever she is (not) aware of something, she knows/believes so. The *invariance* property, the conjunction of preservation and anti-preservation, captures perfect/total awareness introspection. Finally, the *inversion* properties are mathematical variations of the preservation properties: they ask for the accessibility relation to *invert* the 'opinion' of a set towards an object. To the best of our knowledge, none of them has been studied, and yet they can be seen as intuitively appealing in some contexts. For instance, $\mathcal{R}$ inverting $\mathcal{O}$ seems appropriate to talk, in the spirit of [1], about normative violations in a deontic reading of **O**-models: if an agent has a bad habit, then she would prefer not to have it. Analogously, $\mathcal{R}$ anti-inverting $\mathcal{O}$ works well as a formal property for normative demands as those of [28]: if the agent lacks it, then she prefers to have it.

The following proposition states the definability of the listed individual properties in $\mathcal{L}$.

Proposition 1 *Let $\mathbb{P}$ be an individual property (left-hand column of Table 1); let $\Gamma(\mathbb{P})$ be the set of all instances of the corresponding schema in the right-hand column. For any **O**-model $\mathcal{M}$, we have that*

$$\mathcal{M} \textit{ satisfies } \mathbb{P} \quad \textit{iff} \quad \mathcal{M} \models \Gamma(\mathbb{P}). \qquad \blacksquare$$

Group properties These properties express how the set of objects of one agent 'affects'/'influences' the set of objects of other agents in the worlds accessible to the first. As it is explained below, the notion of "a model $\mathcal{M}$ satisfying a group property $\mathbb{P}$" should be parametrised to avoid trivialisations (e.g., all agents are aware of everything). The properties are listed in Table 2, with $f : \mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ a possibly partial function whose domain is non-empty. If $\mathbb{P}$ is a group property, we say that $\mathcal{M}$ *f-satisfies* $\mathbb{P}$ iff for every $i \in Dom(f)$, $\mathcal{M}$ satisfies $\mathbb{P}$ for i and $f(i)$. Moreover, we call *universal (resp. existential) group properties* those that contain "for all" (resp. "for some") in their formulation. Regarding their use, the property of anti-preservation of **O** for everyone in $f(i)$ was first brought up by [35] in the context of epistemic logics for abstract argumentation: if the agent is not aware of an argument, she thinks no one else is. As suggested by [33], this property makes general sense under a *de re* reading of the epistemic possibility of attributing someone else a given item. The remaining versions of preservation and anti-preservation are natural mathematical variations of the first, and it is not difficult to find intuitive readings for them. For instance, in an awareness context, preservation *for all* indicates that each agent i knows/believes that everybody in $f(i)$ is aware of what she is aware of. Analogously, preservation *for some* indicates that each agent i knows/believes that at least someone in $f(i)$ is aware of what she is aware of.

$(\mathcal{W},\mathcal{R},\mathcal{O},\mathcal{V})$ **is s.t.**	**iff, for every** $w,u \in \mathcal{W}$,	**Characterising schema**
$\mathcal{R}_i$ preserves $\mathcal{O}_j$ for all $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \subseteq \bigcap_{j\in f(i)} \mathcal{O}_j(u)$	$\mathrm{O}_i\, o \to \Box_i \bigwedge_{j\in f(i)} \mathrm{O}_j\, o$
$\mathcal{R}_i$ preserves $\mathcal{O}_j$ for some $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \subseteq \bigcup_{j\in f(i)} \mathcal{O}_j(u)$	$\mathrm{O}_i\, o \to \Box_i \bigvee_{j\in f(i)} \mathrm{O}_j\, o$
$\mathcal{R}_i$ anti-preserves $\mathcal{O}_j$ for all $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \bigcup_{j\in f(i)} \mathcal{O}_j(u) \subseteq \mathcal{O}_i(w)$	$\neg \mathrm{O}_i\, o \to \Box_i \bigwedge_{j\in f(i)} \neg \mathrm{O}_j\, o$
$\mathcal{R}_i$ anti-preserves $\mathcal{O}_j$ for some $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \bigcap_{j\in f(i)} \mathcal{O}_j(u) \subseteq \mathcal{O}_i(w)$	$\neg \mathrm{O}_i\, o \to \Box_i \bigvee_{j\in f(i)} \neg \mathrm{O}_j\, o$
$\mathcal{R}_i$ inverts $\mathcal{O}_j$ for all $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cap \bigcup_{j\in f(i)} \mathcal{O}_j(u) = \emptyset$	$\mathrm{O}_i\, o \to \Box_i \bigwedge_{j\in f(i)} \neg \mathrm{O}_j\, o$
$\mathcal{R}_i$ inverts $\mathcal{O}_j$ for some $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cap \bigcap_{j\in f(i)} \mathcal{O}_j(u) = \emptyset$	$\mathrm{O}_i\, o \to \Box_i \bigvee_{j\in f(i)} \neg \mathrm{O}_j\, o$
$\mathcal{R}_i$ anti-inverts $\mathcal{O}_j$ for all $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cup \bigcap_{j\in f(i)} \mathcal{O}_j(u) = \mathbf{O}$	$\neg \mathrm{O}_i\, o \to \Box_i \bigwedge_{j\in f(i)} \mathrm{O}_j\, o$
$\mathcal{R}_i$ anti-inverts $\mathcal{O}_j$ for some $j \in f(i) \subseteq \mathsf{Ag}$	$w\mathcal{R}_i u \Rightarrow \mathcal{O}_i(w) \cup \bigcup_{j\in f(i)} \mathcal{O}_j(u) = \mathbf{O}$	$\neg \mathrm{O}_i\, o \to \Box_i \bigvee_{j\in f(i)} \mathrm{O}_j\, o$

Table 2: Some group properties.

The following proposition justifies the parametrisation of the group properties. In awareness epistemic terms, the first bullet says that, when combined with knowledge (or any other factive epistemic attitude), preservation and anti-preservation together imply that every agent is aware of the same things, and that this is common knowledge among all agents. This is clearly a trivialisation. The second bullet shows that knowledge cannot be combined with the universal group version of inversion or anti-inversion.

Proposition 2 *Let $f_{gen} = \{(i, \mathsf{Ag}) \mid i \in \mathsf{Ag}\}$, let $\mathcal{M}$ be a reflexive* $\mathbf{O}$*-model.*

- *If $\mathcal{M}$ f_{gen}-satisfies universal preservation or anti-preservation, then all agents have available the same objects at each pair of worlds $w, v \in \mathcal{W}$ connected by the transitive closure of $\bigcup_{i\in\mathsf{Ag}} \mathcal{R}_i$.*
- *$\mathcal{M}$ f_{gen}-satisfies neither universal inversion nor universal anti-inversion.* ■

Remark 3 *The individual version of (anti-)preservation and (anti-)inversion properties for $i \in \mathsf{Ag}$ are the group versions (both universal and existential) for $f_{indv} = \{(i, \{i\}) \mid i \in Dom(f)\}$.* ■

Finally, we can characterise the group properties using $\mathcal{L}$.

Proposition 3 *Let $f : \mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ be as described above; let $\mathbb{P}^f_i$ be any of the group properties of the left-hand column of Table 2 (e.g., anti-inversion for i and someone in $f(i)$) and let $\varphi(\mathbb{P}^f_i)$ be its corresponding schema in the right-hand column. Let $\Gamma(\mathbb{P}^f)$ the set of all instances of $\varphi(\mathbb{P}^f_i)$ for all $i \in \mathsf{Ag}$, and let $\mathcal{M}$ be an* $\mathbf{O}$*-model. Then,*

$$\mathcal{M}\ f\text{-satisfies}\ \mathbb{P} \quad \textit{iff} \quad \mathcal{M} \models \Gamma(\mathbb{P}^f). \qquad ■$$

Finally, here is the definition of the class of $\mathbf{O}$-models satisfying a collection of properties.

Definition 3 (Classes of models) *Let $(f_1, \ldots, f_n)$ be a sequence with $f_k : \mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ being a function as described above for every $1 \le k \le n$, and let $(\mathbb{P}_1, \ldots, \mathbb{P}_n)$ be a sequence of group properties. We denote as $\mathfrak{M}(f_1\text{-}\mathbb{P}_1, \ldots, f_n\text{-}\mathbb{P}_n)$ the class of all* $\mathbf{O}$*-models $\mathcal{M}$ s.t. for every k, $\mathcal{M}$ f_k-satisfies $\mathbb{P}_k$.* ◀

2.2 Axiom system

Axiomatizing validities over $\mathfrak{M}^{\mathbf{O}}$ (the class of all $\mathbf{O}$-models) is straightforward, as formulas with the 'owning' modality O_i can be seen as a particular atoms connected to a dedicated valuation function $\mathcal{O}_i$. Since the $\mathcal{O}_i$ sets have no particular requirements, the modal logic axiomatisation is enough.

When the focus is the class of models satisfying a certain collection of properties, additional work is needed; for this, Proposition 3 will be useful. Define the notion of local semantic consequence w.r.t. a given class of models in the standard way [12], denoting it by $\Phi \models_{\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)} \varphi$.

TAUT:	All propositional tautologies	MP:	From φ and $\varphi \to \psi$, infer ψ
K:	$\Box_i(\varphi \to \psi) \to (\Box_i\varphi \to \Box_i\psi)$	NEC:	From φ infer $\Box_i\varphi$

Table 3: The minimal modal logic K.

Definition 4 (Static logics) *The logic* K *is the smallest set containing all instances of the axiom schemas of Table 3 that is moreover closed under both inference rules of the same table. The extension of* K *by* $\Phi \subseteq \mathcal{L}$ *is the smallest set of formulas containing all instances of schemas of Table 3, all formulas in* Φ *and it is closed under both inference rules. Let* $(f_1,\dots,f_n)$ *be a sequence of functions* $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ *as described above, and let* $(\mathbb{P}_1,\dots,\mathbb{P}_n)$ *be a sequence of group properties. Then, we denote by* $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$ *the extension of* K *with* $\bigcup_{1\leq k\leq n}\Gamma(\mathbb{P}_k^{f_k})$*.*[2] *Note that when* $n=0$, $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n) = \mathsf{K}$. ◀

The notions of $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$-proof and $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$-deduction from assumption (noted $\Phi \vdash_{\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)} \varphi$), are the standard ones in modal logic (see e.g., [12]).

Theorem 1 (Static completeness) *Let* $(f_1,\dots,f_n)$ *be a sequence of functions* $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ *as described above, and let* $(\mathbb{P}_1,\dots,\mathbb{P}_n)$ *be a sequence of group properties, we have that:*

$\mathsf{L}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$ *is sound and strongly complete with respect to* $\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$. ■

3 Dynamics of O-models, semantically

Changes in different modal attitudes (knowledge, beliefs, preferences and so on) have been the main topic of DEL. The main feature that distinguishes DEL from other approaches for modelling dynamics (e.g., propositional dynamic logic [24] or automata theory [27]) is that changes are not represented as (binary) relations, but rather as operations that modify the underlying semantic structure. Indeed, DEL can be understood, more broadly, as the study of modal logics of model change [7]. Here we focus on the *event models* of [6, 5]: structures that, when 'applied' to a relational model (by means of a *product update*), produce another relational model. They were initially introduced as a way of modelling non-public acts of communication, and have since then been widely employed to model other forms of informational and factual changes [8, 18, 11, 15]. Besides their versatility, they have an important technical advantage: as proved in [18], any pointed relational model can be turned into any other by means of the product update with some event model that allows factual change.[3] The rest of this section will discuss an extension of these event models that works for describing dynamics of **O**-models.

Definition 5 (Event O-Model) *An* event **O**-model *is a tuple* $\mathcal{E} = (\mathcal{S},\mathcal{T},\mathsf{pre},\mathsf{eff})$ *where* $\mathcal{S} \neq \emptyset$ *is a finite set of events,* $\mathcal{T} : \mathsf{Ag} \to \wp(\mathcal{S}\times\mathcal{S})$ *assigns to each agent a binary relation,* $\mathsf{pre} : \mathcal{S} \to \mathcal{L}$ *assigns a precondition to each event, and* $\mathsf{eff} : (\mathsf{Ag}\times\{+,-\}\times\mathcal{S}) \to \wp(\mathbf{O})$ *is a function indicating, for each event, its (positive and negative) effects on the set of objects available to each agent (write* $\mathsf{eff}(i,\pm,s)$ *as* $\mathsf{eff}_i^{\pm}(s)$ *for* $\pm \in \{+,-\}$*). We assume that, for every* $s \in \mathcal{S}$ *and every* $i \in \mathsf{Ag}$*, the sets* $\mathsf{eff}_i^{+}(s)$ *and* $\mathsf{eff}_i^{-}(s)$ *are finite and disjoint. Note:* $\mathcal{T}(i)$ *abbreviates* $\mathcal{T}_i$*. The set of events of a given* $\mathcal{E}$ *is referred to as* $\mathcal{S}[\mathcal{E}]$ *(and the same*

[2] See propositions 1 and 3 for the meaning of $\Gamma(\mathbb{P}^f)$.

[3] Slightly more precisely, given pointed models $(\mathcal{M},w)$ and $(\mathcal{M}',w')$, there is 'almost always' an event model such that, when applied to $(\mathcal{M},w)$, produces a pointed model $(\mathcal{M}'',w'')$ that is, from the point of view of the language of propositional dynamic logic [24] (an extension of the basic modal language), indistinguishable from $(\mathcal{M}',w')$. See [18] for details.

$\mathcal{E}$ f**-satisfies**	**iff for every** $i \in Dom(f)$**,** $s,t \in \mathcal{S}[\mathcal{E}]$	**is safe for**
$\mathsf{EMP}^{\mathsf{pres}-\forall}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^+(s) \subseteq \bigcap_{j\in f(i)} \mathsf{eff}_j^+(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^-(t) \subseteq \mathsf{eff}_i^-(s)$	preservation for everyone
$\mathsf{EMP}^{\mathsf{pres}-\exists}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^+(s) \subseteq \bigcup_{j\in f(i)} \mathsf{eff}_j^+(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^-(t) \subseteq \mathsf{eff}_i^-(s)$	preservation for someone
$\mathsf{EMP}^{\mathsf{anti}-\mathsf{pres}-\forall}$	$s\mathcal{T}_i t \Rightarrow \bigcup_{j\in f(i)} \mathsf{eff}_j^+(t) \subseteq \mathsf{eff}_i^+(s)$ and $\mathsf{eff}_i^-(s) \subseteq \bigcap_{j\in f(i)} \mathsf{eff}_j^-(t)$	anti-preservation for everyone
$\mathsf{EMP}^{\mathsf{anti}-\mathsf{pres}-\exists}$	$s\mathcal{T}_i t \Rightarrow \bigcup_{j\in f(i)} \mathsf{eff}_j^+(t) \subseteq \mathsf{eff}_i^+(s)$ and $\mathsf{eff}_i^-(s) \subseteq \bigcup_{j\in f(i)} \mathsf{eff}_j^-(t)$	anti-preservation for someone
$\mathsf{EMP}^{\mathsf{inv}-\forall}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^+(s) \subseteq \bigcap_{j\in f(i)} \mathsf{eff}_j^-(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^+(t) \subseteq \mathsf{eff}_i^-(s)$	inversion for everyone
$\mathsf{EMP}^{\mathsf{inv}-\exists}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^+(s) \subseteq \bigcup_{j\in f(i)} \mathsf{eff}_j^-(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^+(t) \subseteq \mathsf{eff}_i^-(s)$	inversion for someone
$\mathsf{EMP}^{\mathsf{anti}-\mathsf{inv}-\forall}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^-(s) \subseteq \bigcap_{j\in f(i)} \mathsf{eff}_j^+(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^-(t) \subseteq \mathsf{eff}_i^+(s)$	anti-inversion for everyone
$\mathsf{EMP}^{\mathsf{anti}-\mathsf{inv}-\exists}$	$s\mathcal{T}_i t \Rightarrow \mathsf{eff}_i^-(s) \subseteq \bigcup_{j\in f(i)} \mathsf{eff}_j^+(t)$ and $\bigcup_{j\in f(i)} \mathsf{eff}_j^-(t) \subseteq \mathsf{eff}_i^+(s)$	anti-inversion for someone

Table 4: Properties of event **O**-models.

convention applies for the other components of $\mathcal{E}$). We use infix notation for each $\mathcal{T}_i$. A pointed event **O**-model *is a tuple $(\mathcal{E}, s)$ with $\mathcal{E} = (\mathcal{S}, \mathcal{T}, \mathsf{pre}, \mathsf{eff})$ an event* **O***-model and $s \in \mathcal{S}[\mathcal{E}]$.* ◀

The above definition does not include the *post-condition function* (see, e.g., [15, 16]), as we want to focus on non-factual changes (i.e., changes on accessibility relations and **O**-sets, but not on atomic valuations). We think, however, that incorporating them does not pose any challenge, since our framework can in fact be seen as a variation of event models for factual change, where one deals with agent-indexed predicates instead of purely atomic propositions.

Definition 6 (Product update) *Let $\mathcal{M} = (\mathcal{W}, \mathcal{R}, \mathcal{O}, \mathcal{V})$ be an* **O***-model and let $\mathcal{E} = (\mathcal{S}, \mathcal{T}, \mathsf{pre}, \mathsf{eff})$ be an event* **O***-model. The* product update *of $\mathcal{M}$ and $\mathcal{E}$ produces the model $\mathcal{M} \otimes \mathcal{E} = (\mathcal{W}', \mathcal{R}', \mathcal{O}', \mathcal{V}')$, where:*

- $\mathcal{W}' := \{(w,s) \in \mathcal{W} \times \mathcal{S} \mid \mathcal{M}, w \models \mathsf{pre}(s)\}$.
- $\mathcal{R}'_i := \{\big((w,s),(u,t)\big) \in \mathcal{W}' \times \mathcal{W}' \mid w\mathcal{R}_i u \,\&\, s\mathcal{T}_i t\}$.
- $\mathcal{O}'_i(w,s) := \big(\mathcal{O}_i(w) \cup \mathsf{eff}_i^+(s)\big) \setminus \mathsf{eff}_i^-(s)$.
- $\mathcal{V}'(p) := \{(w,s) \in \mathcal{W}' \mid w \in \mathcal{V}(p)\}$.

Note: $\mathcal{W}'$ is empty (and thus $\mathcal{M} \otimes \mathcal{E}$ is not defined) when no possible world satisfies any precondition. Thus, $\otimes$ is a partial function. When $\mathcal{W}' \neq \emptyset$, we say that $\mathcal{M} \otimes \mathcal{E}$ is defined. ◀

The closure problem Given a class of **O**-models $\mathfrak{M}$, the *closure* problem [3, 4] asks to find a class of event **O**-models $\mathfrak{E} \neq \emptyset$ s.t., $\mathcal{M} \in \mathfrak{M}$ and $\mathcal{E} \in \mathfrak{E}$ imply $\mathcal{M} \otimes \mathcal{E} \in \mathfrak{M}$. This is not trivial for the properties in Tables 1 and 2: it is clear that executing certain event **O**-models in certain **O**-models leads to the loss of, e.g., individual preservation. This paper focusses on group properties (Remark 3), using $\mathsf{EMP}(\mathbb{P})$ for referring to the event-model property in Table 4 that corresponds to the group property $\mathbb{P}$ in Table 2.[4]

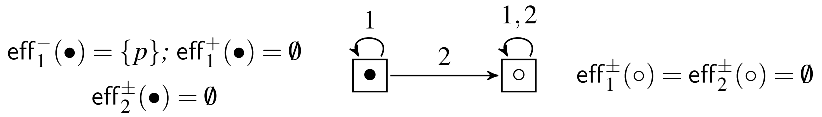
Definition 7 (Classes of event models) *Let $(f_1, \dots, f_n)$ be a sequence of functions $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ as described above, and let $(\mathsf{EMP}_1, \dots, \mathsf{EMP}_n)$ be a sequence of group properties for event models (Table 4). We denote as $\mathfrak{E}(f_1\text{-}\mathsf{EMP}_1, \dots, f_n\text{-}\mathsf{EMP}_n)$ the class of all event* **O***-models $\mathcal{E}$ s.t. for every $1 \leq k \leq n$, $\mathcal{E}$ f_k-satisfies EMP_k.* ◀

With properties of event models defined, here is the main result.

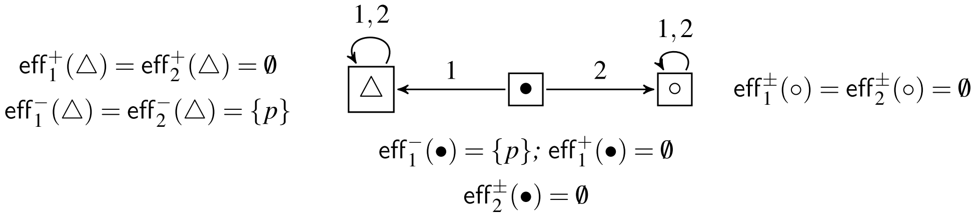
Theorem 2 (Closure for group properties) *Let $f : \mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ be as described above. Let $\mathcal{M}$ be an* **O***-model and $\mathcal{E}$ an event* **O***-model s.t. $\mathcal{M} \otimes \mathcal{E}$ is defined. For any property $\mathbb{P}$ in Table 2, if $\mathcal{M}$ f-satisfies $\mathbb{P}$ and $\mathcal{E}$ f-satisfies $\mathsf{EMP}(\mathbb{P})$, then $\mathcal{M} \otimes \mathcal{E}$ f-satisfies $\mathbb{P}$.* ■

[4] For instance, if $\mathbb{P}$ is anti-inversion for someone, then $\mathsf{EMP}(\mathbb{P}) = \mathsf{EMP}^{\mathsf{anti}-\mathsf{inversion}-\exists}$.

Example 1 (Different forms of forgetting) Theorem 2 helps to test the compatibility between the model of a notion/concept and the model of its dynamics: a single action might be modelled by different event models, and the choice might depend on the specific model requirements. As an example, and in the awareness context, consider an action through which agent i becomes unaware of the atom p without anybody else noticing it. In [11], this action corresponds to the event model $\mathsf{Pri}_i^p = (\{\bullet,\circ\},\mathcal{T},\mathsf{pre},\mathsf{eff})$ with $\mathcal{T}_i = \{(\bullet,\bullet),(\circ,\circ)\}$ and $\mathcal{T}_j = \{(\bullet,\circ),(\circ,\circ)\}$ for $j \neq i$, and with $\mathsf{eff}_i^-(\bullet) = \{p\}$ and $\mathsf{eff}_j^\pm(\bullet) = \mathsf{eff}_j^\pm(\circ) = \mathsf{eff}_i^\pm(\circ) = \emptyset$. When $\mathsf{Ag} = \{1,2\}$ and $i = 1$, the event model can be represented as


$\mathsf{eff}_1^-(\bullet) = \{p\}$; $\mathsf{eff}_1^+(\bullet) = \emptyset$
$\mathsf{eff}_2^\pm(\bullet) = \emptyset$
1
1,2
2
$\mathsf{eff}_1^\pm(\circ) = \mathsf{eff}_2^\pm(\circ) = \emptyset$

This event model does the job when awareness is not required to have special properties.[5] However, it is not appropriate, e.g., when $\mathcal{R}$ is required to f_{gen}-anti-preserve $\mathcal{O}$ for everyone, for $f_{gen} = \{(i,\mathsf{Ag}) \mid i \in \mathsf{Ag}\}$ (as in the case of awareness of arguments of [35, 33]). Fortunately, there is another event model that captures the central intuition of the action (that is, that agent 1 privately looses awareness of p and she is the only one suffering this loss in the actual event $\bullet$) while also preserving the property. Indeed, consider $\mathsf{AIPri}_i^p = (\{\bullet,\circ,\triangle\},\mathcal{T},\mathsf{pre},\mathsf{eff})$ with $\mathcal{T}_i = \{(\bullet,\triangle),(\triangle,\triangle),(\circ,\circ)\}$ and $\mathcal{T}_j = \{(\bullet,\circ),(\triangle,\triangle),(\circ,\circ)\}$ for $j \neq i$, and with $\mathsf{eff}_i^-(\bullet) = \mathsf{eff}_i^-(\triangle) = \mathsf{eff}_j^-(\triangle) = \{p\}$ and $\mathsf{eff}_i^+(\bullet) = \mathsf{eff}_j^\pm(\bullet) = \mathsf{eff}_i^+(\triangle) = \mathsf{eff}_j^+(\triangle) = \mathsf{eff}_i^\pm(\circ) = \mathsf{eff}_j^\pm(\circ) = \emptyset$. When $\mathsf{Ag} = \{1,2\}$ and $i = 1$, the event model is


1,2
1,2
$\mathsf{eff}_1^+(\triangle) = \mathsf{eff}_2^+(\triangle) = \emptyset$
$\mathsf{eff}_1^-(\triangle) = \mathsf{eff}_2^-(\triangle) = \{p\}$
1
2
$\mathsf{eff}_1^\pm(\circ) = \mathsf{eff}_2^\pm(\circ) = \emptyset$
$\mathsf{eff}_1^-(\bullet) = \{p\}$; $\mathsf{eff}_1^+(\bullet) = \emptyset$
$\mathsf{eff}_2^\pm(\bullet) = \emptyset$

Just as before, agent 1 drops p (the effect of $\bullet$), and this change is private, since 2 believes that nothing happened ($\circ$). Additionally, and due to the nature of universal anti-preservation, 1 thinks that everyone loses awareness of p as well (the effects of $\triangle$). Note, moreover, that AIPri_i^p f_{gen}-satisfies $\mathsf{EMP}^{\mathsf{anti-inv-}\forall}$ (our sufficient condition for the preservation of universal anti-preservation). ◀

4 Dynamics of O-models, syntactically

Here is the language used to describe the effect of product updates.

Definition 8 (Language $\mathcal{L}(\star)$) Let $\mathfrak{E}^{\mathbf{O}}$ the class of all event $\mathbf{O}$-models, and let $\star \subseteq \mathfrak{E}^{\mathbf{O}}$ be a non-empty subclass. The dynamic language $\mathcal{L}(\star)$ is given by

$$\varphi ::= \top \mid p \mid \mathrm{O}_i o \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_i \varphi \mid [\mathcal{E},s]\varphi$$

with $p \in \mathsf{At}$, $i \in \mathsf{Ag}$, $o \in \mathbf{O}$, $\mathcal{E} \in \star$ and $s \in \mathcal{S}[\mathcal{E}]$. The truth clause for the new kinds of formulas is:

$$\mathcal{M}, w \models [\mathcal{E},s]\varphi \quad \textit{iff} \quad \mathcal{M}, w \models \mathsf{pre}(s) \textit{ implies } \mathcal{M} \otimes \mathcal{E}, (w,s) \models \varphi. \qquad ◀$$

Definition 9 (Dynamic logics) Let $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ be a static logic of Definition 4. The logic $\mathsf{L}^!(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ extends $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ with all axioms and rules of Table 5 that can be written in $\mathcal{L}(\mathfrak{E}(f_1\text{-}\mathsf{EMP}(\mathbb{P}_1),\ldots,f_n\text{-}\mathsf{EMP}(\mathbb{P}_n)))$. ◀

[5] It even preserves the individual version of invariance (Table 1).

$[\mathcal{E},s]\top \leftrightarrow \top$	$[\mathcal{E},s]\neg\varphi \leftrightarrow (\mathsf{pre}(s) \to \neg[\mathcal{E},s]\varphi)$
$[\mathcal{E},s]p \leftrightarrow (\mathsf{pre}(s) \to p)$	$[\mathcal{E},s](\varphi \wedge \psi) \leftrightarrow ([\mathcal{E},s]\varphi \wedge [\mathcal{E},s]\psi)$
$[\mathcal{E},s]\mathbf{O}_i x \leftrightarrow (\mathsf{pre}(s) \to \mathbf{O}_i x)$ for $x \notin \mathsf{eff}[\mathcal{E}]^+_i(s) \cup \mathsf{eff}[\mathcal{E}]^-_i(s)$	$[\mathcal{E},s]\Box_i\varphi \leftrightarrow (\mathsf{pre}(s) \to \bigwedge_{s\mathcal{T}_i t} \Box_i[\mathcal{E},t]\varphi)$
$[\mathcal{E},s]\mathbf{O}_i x \leftrightarrow \top$ for $x \in \mathsf{eff}[\mathcal{E}]^+_i(s)$	
$[\mathcal{E},s]\mathbf{O}_i x \leftrightarrow \neg\mathsf{pre}(s)$ for $x \in \mathsf{eff}[\mathcal{E}]^-_i(s)$	From $\varphi \leftrightarrow \psi$, infer $\delta \leftrightarrow \delta[\varphi/\psi]$

Table 5: Reduction axioms for dynamic modalities. $\delta[\varphi/\psi]$ is the result of substituting one or more occurrences of φ in δ by ψ. Furthermore, it is assumed for simplicity that these substitutions do not affect the occurrences of formulas inside dynamic modalities, i.e. $([\mathcal{E},s]\delta)[\varphi/\psi] = [\mathcal{E},s](\delta[\varphi/\psi])$.

Then, the completeness result.

Theorem 3 (Dynamic completeness) *Let* $(f_1,\dots,f_n)$ *be a sequence of functions* $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ *as described above, and let* $(\mathbb{P}_1,\dots,\mathbb{P}_n)$ *be a sequence of group properties. We have that:*

$\mathsf{L}^!(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$ *is sound and strongly complete with respect to* $\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\dots,f_n\text{-}\mathbb{P}_n)$. ■

5 Conclusion and future work

The paper provides an abstract look at awareness epistemic models, understanding them not as a representation of the formulas the agents are aware of, but rather as a more general setting for dealing with a notion of ‘owning abstract objects’. As discussed in Section 2, several well-know proposals from different areas can be seen as particular instances of these general type of structures.

When modelling specific phenomena, a general **O**-structure may be asked to satisfy specific requirements. Of particular interest are those that relate $\mathcal{O}$-sets with accessibility relations, and Subsection 2.1 listed some possibilities, together with their characterising formula. Maybe more importantly, these requirements should be preserved by model operations representing dynamics of the modelled phenomena. Section 3 focussed on model operations defined in terms of event models, introducing classes of the latter that, under the product update operation, guarantee the preservation of the specified requirements. This establishes a form of ‘compatibility’ between the represented phenomena and the chosen event models. Section 4 closed the discussion, obtaining complete axiomatisations via reduction axioms.

There are branches open for further exploration; here are two of them. The first one is to work out the details of the instantiations of **O**-models that were sketched in Section 2. The second one is to study more systematically the trivialisation of awareness ($\mathcal{O}$-sets) for extreme cases of f (e.g., for f_{gen}) so as to underpin our notion of f-satisfiability.

References

[1] Alan Ross Anderson (1958): *A reduction of deontic logic to alethic modal logic*. *Mind* 67(265), pp. 100–103, doi:10.1093/mind/LXVII.265.100.

[2] Sergei Artemov & Melvin Fitting (2016): *Justification Logic*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

[3] Guillaume Aucher (2008): *Consistency preservation and crazy formulas in BMS*. In S. Hölldobler, C. Lutz & H. Wansing, editors: *European Workshop on Logics in Artificial Intelligence*, *LNCS* 5293, Springer, pp. 21–33, doi:10.1007/978-3-540-87803-2_4.

[4] Philippe Balbiani, Hans van Ditmarsch, Andreas Herzig & Tiago De Lima (2012): *Some Truths Are Best Left Unsaid.* In T. Bolander, T. Braüner, S. Ghilardi, & L. Moss, editors: *Advances in modal logic*, 9, College Publication, pp. 36–54.

[5] Alexandru Baltag & Lawrence S Moss (2004): *Logics for epistemic programs*. *Synthese* 139(2), pp. 165–224, doi:10.1023/B:SYNT.0000024912.56773.5e.

[6] Alexandru Baltag, Lawrence S Moss & Slawomir Solecki (1998): *The logic of common knowledge, public announcements, and private suspicions*. In I. Gilboa, editor: *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, Morgan Kaufmann Publishers, pp. 43–56, doi:10.1007/978-3-319-20451-2_38.

[7] Alexandru Baltag & Bryan Renne (2016): *Dynamic Epistemic Logic*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Winter 2016 edition, Metaphysics Research Lab, Stanford University.

[8] Alexandru Baltag & Sonja Smets (2008): *A qualitative theory of dynamic interactive belief revision*. In Wiebe van der Hoek, Giacomo Bonanno & Michael Wooldridge, editors: *Logic and the foundations of game and decision theory (LOFT 7)*, *Texts in Logic and Games* 3, Amsterdam University Press, pp. 9–58, doi:10.1007/978-3-319-20451-2_39.

[9] Johan van Benthem (2011): *Logical dynamics of information and interaction*. Cambridge University Press, doi:10.1017/CBO9780511974533.

[10] Johan van Benthem, Jan van Eijck & Barteld Kooi (2006): *Logics of communication and change*. *Information and computation* 204(11), pp. 1620–1662, doi:10.1016/j.ic.2006.04.006.

[11] Johan van Benthem & Fernando R Velázquez-Quesada (2010): *The dynamics of awareness*. *Synthese* 177(1), pp. 5–27, doi:10.1007/s11229-010-9764-9.

[12] Patrick Blackburn, Maarten De Rijke & Yde Venema (2002): *Modal Logic*. Cambridge University Press, doi:10.1017/CBO9781107050884.

[13] Alfredo Burrieza & Antonio Yuste-Ginel (2020): *Basic beliefs and argument-based beliefs in awareness epistemic logic with structured arguments*. In Henry Prakken, Stefano Bistarelli, Francesco Santini & Carlo Taticchi, editors: *Proceedings of the COMMA 2020*, IOS Press, pp. 123–134, doi:10.3233/FAIA200498.

[14] Alfredo Burrieza & Antonio Yuste-Ginel (2021): *An Awareness Epistemic Framework for Belief, Argumentation and Their Dynamics*. In Joseph Y. Halpern & Andrés Perea, editors: *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, *EPTCS* 335, pp. 69–83, doi:10.4204/EPTCS.335.6.

[15] Hans van Ditmarsch, Tim French & Fernando R. Velázquez-Quesada (2012): *Action models for knowledge and awareness*. In Wiebe van der Hoek, Lin Padgham, Vincent Conitzer & Michael Winikoff, editors: *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, IFAAMAS, pp. 1091–1098.

[16] Hans van Ditmarsch, Tim French, Fernando R Velázquez-Quesada & Yì N Wáng (2013): *Knowledge, awareness, and bisimulation*. In Burkhard C. Schipper, editor: *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, doi:10.48550/arXiv.1310.6410.

[17] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic epistemic logic*. Springer, doi:10.1007/978-1-4020-5839-4.

[18] Hans van Ditmarsch & Barteld Kooi (2008): *Semantic results for ontic and epistemic change*. In Wiebe van der Hoek, Giacomo Bonanno & Michael Wooldridge, editors: *Logic and the foundations of game and decision theory (LOFT 7)*, *Texts in Logic and Games* 3, Amsterdam University Press, pp. 9–58, doi:10.48550/arXiv.cs/0610093.

[19] Phan Minh Dung (1995): *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. *Artificial Intelligence* 77(2), pp. 321–357, doi:10.1016/0004-3702(94)00041-X.

[20] Ronald Fagin & Joseph Y Halpern (1987): *Belief, awareness, and limited reasoning*. *Artificial intelligence* 34(1), pp. 39–76, doi:10.1016/0004-3702(87)90003-8.

[21] Davide Grossi & Fernando R. Velázquez-Quesada (2015): *Syntactic awareness in logical dynamics*. Synthese 192(12), pp. 4071–4105, doi:10.1007/s11229-015-0733-1.

[22] Joseph Y Halpern (2001): *Alternative semantics for unawareness*. Games and Economic Behavior 37(2), pp. 321–339, doi:10.1006/game.2000.0832.

[23] Bengt Hansson (1969): *An analysis of some deontic logics*. Nous, pp. 373–398, doi:10.1007/978-94-010-3146-2_5.

[24] David Harel, Dexter Kozen & Jerzy Tiuryn (2001): *Dynamic logic*. In Dov M. Gabbay & Franz Guenthner, editors: Handbook of philosophical logic, 4, Springer, Dordrecht, pp. 99–217, doi:10.1007/978-94-017-0456-4_2.

[25] Aviad Heifetz, Martin Meier & Burkhard C Schipper (2006): *Interactive unawareness*. Journal of economic theory 130(1), pp. 78–94, doi:10.1016/j.jet.2005.02.007.

[26] Jaakko Hintikka (1962): *Knowledge and belief: an introduction to the logic of the two notions*. Cornell University Press.

[27] John E. Hopcroft, Rajeev Motwani & Jeffrey D. Ullman (2003): *Introduction to automata theory, languages, and computation - international edition (2. ed)*. Addison-Wesley.

[28] Stig Kanger (1970): *New foundations for ethical theory (1957). Reprinted*. In . R Hilpinen, editor: Deontic logic: introductory and systematic readings, Springer, doi:10.1007/978-94-010-3146-2_2.

[29] Barteld Kooi (2007): *Expressivity and completeness for public update logics via reduction axioms*. Journal of Applied Non-Classical Logics 17(2), pp. 231–253, doi:10.3166/jancl.17.231-253.

[30] Sanjay Modgil & Henry Prakken (2013): *A general account of argumentation with preferences*. Artificial Intelligence 195, pp. 361–397, doi:10.1016/j.artint.2012.10.008.

[31] Jan Plaza (1989): *Logics of public communications*. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic & Z.W. Ras, editors: Proceedings 4th International Symposium on Methodologies for Intelligent Systems, Oak Ridge National Laboratory, pp. 201–216, doi:10.1007/s11229-007-9168-7.

[32] Carlo Proietti & Antonio Yuste-Ginel (2020): *Persuasive Argumentation and Epistemic Attitudes*. In Luís Soares Barbosa & Alexandru Baltag, editors: Dynamic Logic. New Trends and Applications, LNCS 12005, Springer, pp. 104–123, doi:10.1007/978-3-030-38808-9_7.

[33] Carlo Proietti & Antonio Yuste-Ginel (2021): *Dynamic epistemic logics for abstract argumentation*. Synthese 199(3-4), pp. 8641–8700, doi:10.1007/s11229-021-03178-5.

[34] Burkhard C. Schipper (2015): *Awareness*. In Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek & Barteld Kooi, editors: Handbook of Epistemic Logic, London: College Publications, pp. 77–146, doi:10.2139/ssrn.2401352.

[35] François Schwarzentruber, Srdjan Vesic & Tjitze Rienstra (2012): *Building an Epistemic Logic for Argumentation*. In Luis Fariñas del Cerro, Andreas Herzig & Jérôme Mengin, editors: Logics in Artificial Intelligence, LNCS 7519, Springer, pp. 359–371, doi:10.1007/978-3-642-33353-8_28.

[36] Matthias Thimm (2014): *Strategic argumentation in multi-agent systems*. KI-Künstliche Intelligenz 28(3), pp. 159–168, doi:10.1007/s13218-014-0307-2.

[37] Johan Van Benthem, Davide Grossi & Fenrong Liu (2014): *Priority structures in deontic logic*. Theoria 80(2), pp. 116–152, doi:10.1111/theo.12028.

[38] Fernando Raymundo Velázquez-Quesada (2010): *Small steps in dynamics of information*. Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

[39] Yanjing Wang (2018): *Beyond Knowing That: A New Generation of Epistemic Logics*. In Hans van Ditmarsch & Gabriel Sandu, editors: Jaakko Hintikka on Knowledge and Game-Theoretical Semantics, Outstanding Contributions to Logic 12, Springer International Publishing, pp. 499–533, doi:10.1007/978-3-319-62864-6_21.

[40] Yanjing Wang & Qinxiang Cao (2013): *On axiomatizations of public announcement logic*. Synthese 190(1), pp. 103–134, doi:10.1007/s11229-012-0233-5.

Appendix

Theorem 1 Let $(f_1,\ldots,f_n)$ be a sequence of functions $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ as described above, and let $(\mathbb{P}_1,\ldots,\mathbb{P}_n)$ be a sequence of group properties, we have that:

$$\mathsf{L}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n) \text{ is sound and strongly complete with respect to } \mathfrak{M}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n).$$

Proof. Let $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ and $\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ be arbitrarily fixed from now on, we drop the parameters $(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ for readability.

Soundness follows by induction for the length of L-proofs. For the basic step, one needs to show that every instance of an L-axiom schema is valid in the corresponding class of models. For the inductive step, it is enough to show that both inference rules preserve $\mathfrak{M}$-validity.

As for completeness, the proof follows a canonical model argument. We denote by MC^{L} the class of all maximally L-consistent sets of formulas. The proofs of the Lindenbaum lemma, as well as the closure properties of maximally L-consistent sets, are as usual. The L-canonical model is the **O**-model $\mathcal{M}^{\mathsf{L}} = (\mathcal{W}^{\mathsf{L}}, \mathcal{R}^{\mathsf{L}}, \mathcal{O}^{\mathsf{L}}, \mathcal{V}^{\mathsf{L}})$ where each component is defined as follows:

$$\begin{array}{rl}
\mathcal{W}^{\mathsf{L}} = & \mathsf{MC}^{\mathsf{L}}, \\
\Phi\mathcal{R}^{\mathsf{L}}_i\Delta \quad \text{iff} & \{\varphi \in \mathcal{L} \mid \Box_i\varphi \in \Phi\} \subseteq \Delta, \\
\mathcal{O}^{\mathsf{L}}_i(\Phi) = & \{x \in \mathbf{O} \mid \mathbf{O}_i x \in \Phi\}, \text{ and} \\
\mathcal{V}^{\mathsf{L}}(p) = & \{\Phi \in \mathcal{W}^{\mathsf{L}} \mid p \in \Phi\}.
\end{array}$$

The proof of the Truth Lemma ($\forall\varphi \in \mathcal{L}$, $\varphi \in \Phi$ iff $\mathcal{M}^{\mathsf{L}}, \Phi \models \varphi$) is by induction on the structure of φ. The only difference w.r.t. the proof of the lemma for basic modal logic is the step where $\varphi = \mathbf{O}_i x$, and this is straightforward. For the right-to-left direction of the case $\varphi = \Box_i\psi$, one needs to show that the so-called Existence Lemma holds, namely, that if $\neg\Box_i\delta \in \Phi(\in \mathcal{W}^{\mathsf{L}})$, then there is a $\Delta \in \mathcal{W}^{\mathsf{L}}$ with $\Phi\mathcal{R}^{\mathsf{L}}_i\Delta$ and $\neg\delta \in \Delta$. The final, crucial part is to show that all group properties are canonical, that is to say, if φ^f is an L-axiom schema that defines a group property $\mathbb{P}$, then $\mathcal{M}^{\mathsf{L}}$ f-satisfies $\mathbb{P}$. We leave details for the reader.

As a curiosity, note that if **O**-formulas are considered as special types of atoms (as done, e.g., in [33]), our logic is not normal in the sense of [12], because the rule of uniform substitution does not preserve validity in all classes of models. However, this does not affect the completeness argument. ■

Theorem 2 Let $f : \mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ be as described above. Let $\mathcal{M}$ be an **O**-model and $\mathcal{E}$ an event **O**-model s.t. $\mathcal{M} \otimes \mathcal{E}$ is defined. For any property $\mathbb{P}$ in Table 2, if $\mathcal{M}$ f-satisfies $\mathbb{P}$ and $\mathcal{E}$ f-satisfies $\mathsf{EMP}(\mathbb{P})$, then $\mathcal{M} \otimes \mathcal{E}$ f-satisfies $\mathbb{P}$.

Proof. For space reasons, we just show that the theorem holds for the first and the last property. The rest of the cases are left for the reader:
[$\mathbb{P}$ = preservation for everyone] Take $\mathcal{M}$ and $\mathcal{E}$ s.t.

$$\mathcal{M} \ f\text{-satisfies preservation for everyone} \tag{1}$$

$$\mathcal{E} \text{ f-satisfies } \mathsf{EMP}^{\mathsf{pres}-\forall} \tag{2}$$

We want to show that $\mathcal{M} \otimes \mathcal{E} = (\mathcal{W}', \mathcal{R}', \mathcal{V}', \mathcal{O}')$ f-satisfies preservation for everyone. Let $i \in Dom(f)$ and $(w,s) \in \mathcal{W}'$ and suppose that $(w,s)\mathcal{R}'_i(u,t)$. This is equivalent by the definition of product update to

$$w\mathcal{R}_i u \quad \text{and} \quad s\mathcal{T}_i t \tag{3}$$

Further, suppose that $x \in \mathcal{O}'_i(w,s)$, which is equivalent, by the definition of product update, to $x \in \big(\mathcal{O}_i(w) \cup \mathsf{eff}^+_i(s)\big) \setminus \mathsf{eff}^-_i(s)$. We continue by cases on the membership of x, showing that $x \in \bigcap_{j\in f(i)} \mathcal{O}'_j(u,t)$ always obtains.

Case: $x \in \mathcal{O}_i(w)$ **and** $x \notin \mathsf{eff}^-_i(s)$**.** On the one hand, $x \in \mathcal{O}_i(w)$ implies together with (1) and (3) that $x \in \bigcap_{j\in f(i)} \mathcal{O}_j(u)$. On the other hand, $x \notin \mathsf{eff}^-_i(s)$ implies together with (2) and (3) that $x \notin \bigcup_{j\in f(i)} \mathsf{eff}^-_j(t)$. Both facts imply by set-theoretic reasoning that $x \in \bigcap_{j\in f(i)} \big((\mathcal{O}_j(u) \cup \mathsf{eff}^+_j(t)) \setminus \mathsf{eff}^-_j(t)\big)$, which is equivalent to what we wanted to show (by definition of product update).

Case: $x \in \mathsf{eff}^+_i(s)$ **and** $x \notin \mathsf{eff}^-_i(s)$**.** The latter implies, together with (2) and (3), that $x \in \bigcap_{j\in f(i)} \mathsf{eff}^+_j(t)$ and $x \notin \bigcup_{j\in f(i)} \mathsf{eff}^-_j(t)$, which implies by set-theoretic reasoning that $x \in \bigcap_{j\in f(i)} \big((\mathcal{O}_j(u) \cup \mathsf{eff}^+_j(t)) \setminus \mathsf{eff}^-_j(t)\big)$, which is equivalent to what we wanted to show (by definition of product update).

[$\mathbb{P}$ = anti-inversion for someone] Take $\mathcal{M}$ and $\mathcal{E}$ s.t.

$$\begin{aligned} \mathcal{M}\ f\text{-satisfies anti-inversion for someone} \quad (1)\\ \mathcal{E}\ \text{f-satisfies}\ \mathsf{EMP}^{\mathsf{anti-inv-}\exists} \quad (2)\end{aligned}$$

We want to show that $\mathcal{M} \otimes \mathcal{E} = (\mathcal{W}', \mathcal{R}', \mathcal{V}', \mathcal{O}')$ f-satisfies anti-inversion for someone. Let $i \in Dom(f)$ and $(w,s) \in \mathcal{W}'$ and suppose that $(w,s)\mathcal{R}'_i(u,t)$. This is equivalent by the definition of product update to

$$w\mathcal{R}_i u \quad \text{and} \quad s\mathcal{T}_i t \quad (3).$$

Further, suppose that $x \notin \mathcal{O}'_i(w,s)$, which is equivalent, by the definition of product update, to $x \notin \mathcal{O}_i(w) \cup \mathsf{eff}^+_i(s) \setminus \mathsf{eff}^-_i(s)$. We want to show $x \in \bigcup_{j\in f(i)} \mathcal{O}'_j(u,t)$, which is equivalent to $x \in \bigcup_{j\in f(i)} (\mathcal{O}_j(u) \cup \mathsf{eff}^+_j(t)) \setminus \mathsf{eff}^-_j(t)$. We continue by cases on $x \notin \mathcal{O}_i(w) \cup \mathsf{eff}^+_i(s) \setminus \mathsf{eff}^-_i(s)$, showing that the latter claim always obtains.

Case: $x \notin \mathcal{O}_i(w)$ **and** $x \notin \mathsf{eff}^+_i(s)$**.** On the one hand, $x \notin \mathcal{O}_i(w)$ implies together with (1) and (3) that $x \in \bigcup_{j\in f(i)} \mathcal{O}_j(u)$. On the other hand, $x \notin \mathsf{eff}^+_i(s)$ implies together with (2) and (3) that $x \notin \bigcup_{j\in f(i)} \mathsf{eff}^-_j(t)$. Both facts imply by set-theoretic reasoning that $x \in \bigcup_{j\in f(i)} (\mathcal{O}_j(u) \cup \mathsf{eff}^+_j(t)) \setminus \mathsf{eff}^-_j(t)$.

Case: $x \in \mathsf{eff}^-_i(s)$**.** The latter implies, together with (2) and (3), that $x \in \bigcup_{j\in f(i)} \mathsf{eff}^+_j(t)$ which implies that $x \notin \bigcap_{j\in f(i)} \mathsf{eff}^-_i(t)$ (by definition of event **O**-model, because $\mathsf{eff}^+_k(t) \cap \mathsf{eff}^-_k(t) = \emptyset$ for every $k \in \mathsf{Ag}$). The latter two claims implies by set-theoretical reasoning that $x \in \bigcap_{j\in f(i)} (\mathcal{O}_j(u) \cup \mathsf{eff}^+_j(t)) \setminus \mathsf{eff}^-_j(t)$. ■

Theorem 3 Let $(f_1,\ldots,f_n)$ be a sequence of functions $\mathsf{Ag} \to \wp(\mathsf{Ag}) \setminus \{\emptyset\}$ as described above, and let $(\mathbb{P}_1,\ldots,\mathbb{P}_n)$ be a sequence of group properties. We have that:

$\mathsf{L}^!(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ is sound and strongly complete with respect to $\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$.

Proof. (Sketched) Let $\mathfrak{M}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$, $\mathcal{L}(\mathfrak{E}(f_1\text{-}\mathsf{EMP}(\mathbb{P}_1),\ldots,f_n\text{-}\mathsf{EMP}(\mathbb{P}_n)))$, $\mathsf{L}^!(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$, and $\mathsf{L}(f_1\text{-}\mathbb{P}_1,\ldots,f_n\text{-}\mathbb{P}_n)$ be arbitrarily fixed from now on. We drop the parameters and denote them by $\mathfrak{M}$, $\mathcal{L}(\mathfrak{E})$, $\mathsf{L}^!$, and L for the sake of readability, but note that the parametrisation of each of the components is crucial for our argument.

The soundness of $\mathsf{L}^!$ follows from soundness of its static base L (Theorem 1), the validity of axioms of Table 5, and the validity-preserving character of the only rule present in the same table. For proving the latter, i.e., that the application of the rule preserves validity within $\mathfrak{M}$, Theorem 2 is necessary. In more detail, the validity-preservation of the rule is proven by induction on δ, and Theorem 2 is crucial when we arrive at the step where δ has the shape $[\mathcal{E},s]\alpha$. Moreover, and in the same inductive step, the simplification shown in the caption of Table 5 is needed.

We can then prove strong completeness via a reduction argument (see [29, 17, 10, 40]). For doing so, we use two numeric measures for formulas, the *depth* of φ, noted $d(\varphi)$, and the number of *nested dynamic modalities* in φ, noted $Od(\varphi)$. More formally:

- Define $d : \mathcal{L}(\mathfrak{E}) \to \mathbb{N}$ as $d(p) = 0$ for every $p \in \mathsf{At}$, $d(\mathbf{O}_i x) = 0$ for every $x \in \mathbf{O}$, $i \in \mathsf{Ag}$, $d(\circledast\varphi) = 1 + d(\varphi)$ where $\circledast \in \{\neg, \Box_i, [\mathcal{E}, s]\}$ and $d(\varphi \wedge \psi) = 1 + max(d(\varphi), d(\psi))$.
- Define $Od : \mathcal{L}(\mathfrak{E}) \to \mathbb{N}$ as $Od(p) = 0$, $d(\mathbf{O}_i x) = 0$ for every $x \in \mathbf{O}$, $i \in \mathsf{Ag}$, $Od(\neg\varphi) = Od(\Box_i\varphi) = Od(\varphi)$, $Od(\varphi \wedge \psi) = max(Od(\varphi), Od(\psi))$, and $Od([\mathcal{E}, s]\varphi) = 1 + Od(\varphi)$.

Now, we define the following function, translating formulas from each dynamic language $\mathcal{L}(\mathfrak{E})$ to the its static fragment $\mathcal{L}$:

$$\begin{array}{ll}
\tau(p) = p & \tau([\mathcal{E}, s]p) = \mathsf{pre}(s) \to p \\
\tau(\mathbf{O}_i x) = \mathbf{O}_i x & \tau([\mathcal{E}, s]\mathbf{O}_i x) = \mathsf{pre}(s) \to \mathbf{O}_i x \text{ if } x \notin \mathsf{eff}[\mathcal{E}]_i^+(s) \cup \mathsf{eff}[\mathcal{E}]_i^-(s) \\
 & \tau([\mathcal{E}, s]\mathbf{O}_i x) = \top \text{ if } x \in \mathsf{eff}[\mathcal{E}]_i^+(s) \\
 & \tau([\mathcal{E}, s]\mathbf{O}_i x) = \neg\mathsf{pre}(s) \text{ if } x \in \mathsf{eff}[\mathcal{E}]_i^-(s) \\
\tau(\neg\varphi) = \neg\tau(\varphi) & \tau([\mathcal{E}, s]\neg\varphi) = \mathsf{pre}(s) \to \neg\tau([\mathcal{M}, s]\varphi) \\
\tau(\varphi \wedge \psi) = \tau(\varphi) \wedge \tau(\psi) & \tau([\mathcal{E}, s](\varphi \wedge \psi)) = \tau([\mathcal{E}, s]\varphi) \wedge \tau([\mathcal{E}, s]\psi) \\
\tau(\Box_i\varphi) = \Box_i\tau(\varphi) & \tau([\mathcal{E}, s]\Box_i\varphi) = \mathsf{pre}(s) \to \bigwedge_{s\mathcal{T}_i t} \Box_i\tau([E, t]\varphi) \\
 & \tau([\mathcal{E}, s][\mathcal{F}, s]\varphi) = \tau([\mathcal{E}, s]\tau([\mathcal{F}, s]\varphi))
\end{array}$$

The next step is showing that the co-domain of τ is in fact $\mathcal{L}$. This is proven in two phases. First, one can show that it holds for the special case where $O(\varphi) = 1$, and this is done by induction on $d(\varphi)$. Then it can be proven for the general case (and the previous claim is needed). Note that this translation amounts to what [40] coined as an *inside-out* reduction because, when dealing with a formula δ with more that one nested dynamic operator (i.e., with $Od(\delta) \geq 2$), we first take care of the innermost occurrence due to the definition of τ.

Finally, we can establish and prove the key Reduction Lemma, namely, that for every $\varphi \in \mathcal{L}(\mathfrak{E})$:

$$\vdash_{\mathsf{L}^!} \varphi \leftrightarrow \tau(\varphi).$$

This is done through a complex inductive argument. Again, one first needs to prove the claim for the special case $Od(\varphi) = 1$ by induction on φ. Then, the claim can be proven for the general case. This second proof requires a double induction, first on $d(\varphi)$ and, we arrive at the step $\varphi = [\mathcal{E}, s]\psi$, we continue by induction on $d(\psi)$. Note that the validity-preservation character of the rule of Table 5 is strongly needed for all cases (which in turn requires Theorem 2, as we mentioned). ■

A Logic-Based Analysis of Responsibility

Aldo Iván Ramírez Abarca
Utrecht University
Utrecht, The Netherlands
nadabundo@gmail.com

This paper presents a logic-based framework to analyze responsibility, which I refer to as intentional epistemic act-utilitarian stit theory (IEAUST). To be precise, IEAUST is used to model and syntactically characterize various modes of responsibility, where by 'modes of responsibility' I mean instances of Broersen's three categories of responsibility (causal, informational, and motivational responsibility), cast against the background of particular deontic contexts. IEAUST is obtained by integrating a modal language to express the following components of responsibility on stit models: agency, epistemic notions, intentionality, and different senses of obligation. With such a language, I characterize the components of responsibility using particular formulas. Then, adopting a compositional approach—where complex modalities are built out of more basic ones—these characterizations of the components are used to formalize the aforementioned modes of responsibility.

1 Introduction

The study of responsibility is a complicated matter. The term is used in different ways in different fields, and it is easy to engage in everyday discussions as to why someone should be considered responsible for something. Typically, the backdrop of these discussions involves social, legal, moral, or philosophical problems, each with slightly different meanings for expressions like *being responsible for...*, *being held responsible for...*, or *having the responsibility of...*, among others. Therefore—to approach such problems efficiently—there is a demand for clear, taxonomical definitions of responsibility.

For instance, suppose that you are a judge in Texas. You are presiding over a trial where the defendant is being charged with first-degree murder. The alleged crime is horrible, and the prosecution seeks capital punishment. The case is as follows: driving her car, the defendant ran over a traffic officer that was holding a stop-sign at a crossing walk, while school children were crossing the street. The traffic officer was killed, and some of the children were severely injured. A highly complicated case, the possibility of a death-penalty sentence means that the life of the defendant is at stake. More than ever, due process is imperative. As the presiding judge, you must abide by the prevailing definitions of criminal liability with precision. In other words, there is little to no room for ambiguity in the ruling, and your handling of the notions associated with responsibility in criminal law should be impeccable.

As this example suggests, a framework with intelligible, realistically applicable definitions of responsibility is paramount in the field of law. However, responsibility-related problems arise across many other disciplines—social psychology, philosophy of emotion, legal theory, and ethics, to name a few [17, 24]. A clear pattern in all these is the intent of issuing standards for when—and to what extent—an agent should be held responsible for a state of affairs.

This is where Logic lends a hand. The development of expressive logics—to reason about agents' decisions in situations with moral consequences—involves devising unequivocal representations of components of behavior that are highly relevant to systematic responsibility attribution and to systematic

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 470–486, doi:10.4204/EPTCS.379.36

blame-or-praise assignment. To put it plainly, expressive syntactic-and-semantic frameworks help us analyze responsibility-related problems in a methodical way.[1]

The main goal of this paper is to present a proposal for a formal theory of responsibility. Such a proposal relies on (a) a *decomposition* of responsibility into specific components and (b) a functional *classification* of responsibility, where the different categories directly correlate with the components of the decomposition. As for the decomposition, it is given by the following list:

- **Agency**: the process by which agents bring about states of affairs in the environment. In other words, the phenomenon by which agents choose and perform actions, with accompanying mental states, that change the environment.
- **Knowledge and belief**: mental states that concern the information available in the environment and that explain agents' particular choices of action.
- **Intentions**: mental states that determine whether an action was done with the purpose of bringing about its effects.
- **Ought-to-do's**: the actions that agents should perform, complying to the codes of a normative system. Oughts-to-do's make up contexts that provide a criterion for deciding whether an agent should be blamed or praised. I refer to these contexts as the *deontic contexts* of responsibility.

As for the classification, it is a refinement of Broersen's three categories of responsibility: *causal, informational*, and *motivational* responsibility [4, 9, 14]. I will discuss these categories at length in Section 2. On the basis of both the decomposition and the classification, here I introduce a very rich stit logic to analyze responsibility, which I refer to as *intentional epistemic act-utilitarian stit theory* (*IEAUST*). More precisely, I use *IEAUST* to model and syntactically characterize various modes of responsibility. By 'modes of responsibility' I mean combinations of sub-categories of the three ones mentioned above, cast against the background of particular deontic contexts. On the one hand, the sub-categories correspond to the different versions of responsibility that one can consider according to the *active* and *passive* forms of the notion: while the active form involves contributions—in terms of explicitly bringing about outcomes—the passive form involves omissions—which are interpreted as the processes by which agents allow that an outcome happens while being able, to some extent, to prevent it. On the other hand, the deontic context of a mode establishes whether and to what degree the combination of sub-categories involves either blameworthiness or praiseworthiness.

The logic *IEAUST* includes a language that expresses agency, epistemic notions, intentionality, and different senses of obligation. With this language, I characterize the components of responsibility using particular formulas. Then, adopting a compositional approach—where complex modalities are built out of more basic ones—I use these characterizations of the components to formalize the aforementioned modes of responsibility. An outline of the paper is included below.

- Section 2 presents an operational definition for responsibility and addresses the philosophical perspective adopted in my study of the notion.
- Section 3 introduces *IEAUST* and uses this logic to provide stit-theoretic characterizations of different modes of responsibility.
- Section 4 presents Hilbert-style proof systems both for *IEAUST* and for a technical extension, addressing the status of their soundness & completeness results.

[1] Most likely, this is why the logic-based formalization of responsibility has become such an important topic in, for instance, normative multi-agent systems, responsible autonomous agents, and machine ethics for AI [21, 6]

2 Categories of Responsibility

To make a start on formally analyzing responsibility, I identify (a) two *viewpoints* for the philosophical study of responsibility, (b) three main *categories* for the viewpoint that I focus on, and (c) two *forms* in which the elements of the categories can be interpreted.

As for point (a), the philosophical literature on responsibility usually distinguishes two *viewpoints* on the notion [22]: *backward-looking responsibility* and *forward-looking responsibility*. By backward-looking responsibility one refers to the viewpoint according to which an agent is considered to have produced a state of affairs that has already ensued and lies in the past. This is the viewpoint taken by a judge when, while trying a murder case, she wants to get to the bottom of things and find out who is responsible for doing the killing. In contrast, by forward-looking responsibility one refers to the viewpoint according to which which an agent is expected to comply with the duty of bringing about a state of affairs in the future. When one thinks of a student that has to write an essay before its due date, for instance, this is the view that is being used. In other words, the writing and the handing in of the essay before the deadline are seen as responsibilities of the student.

From here on, I will focus on backward-looking responsibility. I work with the following operational definition: *responsibility* is a relation between the agents and the states of affairs of an environment, such that an agent is responsible for a state of affairs iff the agent's degree of involvement in the realization of that state of affairs warrants blame or praise (in light of a given normative system). As for point (b), I follow [10] and [14] and distinguish three main *categories* of responsibility, where each category can be correlated with the components of responsibility that it involves:[2]

1. ***Causal responsibility***: an agent is causally responsible for a state of affairs iff the agent is the material author of such a state of affairs. The component that this category involves is agency.
2. ***Informational responsibility***: an agent is informationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly, or consciously, while bringing about the state of affairs. The components that this category involves are agency, knowledge, and belief.
3. ***Motivational responsibility***: an agent is motivationally responsible for a state of affairs iff the agent is the material author and it behaved knowingly and intentionally while bringing about the state of affairs. The components that this category involves are agency, knowledge, and intentions.

Finally, as for point (c), the two *forms* of responsibility are the *active* form and the *passive* form. The active form of responsibility concerns contributions, and the passive form of responsibility concerns omissions.

Now, key elements in my operational definition of responsibility are the notions of blame and praise. Intuitively, responsibility can be measured by how much blame or how much praise an agent gets for its participation in bringing about a state of affairs. As mentioned before, *ought-to-do's* can provide a criterion for deciding when agents should be blamed and when agents should be praised. The main idea is as follows: if agent α ought to have done ϕ, then having seen to it that ϕ makes α praiseworthy, while having refrained from seeing to it that ϕ makes α blameworthy. For a given ϕ, then, the degrees of α's praiseworthiness/blameworthiness correspond to the possible combinations between (a) an agent's ought-to-do's and (b) the active/passive forms of the three categories of responsibility.

[2] These categories extend the literature's common distinction between *causal* and *agentive* responsibility [17, 23, 13], and they were derived by [10] on the basis of his analysis of the modes of *mens rea*.

3 A Logic of Responsibility

We are ready to introduce *intentional epistemic act-utilitarian stit theory* (*IEAUST*), a stit-theoretic logic of responsibility. Without further ado, let me address the syntax and semantics of this expressive framework.

3.1 Syntax & Semantics

Definition 3.1 (Syntax of intentional epistemic act-utilitarian stit theory)**.** Given a finite set Ags of agent names and a countable set of propositions P, the grammar for the formal language $\mathscr{L}_{\mathrm{R}}$ is given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi \mid [\alpha]\phi \mid K_\alpha\phi \mid I_\alpha\phi \mid \odot_\alpha\phi \mid \odot_\alpha^{\mathscr{S}}\phi,$$

where p ranges over P and α ranges over Ags.

In this language, $\Box\varphi$ is meant to express the historical necessity of φ ($\Diamond\varphi$ abbreviates $\neg\Box\neg\varphi$); $[\alpha]\varphi$ expresses that 'agent α has seen to it that φ'; $K_\alpha\phi$ expresses that 'α knew φ'; $I_\alpha\phi$ expresses that 'α had a present-directed intention toward the realization of φ'; $\odot_\alpha\phi$ expresses that 'α objectively ought to have seen to it that ϕ'; and $\odot_\alpha^{\mathscr{S}}\phi$ expresses that 'α subjectively ought to have seen to it that ϕ.' As for the semantics, the structures on which the formulas of $\mathscr{L}_{\mathrm{R}}$ are evaluated are based on what I call *knowledge-intentions-oughts branching-time frames*. Let me first present the formal definition of these frames and then review the intuitions behind the extensions.

Definition 3.2 (*Kiobt*-frames & models)**.** A tuple $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \tau, \mathbf{Value}\right\rangle$ is called a *knowledge-intention-oughts branching-time frame* (*kiobt*-frame for short) iff

- M is a non-empty set of moments and $\sqsubset$ is a strict partial ordering on M satisfying 'no backward branching.' Each maximal $\sqsubset$-chain of moments is called a history, where each history represents a complete temporal evolution of the world. H denotes the set of all histories, and for each $m \in M$, $H_m := \{h \in H; m \in h\}$. Tuples $\langle m,h\rangle$ such that $m \in M$, $h \in H$, and $m \in h$, are called *indices*, and the set of indices is denoted by $I(M\times H)$. **Choice** is a function that maps each agent α and moment m to a partition $\mathbf{Choice}_\alpha^m$ of H_m, where the cells of such a partition represent α's available actions at m. For $m \in M$ and $h \in H_m$, we denote the equivalence class of h in $\mathbf{Choice}_\alpha^m$ by $\mathbf{Choice}_\alpha^m(h)$. **Choice** satisfies two constraints:
  - (NC) *No choice between undivided histories*: For all $h,h' \in H_m$, if $m' \in h\cap h'$ for some $m' \sqsupset m$, then $h \in L$ iff $h' \in L$ for every $L \in \mathbf{Choice}_\alpha^m$.
  - (IA) *Independence of agency*: A function s on Ags is called a *selection function* at m if it assigns to each α a member of $\mathbf{Choice}_\alpha^m$. If we denote by $\mathbf{Select}^m$ the set of all selection functions at m, then we have that for every $m \in M$ and $s \in \mathbf{Select}^m$, $\bigcap_{\alpha\in Ags} s(\alpha) \neq \emptyset$ (see [8] for a discussion of the property).
- For $\alpha \in Ags$, $\sim_\alpha$ is the epistemic indistinguishability equivalence relation for agent α, which satisfies the following constraints:
  - (OAC) *Own action condition*: if $\langle m_*,h_*\rangle \sim_\alpha \langle m,h\rangle$, then $\langle m_*,h'_*\rangle \sim_\alpha \langle m,h\rangle$ for every $h'_* \in \mathbf{Choice}_\alpha^{m_*}(h_*)$. We refer to this constraint as the 'own action condition' because it implies that agents do not know more than what they perform.
  - $(\mathtt{Unif}-\mathtt{H})$ *Uniformity of historical possibility*: if $\langle m_*,h_*\rangle \sim_\alpha \langle m,h\rangle$, then for every $h'_* \in H_{m_*}$ there exists $h' \in H_m$ such that $\langle m_*,h'_*\rangle \sim_\alpha \langle m,h'\rangle$. Combined with (OAC), this constraint is meant to capture a notion of uniformity of strategies, where epistemically indistinguishable indices should have the same available actions for the agent to choose upon.

For $\langle m,h\rangle$ and $\alpha \in Ags$, the set $\pi^{\Box}_{\alpha}[\langle m,h\rangle] := \{\langle m',h'\rangle; \exists h'' \in H_{m'} s.t. \langle m,h\rangle \sim_{\alpha} \langle m',h''\rangle\}$ is known as α's *ex ante information set*.

- τ is a function that assigns to each $\alpha \in Ags$ and index $\langle m,h\rangle$ a topology $\tau_{\alpha}^{\langle m,h\rangle}$ on $\pi^{\Box}_{\alpha}[\langle m,h\rangle]$. This is the *topology of α's intentionality at* $\langle m,h\rangle$, where any non-empty open set is interpreted as a *present-directed intention*, written 'p-d intention' from here on, of α at $\langle m,h\rangle$. Additionally, τ must satisfy the following conditions:
  - (CI) *Finitary consistency of intention*: for every $\alpha \in Ags$ and index $\langle m,h\rangle$, every non-empty $U,V \in \tau_{\alpha}^{\langle m,h\rangle}$ are such that $U \cap V \neq \emptyset$. In other words, every non-empty $U \in \tau_{\alpha}^{\langle m,h\rangle}$ is $\tau_{\alpha}^{\langle m,h\rangle}$-dense.
  - (KI) *Knowledge of intention*: for every $\alpha \in Ags$ and index $\langle m,h\rangle$, $\tau_{\alpha}^{\langle m,h\rangle} = \tau_{\alpha}^{\langle m',h'\rangle}$ for every $\langle m',h'\rangle$ such that $\pi^{\Box}_{\alpha}[\langle m,h\rangle] = \pi^{\Box}_{\alpha}[\langle m',h'\rangle]$. In other words, α has the same topology of p-d intentions at all indices lying within α's current *ex ante* information set.
- **Value** is a deontic function that assigns to each history $h \in H$ a real number, representing the utility of h.

A *kiobt*-model $\mathscr{M}$, then, results from adding a valuation function $\mathscr{V}$ to a *kiobt*-frame, where $\mathscr{V} : P \to 2^{I(M\times H)}$ assigns to each atomic proposition a set of indices.

For $\alpha \in Ags$, the equivalence relation $\sim_{\alpha}$ is the usual indistinguishability relation, borrowed from epistemic logic, that represents α's uncertainty: whatever holds at all epistemically accessible indices is what an agent knows. As for the function τ, it assigns to each agent the topology of intentions, according to the ideas presented by [3]. The open sets of any such topology are taken to be p-d intentions for bringing about circumstances. At each moment, the fact that the non-empty open sets of the topologies are dense implies that an agent's intentions are consistent.

Regarding the deontic dimension, the idea is that objective, subjective, and doxastic ought-to-do's stem from the optimal actions for an agent: to have seen to it that ϕ is taken to be an obligation of an agent at an index iff ϕ is an effect of all the optimal actions for that agent and index, where the notion of optimality is based on the deontic value of the histories in those actions—provided by **Value**. The semantics for formulas involving the deontic operators require some previous definitions. For $m \in M$ and $\beta \in Ags$, we define $\mathbf{State}^{m}_{\beta} = \left\{S \subseteq H_m; S = \bigcap_{\alpha \in Ags-\{\beta\}} s(\alpha), \text{ where } s \in \mathbf{Select}^{m}\right\}$. For $\alpha \in Ags$ and $m_* \in M$, we first define a general ordering $\leq$ on $\mathscr{P}(H_{m_*})$ such that for $X,Y \subseteq H_{m_*}$, $X \leq Y$ iff $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in X, h' \in Y$. The *objective* dominance ordering $\preceq$ is defined such that for $L,L' \in \mathbf{Choice}^{m_*}_{\alpha}$, $L \preceq L'$ iff for each $S \in \mathbf{State}^{m_*}_{\alpha}, L \cap S \leq L' \cap S$. The optimal set of actions is taken as $\mathbf{Optimal}^{m_*}_{\alpha} := \{L \in \mathbf{Choice}^{m_*}_{\alpha}; \text{there is no } L' \in \mathbf{Choice}^{m_*}_{\alpha} \text{such that } L \prec L'\}$.

Subjective ought-to-do's involve a different dominance ordering. To define it, [11] and [1] introduce the so-called *epistemic clusters*, which are nothing more than a given action's epistemic equivalents in indices that are indistinguishable to the one of evaluation. Formally, we have that for $\alpha \in Ags$, $m_*, m \in M$, and $L \subseteq H_{m_*}$, L's *epistemic cluster* at m is the set $[L]^{m}_{\alpha} := \{h \in H_m; \exists h_* \in L \text{ s.t. } \langle m_*,h_*\rangle \sim_{\alpha} \langle m,h\rangle\}$. A subjective dominance ordering $\preceq_s$ on $\mathbf{Choice}^{m_*}_{\alpha}$ is then defined by the following rule: for $L,L' \subseteq H_{m_*}$, $L \preceq_s L'$ iff for each m such that $m_* \sim_{\alpha} m$, for each $S \in \mathbf{State}^{m}_{\alpha}, [L]^{m}_{\alpha} \cap S \leq [L']^{m}_{\alpha} \cap S$.[3] Just as in the case of objective ought-to-do's, this ordering allows us to define a subjectively optimal set of actions $\mathbf{SOptimal}^{m_*}_{\alpha} := \{L \in \mathbf{Choice}^{m_*}_{\alpha}; \text{ there is no } L' \in \mathbf{Choice}^{m_*}_{\alpha} \text{ s. t. } L \prec_s L'\}$, where I write $L \prec_s L'$ iff $L \preceq_s L'$ and $L' \not\preceq_s L$.

[3] As a convention, I write $m \sim_{\alpha} m'$ if there exist $h \in H_m$, $h' \in H_{m'}$ such that $\langle m,h\rangle \sim_{\alpha} \langle m',h'\rangle$.

Therefore, *kiobt*-frames allow us to represent the components of responsibility discussed in the introduction: agency, knowledge, intentions, and ought-to-do's. More precisely, they allow us to provide semantics for the modalities of $\mathscr{L}_{\mathsf{R}}$:

Definition 3.3 (Evaluation rules for *IEAUST*)**.** Let $\mathscr{M}$ be a finite-choice *kiobt*-model.[4] The semantics on $\mathscr{M}$ for the formulas of $\mathscr{L}_{\mathsf{R}}$ are recursively defined by the following truth conditions:

$$\begin{array}{lll}
\mathscr{M},\langle m,h\rangle \models p & \text{iff} & \langle m,h\rangle \in \mathscr{V}(p) \\
\mathscr{M},\langle m,h\rangle \models \neg\phi & \text{iff} & \mathscr{M},\langle m,h\rangle \not\models \phi \\
\mathscr{M},\langle m,h\rangle \models \phi\wedge\psi & \text{iff} & \mathscr{M},\langle m,h\rangle \models \phi \text{ and } \mathscr{M},\langle m,h\rangle \models \psi \\
\mathscr{M},\langle m,h\rangle \models \Box\phi & \text{iff} & \text{for all } h'\in H_m, \mathscr{M},\langle m,h'\rangle \models \phi \\
\mathscr{M},\langle m,h\rangle \models [\alpha]\phi & \text{iff} & \text{for all } h'\in \mathbf{Choice}^m_\alpha(h), \mathscr{M},\langle m,h'\rangle \models \phi \\
\mathscr{M},\langle m,h\rangle \models K_\alpha\phi & \text{iff} & \text{for all } \langle m',h'\rangle \text{ s. t. } \langle m,h\rangle \sim_\alpha \langle m',h'\rangle, \\
 & & \mathscr{M},\langle m',h'\rangle \models \phi \\
\mathscr{M},\langle m,h\rangle \models I_\alpha\phi & \text{iff} & \text{there exists } U\in \tau_\alpha^{\langle m,h\rangle} \text{ s. t. } U\subseteq \|\phi\| \\
\mathscr{M},\langle m,h\rangle \models \odot_\alpha\phi & \text{iff} & \text{for all } L\in \mathbf{Optimal}^m_\alpha, \mathscr{M},\langle m,h'\rangle \models \varphi \\
 & & \text{for every } h'\in L \\
\mathscr{M},\langle m,h\rangle \models \odot^{\mathscr{S}}_\alpha\varphi & \text{iff} & \text{for all } L\in \mathbf{SOptimal}^m_\alpha, \mathscr{M},\langle m',h'\rangle \models \varphi \\
 & & \text{for every } m' \text{ s. t. } m\sim_\alpha m' \text{ and every } h'\in [L]^{m''}_\alpha.
\end{array}$$

where $\|\phi\|$ refers to the set $\{\langle m,h\rangle \in I(M\times H); \mathscr{M},\langle m,h\rangle \models \phi\}$.

3.2 Formalization of Sub-Categories of Responsibility

The logic introduced in the previous subsection allows us to formalize different modes of responsibility by means of formulas of $\mathscr{L}_{\mathsf{R}}$. Before diving into the formulas, let me present an operational definition for the expression 'mode of responsibility.' For $\alpha\in Ags$, index $\langle m,h\rangle$, and ϕ of $\mathscr{L}_{\mathsf{R}}$, a *mode of α's responsibility with respect to ϕ at $\langle m,h\rangle$* is a tuple consisting of three constituents: (1) a set of categories, taken from Broersen's three categories of responsibility, that applies to the relation between α and ϕ at $\langle m,h\rangle$, (2) the forms of responsibility—active or passive—that apply to the categories in said set, and (3) a deontic context, determining whether the forms of the categories are either blameworthy, praiseworthy, or neutral. As for constituents (1) and (2), observe that the active and passive forms of the three categories of responsibility lead to sub-categories of the notion. For clarity, first I will introduce the stit-theoretic characterizations of these sub-categories; afterwards, in Subsection 3.3, these sub-categories will be discussed against the backdrop of the deontic contexts that will decide their degree of blameworthiness or praiseworthiness (constituent (3) in a given mode).

A maxim usually endorsed in the philosophical literature on moral responsibility is the *principle of alternate possibilities*. According to this principle, "a person is morally responsible for what he has done only if he could have done otherwise" [15]. Following the example of [17], then, I adopt the intuitions behind deliberative agency and restrict my view on responsibility to situations where agents can be said to actually have had a hand in bringing about states of affairs. Therefore, each sub-category of α's responsibility with respect to ϕ at $\langle m,h\rangle$ will include a positive condition—concerning the realization of ϕ—and a negative condition—concerning the realization of $\neg\phi$. For $\alpha\in Ags$ and ϕ of $\mathscr{L}_{\mathsf{R}}$, the main sub-categories of α's responsibility with respect to ϕ are displayed in Table 1.

[4] Finite-choice *bt*-models are those for which function **Choice** is such that $\mathbf{Choice}^m_\alpha$ is finite for every $\alpha\in Ags$ and $m\in M$. I focus on finite-choice models to simplify the evaluation rules for objective and subjective ought-to-do's. The reader is referred to [1] for the evaluation rules in the case of infinite-choice models.

Form *Category*	Active (contributions)	Passive (omissions)
Causal	$[\alpha]\phi \wedge \Diamond[\alpha]\neg\phi$	$\phi \wedge \Diamond[\alpha]\neg\phi$
Informational	$K_\alpha[\alpha]\phi \wedge \Diamond K_\alpha[\alpha]\neg\phi$	$\phi \wedge K_\alpha\neg[\alpha]\neg\phi \wedge \Diamond K_\alpha[\alpha]\neg\phi$
Motivational	$K_\alpha[\alpha]\phi \wedge I_\alpha[\alpha]\phi \wedge \Diamond K_\alpha[\alpha]\neg\phi$	$\phi \wedge K_\alpha\neg[\alpha]\neg\phi \wedge I_\alpha\neg[\alpha]\neg\phi \wedge \Diamond K_\alpha[\alpha]\neg\phi$

Table 1: Main sub-categories.

Let me explain and discuss Table 1. Let $\mathscr{M}$ be a *kiobt*-model. For $\alpha \in Ags$ and index $\langle m,h\rangle$, the sub-categories of α's responsibility with respect to ϕ at $\langle m,h\rangle$ are defined as follows:

- α was *causal-active responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ α has seen to it that ϕ (the positive condition) and it was possible for α to prevent ϕ (the negative condition). As such, I refer to state of affairs ϕ as a causal contribution of α at $\langle m,h\rangle$. α was *causal-passive responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ ϕ was the case (the positive condition), and α refrained from preventing ϕ while it was possible for α to prevent ϕ (the negative conditions). To clarify, formula $\phi \to \neg[\alpha]\neg\phi$ is valid, so that if ϕ was the case then α refrained from preventing ϕ. I refer to $\neg\phi$ as a causal omission of α at $\langle m,h\rangle$.
- α was *informational-active responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ α has knowingly seen to it that ϕ (the positive condition) and it was possible for α to knowingly prevent ϕ (the negative condition). I refer to ϕ as a conscious contribution of α at $\langle m,h\rangle$. α was *informational-passive responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ ϕ was the case (the positive condition), and α knowingly refrained from preventing ϕ while it was possible for α to knowingly prevent ϕ (the negative conditions). I refer to $\neg\phi$ as a conscious omission of α at $\langle m,h\rangle$.
- α was *motivational-active responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ α has both knowingly and intentionally seen to it that ϕ (the positive conditions) and it was possible for α to knowingly prevent ϕ (the negative condition). I refer to ϕ as a motivational contribution of α at $\langle m,h\rangle$. α was *motivational-passive responsible* for ϕ at $\langle m,h\rangle$ iff at $\langle m,h\rangle$ ϕ was the case (the positive condition), and α both knowingly and intentionally refrained from preventing ϕ while it was possible for α to knowingly prevent ϕ (the negative conditions). I refer to $\neg\phi$ as a motivational omission of α at $\langle m,h\rangle$.

The main reason for setting the negative conditions as stated in Table 1 is that it greatly simplifies the relation between the active and the passive forms of responsibility. That said, it is important to mention that these negative conditions lead to a policy that I call *leniency on blameworthy agents*.

Two important observations concerning the relations between these sub-categories are the following:

1. (a) If α was informational-active, resp. informational-passive, responsible for ϕ at $\langle m,h\rangle$, then α was causal-active, resp. causal-passive, responsible for ϕ at $\langle m,h\rangle$; the converse is not true.
   (b) If α was motivational-active, resp. motivational-passive, responsible for ϕ at $\langle m,h\rangle$, then α was informational-active, resp. informational-passive, responsible for ϕ at $\langle m,h\rangle$; the converse is not true.
2. For all three categories, the active form of responsibility with respect to ϕ implies the passive form.

3.3 Formalization of Modes of Responsibility

In Section 2 I explained that obligations provide the deontic contexts of responsibility, which in turn determine degrees of praiseworthiness/blameworthiness for instances of the notion. Let $\mathscr{M}$ be a *kiobt*-model. Take $\alpha \in Ags$, and let ϕ be a formula of $\mathscr{L}_{\mathsf{R}}$. For each index $\langle m,h \rangle$, there are 4 main possibilities for conjunctions of deontic modalities holding at $\langle m,h \rangle$, according to whether $\Delta\phi$ or $\neg\Delta\phi$ is satisfied at the index, where $\Delta \in \left\{\odot_{\alpha}, \odot_{\alpha}^{\mathscr{S}}\right\}$. I refer to any such conjunction as a *deontic context for α's responsibility with respect to ϕ at $\langle m,h \rangle$*. Thus, these contexts render 4 main levels of praiseworthiness, resp. blameworthiness, under the premise that bringing about ϕ is praiseworthy and refraining from bringing about ϕ is blameworthy. I use numbers 1–4 to refer to these levels, so that *Level* 1 corresponds the highest level of praiseworthiness, resp. blameworthiness, and *Level* 4 corresponds to the lowest level.

Level 1: when deontic context $\odot_{\alpha}\phi \wedge \odot_{\alpha}^{\mathscr{S}} \phi$ holds at $\langle m,h \rangle$, which occurs iff at $\langle m,h \rangle$ α objectively and subjectively ought to have seen to it that ϕ. ***Level 2***: when deontic context $\neg \odot_{\alpha} \phi \wedge \odot_{\alpha}^{\mathscr{S}} \phi$ holds at $\langle m,h \rangle$, which occurs iff at $\langle m,h \rangle$ α subjectively ought to have seen to it that ϕ, but α did not objectively ought to have seen to it that ϕ. ***Level 3***: when deontic context $\odot_{\alpha}\phi \wedge \neg \odot_{\alpha}^{\mathscr{S}} \phi$ holds at $\langle m,h \rangle$, which occurs iff at $\langle m,h \rangle$ α objectively ought to have seen to it that ϕ, but α did not subjectively ought to have seen to it that ϕ. ***Level 4***: when deontic context $\neg \odot_{\alpha} \phi \wedge \neg \odot_{\alpha}^{\mathscr{S}} \phi$ holds at $\langle m,h \rangle$, where, unless α either objectively or subjectively ought have seen to it that $\neg\phi$ at $\langle m,h \rangle$ (which would imply that a deontic context of the previous levels holds with respect to $\neg\phi$), neither bringing about ϕ nor refraining from doing so elicits any interest in terms of blame-or-praise assignment.

For each of these deontic contexts, the *basic modes of α's active responsibility with respect to ϕ at $\langle m,h \rangle$* are displayed in Table 2, and the *basic modes of α's passive responsibility* are obtained by substituting the term 'passive' for 'active' in such a table.

Att. *Deg.*	**Praiseworthiness**	**Blameworthiness**
Low_A	Causal-active for ϕ ✓ Infor.-active for ϕ ✗ Motiv.-active for ϕ ✗	Causal-active for $\neg\phi$ ✓ Infor.-active for $\neg\phi$ ✗ Motiv.-active for $\neg\phi$ ✗
$Middle_A$	Causal-active for ϕ ✓ Infor.-active for ϕ ✓ Motiv.-active for ϕ ✗	Causal-active for $\neg\phi$ ✓ Infor.-active for $\neg\phi$ ✓ Motiv.-active for $\neg\phi$ ✗
$High_A$	Causal-active for ϕ ✓ Infor.-active for ϕ ✓ Motiv.-active for ϕ ✓	Causal-active for $\neg\phi$ ✓ Infor.-active for $\neg\phi$ ✓ Motiv.-active for $\neg\phi$ ✓

Table 2: Modes of α's active responsibility with respect to ϕ.

4 Axiomatization

This section is devoted to introducing proof systems for *IEAUST*. More precisely, I present two systems:

- A sound system for *IEAUST*, for which achieving a completeness result is still an open problem.

- A sound and complete system for a technical extension of *IEAUST* that I refer to as *bi-valued IEAUST*. Bi-valued *IEAUST* was devised with the aim of having a completeness result for a logic that would be reasonably similar to the one presented in Section 3.

As for the first bullet point, a proof system for *IEAUST* is defined as follows:

Definition 4.1 (Proof system for *IEAUST*)**.** Let Λ_R be the proof system defined by the following axioms and rules of inference:

- *(Axioms)* All classical tautologies from propositional logic; the **S5** schemata for $\Box$, $[\alpha]$, and K_α; the **KD** schemata for I_α; and the schemata given in Table 3.

Basic-stit-theory schemata:	*Schemata for knowledge:*
$\Box\phi \to [\alpha]\phi$ (SET) For distinct $\alpha_1, \dots, \alpha_m$, $\bigwedge_{1\leq k\leq m} \Diamond[\alpha_i]\phi_i \to \Diamond\left(\bigwedge_{1\leq k\leq m}[\alpha_i]\phi_i\right)$ (IA)	$K_\alpha\phi \to [\alpha]\phi$ (OAC) $\Diamond K_\alpha\phi \to K_\alpha\Diamond\phi$ $(Unif-H)$
Schemata for objective ought-to-do's:	*Schemata for subjective ought-to-do's:*
$\odot_\alpha(\phi \to \psi) \to (\odot_\alpha\phi \to \odot_\alpha\psi)$ $(A1)$ $\Box\phi \to \odot_\alpha\phi$ $(A2)$ $\odot_\alpha\phi \to \Box\odot_\alpha\phi$ $(A3)$ $\odot_\alpha\phi \to \odot_\alpha([\alpha]\phi)$ $(A4)$ $\odot_\alpha\phi \to \Diamond[\alpha]\phi$ (Oic)	$\odot_\alpha^{\mathscr{S}}(\phi \to \psi) \to (\odot_\alpha^{\mathscr{S}}\phi \to \odot_\alpha^{\mathscr{S}}\psi)$ $(A5)$ $\odot_\alpha^{\mathscr{S}}\phi \to \odot_\alpha^{\mathscr{S}}(K_\alpha\phi)$ $(A6)$ $K_\alpha\Box\phi \to \odot_\alpha^{\mathscr{S}}\phi$ (SuN) $\odot_\alpha^{\mathscr{S}}\phi \to \Diamond K_\alpha\phi$ $(s.Oic)$ $\odot_\alpha^{\mathscr{S}}\phi \to K_\alpha\Box\odot_\alpha^{\mathscr{S}}\phi$ $(s.Cl)$ $\odot_\alpha^{\mathscr{S}}\phi \to \neg\odot_\alpha\neg\phi$ $(ConSO)$
Schemata for intentionality:	
$\Box K_\alpha\phi \to I_\alpha\phi$ (InN) $I_\alpha\phi \to \Box K_\alpha I_\alpha\phi$ (KI)	

Table 3: Axioms for the modalities' interactions.

- *(Rules of inference) Modus Ponens*, Substitution, and Necessitation for all modal operators.

For a discussion of all these axioms and schemas, the reader is referred to [1, 18, 3]. An important result for Λ_R, then, is the following proposition, whose proof is relegated to Appendix A.

Proposition 4.2 (Soundness of Λ_R)**.** *The proof system Λ_R is sound with respect to the class of* kiobt-*models.*

Unfortunately, the question of whether Λ_R is complete with respect to the class of *kiobt*-models is still an open problem. Now, in the search for a complete proof system for *IEAUST*, and following a strategy found in my joint works with Jan Broersen [1, 2], I tried to first prove completeness of Λ_R with respect to a class of more general models, that I refer to as *bi-valued kiobt*-models (Definition 4.3 below). This strategy led to the need of dropping one of the schemata in Λ_R: $(ConSO)$. More precisely, if Λ_R' is obtained from Λ_R by eliminating $(ConSO)$ in Definition 4.1, then Λ_R' turns out to be sound and complete with respect to the class of *bi-valued kiobt*-models. The formal statements are included below.

Definition 4.3 (Bi-valued *kiobt*-frames & models)**.** $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \tau, \mathbf{Value}_{\mathscr{O}}, \mathbf{Value}_{\mathscr{S}}\right\rangle$ is called a *bi-valued kiobt*-frame iff

- $M, \sqsubset, Ags, \mathbf{Choice}$, $\{\sim_\alpha\}_{\alpha\in Ags}$, and τ are defined just as in Definition 3.2.
- $\mathbf{Value}_{\mathscr{O}}$ and $\mathbf{Value}_{\mathscr{S}}$ are functions that independently assign to each history $h \in H$ a real number.

A *bi-valued kiobt*-model $\mathscr{M}$, then, results from adding a valuation function $\mathscr{V}$ to a bi-valued *kiobt*-frame, where $\mathscr{V} : P \to 2^{I(M \times H)}$ assigns to each atomic proposition of $\mathscr{L}_{\mathsf{R}}$ a set of indices (recall that P is the set of propositions in $\mathscr{L}_{\mathsf{R}}$).

The two value functions in bi-valued *kiobt*-frames allow us to redefine the dominance orderings so that they are independent from one another, something that proves useful in achieving a completeness result in the style of [1]. For $\alpha \in Ags$ and $m \in M$, two general orderings $\leq$ and $\leq_s$ are first defined on 2^{H_m}: for $X, Y \subseteq H_m$, $X \leq Y$, resp. $X \leq_s Y$, iff $\texttt{Value}_{\mathscr{O}}(h) \leq \texttt{Value}_{\mathscr{O}}(h')$, resp. $\texttt{Value}_{\mathscr{S}}(h) \leq \texttt{Value}_{\mathscr{S}}(h')$, for every $h \in X$ and $h' \in Y$. Then, for $\alpha \in Ags$ and $m \in M$, an objective dominance ordering $\preceq$ is now defined on $\mathbf{Choice}^m_\alpha$ by the rule: $L \preceq L'$ iff for every $S \in \mathbf{State}^m_\alpha, L \cap S \leq L' \cap S$. In turn, for $\alpha \in Ags$ and $m \in M$, a subjective dominance ordering $\preceq_s$ is now defined on $\mathbf{Choice}^m_\alpha$ by the rule: $L \preceq_s L'$ iff for all m' such that $m \sim_\alpha m'$ and each $S \in \mathbf{State}^m_\alpha$, $[L]^{m'}_\alpha \cap S \leq_s [L']^{m'}_\alpha \cap S$. With these new notions, the sets $\mathbf{Optimal}^m_\alpha$ and $\mathbf{SOptimal}^m_\alpha$ are redefined accordingly, and the evaluation rules for the formulas of $\mathscr{L}_{\mathsf{R}}$ (with respect to bi-valued *kiobt*-models) are given just as in Definition 3.3. As mentioned before, I refer to the resulting logic as *bi-valued IEAUST*. Bi-valued *IEAUST*, then, admits the following metalogic result, whose proof is sketched in Appendix A.

Theorem 4.4 (Soundness & Completeness of Λ'_R)**.** *Let Λ'_R be the proof system obtained from Λ_R by eliminating (ConSO) in Definition 4.1. Then Λ'_R is sound and complete with respect to the class of bi-valued* kiobt-*models.*

5 Conclusion

This paper built a formal theory of responsibility by means of stit-theoretic models and languages that were designed to explore the interplay between the following components of responsibility: agency, knowledge, beliefs, intentions, and obligations. Said models were integrated into a framework that is rich enough to provide logic-based characterizations for different instances of three categories of responsibility: causal, informational, and motivational responsibility.

The developed theory belongs to a relatively recent tradition in the philosophical literature, that seeks to formalize responsibility allocation by means of models of agency and logic-based languages (see, for instance, [16], [17], [5], [19], [20], and [7]). Most of these frameworks characterize different forms of responsibility as combinations of causal agency, knowledge, and the principle of alternate possibilities. The novelty of the present approach, then, lies in the introduction of intentionality and ought-to-do's. Such an introduction gives rise to a taxonomy that distinguishes various kinds of responsibility and blameworthiness/praiseworthiness in a methodical, meticulous way. Interesting directions for future work, then, involve extending these models with beliefs and rational decision-making, group notions (coalitions, group knowledge & belief, collective intentionality, collective responsibility), temporal modalities, and long-term strategies, for instance. As for the technical aspects of the formal theory, an important directions for future work involve checking whether the logic is decidable, checking for the complexity of its satisfiability problem, and figuring out its applicability for implementation.[5]

[5] Implementing logics of responsibility might prove relevant in the design, formal verification, and explainability of ethical AI (see, for instance, [12]).

References

[1] Aldo Iván Ramírez Abarca & Jan Broersen (2019): *A Logic of Objective and Subjective Oughts*. In: *European Conference on Logics in Artificial Intelligence*, Springer, pp. 629–641, doi:10.1007/978-3-030-19570-0_41.

[2] Aldo Iván Ramírez Abarca & Jan Broersen (2021): *A Deontic Stit Logic Based on Beliefs and Expected Utility*. *Electronic Proceedings in Theoretical Computer Science* 335, pp. 281–294, doi:10.4204%2Feptcs.335.27.

[3] Aldo Iván Ramírez Abarca & Jan Broersen (2023): *A stit logic of intentionality*. In Carlos Areces & Diana Costa, editors: *Dynamic Logic. New Trends and Applications*, Springer, pp. 125–153, doi:10.1007/978-3-031-26622-5_8.

[4] Thomas Ågotnes (2006): *Action and knowledge in alternating-time temporal logic*. *Synthese* 149(2), pp. 375–407, doi:10.1007/s11229-005-3875-8.

[5] Natasha Alechina, Joseph Y Halpern & Brian Logan (2017): *Causality, Responsibility and Blame in Team Plans*. In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1091–1099. Available at https://dl.acm.org/doi/10.5555/3091125.3091279.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina & Richard Benjamins (2020): *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. *Information Fusion* 58, pp. 82–115, doi:10.1016/j.inffus.2019.12.012.

[7] Christel Baier, Florian Funke & Rupak Majumdar (2021): *A Game-Theoretic Account of Responsibility Allocation*. In: *Thirtieth International Joint Conference on Artificial Intelligence*, IJCAI, pp. 1773–1779, doi:10.24963/ijcai.2021/244.

[8] N. Belnap, M. Perloff & M. Xu (2001): *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.

[9] Jan Broersen (2008): *A complete stit logic for knowledge and action, and some of its applications*. In: *International Workshop on Declarative Agent Languages and Technologies*, Springer, pp. 47–59, doi:10.1007/978-3-540-93920-7_4.

[10] Jan Broersen (2011): *Deontic epistemic stit logic distinguishing modes of mens rea*. *Journal of Applied Logic* 9(2), pp. 137–152, doi:10.1016/j.jal.2010.06.002.

[11] Jan Broersen & Aldo Iván Ramírez Abarca (2018): *Formalising Oughts and Practical Knowledge without Resorting to Action Types*. In: *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1877–1879. Available at https://dl.acm.org/doi/abs/10.5555/3237383.3238009.

[12] Roberta Calegari, Giovanni Ciatto, Enrico Denti & Andrea Omicini (2020): *Logic-based technologies for intelligent systems: State of the art and perspectives*. *Information* 11(3), p. 167, doi:10.3390/info11030167.

[13] Roger Crisp (2014): *Aristotle: Nicomachean Ethics*. Cambridge University Press.

[14] Hein Duijf (2018): *Let's do it!: Collective responsibility, joint action, and participation*. Ph.D. thesis, Utrecht University.

[15] Harry Frankfurt (2018): *Alternate possibilities and moral responsibility*. In: *Moral Responsibility and Alternative Possibilities*, Routledge, pp. 17–25, doi:10.2307/2023833.

[16] Tiago de Lima, Lambér Royakkers & Frank Dignum (2010): *A logic for reasoning about responsibility*. *Logic Journal of the IGPL* 18(1), pp. 99–117, doi:10.1093/jigpal/jzp073.

[17] Emiliano Lorini, Dominique Longin & Eunate Mayor (2014): *A logical analysis of responsibility attribution: emotions, individuals and collectives*. *Journal of Logic and Computation* 24(6), pp. 1313–1339, doi:10.1093/logcom/ext072.

[18] Yuko Murakami (2004): *Utilitarian deontic logic*. *Advances in Modal Logic* 287.

[19] Pavel Naumov & Jia Tao (2019): *Blameworthiness in strategic games.* In: Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp. 3011–3018, doi:`10.1609/aaai.v33i01.33013011`.

[20] Pavel Naumov & Jia Tao (2020): *An epistemic logic of blameworthiness.* Artificial Intelligence 283, p. 103269, doi:`10.1016/j.artint.2020.103269`.

[21] Luís Moniz Pereira & Ari Saptawijaya (2016): *Programming machine ethics.* 26, Springer, doi:`10.1007/978-3-319-29354-7_8`.

[22] Ibo van de Poel (2011): *The relation between forward-looking and backward-looking responsibility.* In: Moral Responsibility, Springer, pp. 37–52, doi:`10.1007/978-94-007-1878-4_3`.

[23] Gary Watson (1996): *Two faces of responsibility.* Philosophical Topics 24(2), pp. 227–248, doi:`10.5840/philtopics199624222`.

[24] Bernard Weiner (1995): *Judgments of responsibility: A foundation for a theory of social conduct.* Guilford Press.

[25] Ming Xu (1994): *Decidability of deliberative stit theories with multiple agents.* In: International Conference on Temporal Logic, Springer, pp. 332–348, doi:`10.1007/BFb0013997`.

A Metalogic Results for *IEAUST*

Appendix A.A Soundness

Proposition A.1 (Soundness of Λ_R). *The system Λ_R (Definition 4) is sound with respect to the class of* kiobt-*models.*

Proof. The proof of soundness is routine: the validity of the **S5** schemata for $\Box$ and $[\alpha]$, as well as that of (SET) and (IA), is standard from [25]; the validity of the **S5** schemata for K_α is standard from epistemic logic; the validity of schemata $(A1)$–$(A4)$, as well as that of (Oic), is standard from [18]; the validity of the **KD** schemata for I_α, as well as that of (InN), follows from Definitions 3.2 and 3.3; and the validity of (KI) follows from frame condition $(\mathtt{KI})$; and the validity of schemata (OAC), $(Unif-H)$, $(A5)$ and $(A6)$, as well as that of (SuN), $(s.Oic)$, $(s.Cl)$, and $(ConSO)$ can be shown as follows:

To see that $\mathscr{M} \models (OAC)$, take $\langle m,h\rangle$ such that $\mathscr{M},\langle m,h\rangle \models K_\alpha\varphi$. Take $h' \in \mathbf{Choice}^m_\alpha(h)$. Frame condition $(\mathtt{OAC})$ implies that $\langle m,h\rangle \sim_\alpha \langle m,h'\rangle$. The assumption that $\mathscr{M},\langle m,h\rangle \models K_\alpha\varphi$ then implies that $\mathscr{M},\langle m,h'\rangle \models \varphi$. Therefore, for any $h' \in \mathbf{Choice}^m_\alpha(h)$, $\mathscr{M},\langle m,h'\rangle \models \varphi$, which implies that $\mathscr{M},\langle m,h\rangle \models [\alpha]\varphi$.

To see that $\mathscr{M} \models (Unif-H)$, take $\langle m,h\rangle$ such that $\mathscr{M},\langle m,h\rangle \models \Diamond K_\alpha\varphi$. Let $\langle m',h'\rangle$ be an index such that $\langle m,h\rangle \sim_\alpha \langle m',h'\rangle$. We want to show that $\mathscr{M},\langle m',h'\rangle \models \Diamond\varphi$. The fact that $\mathscr{M},\langle m,h\rangle \models \Diamond K_\alpha\varphi$ implies that there exists $h_* \in H_m$ such that $(\star)$ $\mathscr{M},\langle m,h_*\rangle \models K_\alpha\varphi$. Frame condition $(\mathtt{Unif-H})$ implies that there exists $h'_* \in H_{m'}$ such that $\langle m,h_*\rangle \sim_\alpha \langle m',h'_*\rangle$. With $(\star)$, this last fact implies that $\mathscr{M},\langle m',h'_*\rangle \models \varphi$, which in turn implies that $\mathscr{M},\langle m',h'\rangle \models \Diamond\varphi$. Therefore, $\mathscr{M},\langle m,h\rangle \models K_\alpha\Diamond\varphi$.

To see that $\mathscr{M} \models (A6)$, take $\langle m,h\rangle$ such that $\mathscr{M},\langle m,h\rangle \models \odot^{\mathscr{S}}_\alpha\varphi$. We want to show that, for every $L \in \mathbf{Choice}^m_\alpha$ such that $[L]^{m'} \nsubseteq |K_\alpha\varphi|^{m'}$ (for some m' such that $m \sim_\alpha m'$), there is $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']^{m''}_\alpha \subseteq |K_\alpha\varphi|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Take $L \in \mathbf{Choice}^m_\alpha$ such that there exists $m' \in M$ such that $m \sim_\alpha m'$ and $[L]^{m'} \nsubseteq |K_\alpha\varphi|^{m'}$. This implies that $[L]^{m'''} \nsubseteq |\phi|^{m'''}$ for some m''' such that $m' \sim_\alpha m'''$. Now, transitivity of $\sim_\alpha$ implies that $m \sim_\alpha m'''$. Therefore, the assumption that $\mathscr{M},\langle m,h\rangle \models \odot^{\mathscr{S}}_\alpha\varphi$ implies that there is $L' \in \mathbf{Choice}^m_\alpha$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']^{m''}_\alpha \subseteq |\varphi|^{m''}$ for every m'' such that $m \sim_\alpha m''$. By definition of epistemic clusters and transitivity of $\sim_\alpha$, this last clause implies that if $L'' = L'$ or $L' \preceq_s L''$ then $[L'']_\alpha^{m''} \subseteq |K_\alpha \varphi|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Thus, L' attests to the fact that $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha^{\mathscr{S}} (K_\alpha \varphi)$.

To see that $\mathscr{M} \models (SuN)$, take $\langle m,h \rangle$ such that $\mathscr{M}, \langle m,h \rangle \models K_\alpha \Box \varphi$. Take $L \in \mathbf{Choice}_\alpha^m$, and let $m' \in M$ be such that $m \sim_\alpha m'$ (which means that there exist $j \in H_m$, $j' \in H_{m'}$ such that $\langle m,j \rangle \sim_\alpha \langle m',j' \rangle$). Condition $(\mathtt{Unif}-\mathtt{H})$ ensures that there exists $h' \in H_{m'}$ such that $\langle m,h \rangle \sim_\alpha \langle m',h' \rangle$. The assumption that $\mathscr{M}, \langle m,h \rangle \models K_\alpha \Box \varphi$ then implies that $\mathscr{M}, \langle m',h' \rangle \models \Box \varphi$. Thus, for any $h'' \in [L]_\alpha^{m'}$, the fact that $h'' \in H_{m'}$ yields that $\mathscr{M}, \langle m',h'' \rangle \models \varphi$. Therefore, for all $L \in \mathbf{Choice}_\alpha^m$ and m' such that $m \sim_\alpha m'$, $[L]_\alpha^{m'} \subseteq |\phi|^{m'}$, which vacuously implies that $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$.

To see that $\mathscr{M} \models (s.Oic)$, take $\langle m,h \rangle$ such that $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$. This implies that there exists $L \subseteq H_m$ such that $[L]_\alpha^{m''} \subseteq |\phi|^{m''}$ for every $m'' \in M$ such that $m \sim_\alpha m''$. Since $\sim_\alpha$ is reflexive, $[L]_\alpha^m \subseteq |\phi|^m$. Now, take $h_0 \in L$. Let $\langle m',h' \rangle$ be an index such that $\langle m,h_0 \rangle \sim_\alpha \langle m',h' \rangle$. From the definition of epistemic clusters, $h' \in [L]_\alpha^{m'}$, so the fact that $[L]_\alpha^{m'} \subseteq |\phi|^{m'}$ implies that $\mathscr{M}, \langle m',h' \rangle \models \varphi$. Therefore, history $h_0 \in H_m$ is such that, for every $\langle m',h' \rangle$ with $\langle m,h_0 \rangle \sim_\alpha \langle m',h' \rangle$, $\mathscr{M}, \langle m',h' \rangle \models \varphi$. This means that $\mathscr{M}, \langle m,h_0 \rangle \models K_\alpha \varphi$, which implies that $\mathscr{M}, \langle m,h \rangle \models \Diamond K_\alpha \varphi$.

To see that $\mathscr{M} \models (s.Cl)$, take $\langle m_*,h_* \rangle$ such that $\mathscr{M}, \langle m_*,h_* \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$. Let $\langle m,j \rangle$ be such that $\langle m_*,h_* \rangle \sim_\alpha \langle m,j \rangle$. Take $h \in H_m$. We want to show that, for every $L \in \mathbf{Choice}_\alpha^m$ such that $[L]^{m'} \not\subseteq |\phi|^{m'}$ (for some m' such that $m \sim_\alpha m'$), there is $L' \in \mathbf{Choice}_\alpha^m$ such that $L \prec_s L'$ and, if $L'' = L'$ or $L' \preceq_s L''$, then $[L'']_\alpha^{m''} \subseteq |\phi|^{m''}$ for every m'' such that $m \sim_\alpha m''$. Take $L \in \mathbf{Choice}_\alpha^m$ such that there exists $m' \in M$ such that $m \sim_\alpha m'$ and $[L]^{m'} \not\subseteq |\phi|^{m'}$. Let N_L be an action in $\mathbf{Choice}_\alpha^{m_*}$ such that $N_L \subseteq [L]_\alpha^{m_*}$, where we know that such an action exists in virtue of $(\mathtt{Unif}-\mathtt{H})$ and $(\mathtt{OAC})$. Notice that transitivity of $\sim_\alpha$ entails that $[N_L]_\alpha^o = [L]_\alpha^o$ for any moment o, so that $[N_L]_\alpha^{m'} \not\subseteq |\phi|^{m'}$. Since $\mathscr{M}, \langle m_*,h_* \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$, there must exist $N \in \mathbf{Choice}_\alpha^{m_*}$ such that $N_L \prec_s N$ and, if $N' = N$ or $N \preceq_s N'$, then $[N']_\alpha^{m''} \subseteq |\phi|^{m''}$ for every m'' such that $m_* \sim_\alpha m''$. Now, let L_N be an action in $\mathbf{Choice}_\alpha^m$ such that $L_N \subseteq [N]_\alpha^m$ (which implies that $[L_N]_\alpha^o = [N]_\alpha^o$ for any moment o). We claim that $L \prec_s L_N$, and show our claim with the following argument: let $m'' \in M$ be such that $m \sim_\alpha m''$, and take $S \in \mathbf{State}_\alpha^{m''}$; on the one hand, $(\star)$ $[L]_\alpha^{m''} \cap S = [N_L]_\alpha^{m''} \cap S \leq [N]_\alpha^{m''} \cap S = [L_N]_\alpha^{m''} \cap S$; on the other hand, we know that there exist a moment m''' and a state $S_0 \in \mathbf{State}_\alpha^{m'''}$ such that $m_* \sim_\alpha m'''$ and such that $[N]_\alpha^{m'''} \cap S_0 \not\leq [N_L]_\alpha^{m'''} \cap S_0$; therefore, $(\star\star)$ $[L_N]_\alpha^{m'''} \cap S_0 = [N]_\alpha^{m'''} \cap S_0 \not\leq [N_L]_\alpha^{m'''} \cap S_0 = [L]_\alpha^{m'''} \cap S_0$. Together, $(\star)$ and $(\star\star)$ entail that $L \prec_s L_N$, proving our claim. Now, let $L'' \in \mathbf{Choice}_\alpha^m$ be such that $L'' = L_N$ or $L_N \preceq_s L''$. If $L'' = L_N$, then $[L'']_\alpha^{m''} = [N]_\alpha^{m''} \subseteq |\phi|^{m''}$ for every m'' such that $m \sim_\alpha m''$. If $L_N \prec_s L''$, then an argument similar to the one used to show that our claim was true renders that there is an action $N_{L''} \in \mathbf{Choice}_\alpha^{m_*}$ such that $N_{L''} \subseteq [L'']_\alpha^{m_*}$ and $N \preceq_s N_{L''}$. Thus, $[L'']_\alpha^{m''} = [N_{L''}]_\alpha^{m''} \subseteq |\phi|^{m''}$. With this, we have shown that $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$ for every $h \in H_m$, so that $\mathscr{M}, \langle m,j \rangle \models \Box \odot_\alpha^{\mathscr{S}} \varphi$. But $\langle m,j \rangle$ was an arbitrary index such that $\langle m_*,h_* \rangle \sim_\alpha \langle m,j \rangle$. Thus, $\mathscr{M}, \langle m_*,h_* \rangle \models K_\alpha \Box \odot_\alpha^{\mathscr{S}} \varphi$.

Let us show that $\mathscr{M} \models (ConSO)$. First of all, let us show that, for all $L, L' \in \mathbf{Choice}_\alpha^m$, if $L \preceq_s L'$, then $L \preceq L'$. Take $L, L' \in \mathbf{Choice}_\alpha^m$. If $L \preceq_s L'$, then, for each m' such that $m \sim_\alpha m'$, $\mathbf{Value}(h) \leq \mathbf{Value}(h')$ for every $h \in [L]_\alpha^{m'}, h' \in [L']_\alpha^{m'}$. Reflexivity of $\sim_\alpha$ implies both that $m \sim_\alpha m'$ and that $L \subseteq [L]_\alpha^m$ and $L' \subseteq [L']_\alpha^m$. Therefore, for all $h'' \in L$ and $h''' \in L'$, $\mathbf{Value}(h'') \leq \mathbf{Value}(h''')$, which implies that $L \preceq L'$.

Now, let $\langle m,h \rangle$ be an index. Assume for a contradiction that $(\star)$ $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha^{\mathscr{S}} \varphi$ and that $(\star\star)$ $\mathscr{M}, \langle m,h \rangle \models \odot_\alpha \neg \varphi$. On the one hand, assumption $(\star)$ implies that there is $L_* \in \mathbf{Choice}_\alpha^m$ such that $L_* \subseteq |\phi|^m$. Thus, assumption $(\star\star)$ yields that there is $L'_* \in \mathbf{Choice}_\alpha^m$ such that $L_* \prec L'_*$ and, if $N = L'_*$ or $L'_* \preceq N$, then $N \subseteq |\neg\phi|^m$. In particular, $L'_* \subseteq |\neg\phi|^m$. Assumption $(\star)$ then implies that there is $L''_* \in \mathbf{Choice}_\alpha^m$ such that $L'_* \prec_s L''_*$ and, if $N = L''_*$ or $L''_* \preceq_s N$, then $N \subseteq [N]_\alpha^m \subseteq |\phi|^m$. In particular, $L''_* \subseteq |\phi|^m$. On the other hand, by the first observation in the proof, the fact that $L'_* \prec_s L''_*$ implies that $L'_* \preceq L''_*$, so that assumption ($\star\star$) yields that $L''_* \subseteq |\neg\phi|^m$, which contradicts the previously shown fact that $L''_* \subseteq |\phi|^m$. Thus, $\mathscr{M}, \langle m,h\rangle \models \odot^{\mathscr{S}}_\alpha \varphi \rightarrow \neg \odot_\alpha \neg\varphi$ for every index $\langle m,h\rangle$, so that $\odot^{\mathscr{S}}_\alpha \varphi \rightarrow \neg \odot_\alpha \neg\varphi$ is indeed valid.

It is clear that the rules of inference *Modus Ponens*, Substitution, and Necessitation for the modal operators all preserve validity. □

Appendix A.B Completeness

As mentioned in the main body of the paper, whether Λ_R is complete with respect to the class of *kiobt*-models is still an open problem. However, the proof system Λ'_R—obtained from Λ_R by eliminating $(ConSO)$ in Definition 4.1—is sound and complete with respect to the class of bi-valued *kiobt*-models (Definition 4.3). Soundness follows from Proposition A.1, and the proof of completeness is obtained by integrating the proofs of completeness in [1] and [3]. More precisely, the proof of completeness will be sketched below as a two-step process. First, I introduce a Kripke semantics for *bi-valued IEAUST*, where the formulas of $\mathscr{L}_\mathrm{R}$ are evaluated on bi-valued Kripke-*kios*-models (Definition A.2). Completeness of Λ_R' with respect to the class of these structures is shown via the well-known technique of canonical models. Secondly, a truth-preserving correspondence between bi-valued Kripke-*kios*-models and a subclass of bi-valued *kiobt*-models is used to prove completeness with respect to bi-valued *kiobt*-models via completeness with respect to bi-valued Kripke-*kios*-models.

A Kripke semantics for *IEAUST* is defined as follows:

Definition A.2 (Bi-valued Kripke-*kios*-frames & models)**.** A tuple

$$\left\langle W, Ags, R_\Box, \textbf{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R^I_\alpha\right\}_{\alpha\in Ags}, \mathtt{Value}_{\mathscr{O}}, \mathtt{Value}_{\mathscr{S}}\right\rangle$$

is called a *bi-valued* Kripke-*kios*-frame iff

- W is a set of possible worlds. $R_\Box$ is an equivalence relation over W. For $w \in W$, the class of w under $R_\Box$ is denoted by $\overline{w}$. $\mathtt{Choice}$ is a function that assigns to each $\alpha \in Ags$ and $\Box$-class $\overline{w}$ a partition $\mathtt{Choice}^{\overline{w}}_\alpha$ of $\overline{w}$ given by an equivalence relation denoted by $R^{\overline{w}}_\alpha$. $\mathtt{Choice}$ must satisfy the following constraint:
  - $(\mathtt{IA})_\mathtt{K}$ For all $w \in W$, each function $s : Ags \rightarrow 2^{\overline{w}}$ that maps α to a member of $\mathtt{Choice}^{\overline{w}}_\alpha$ is such that $\bigcap_{\alpha\in Ags} s(\alpha) \neq \emptyset$ (where the set of all functions s that map α to a member of $\mathtt{Choice}^{\overline{w}}_\alpha$ is denoted by $\mathtt{Select}^{\overline{w}}$).

  For $\alpha \in Ags$, $w \in W$, and $v \in \overline{w}$, the class of v in the partition $\mathtt{Choice}^{\overline{w}}_\alpha$ is denoted by $\mathtt{Choice}^{\overline{w}}_\alpha(v)$. Now, for $\beta \in Ags$ and $w \in W$, $\mathtt{State}^{\overline{w}}_\beta := \left\{S \subseteq \overline{w}; S = \bigcap_{\alpha\in Ags-\{\beta\}} s(\alpha), \text{ for } s \in \mathtt{Select}^{\overline{w}}\right\}$, where $\mathtt{Select}^{\overline{w}}$ denotes the set of all selection functions at $\overline{w}$ (i.e., functions that assign to each α a member of $\mathtt{Choice}^{\overline{w}}_\alpha$).
- For all $\alpha \in Ags$, $\approx_\alpha$ is an (epistemic) equivalence relation on W. The following conditions must be satisfied:
  - $(\mathtt{OAC})_\mathtt{K}$ For all $\alpha \in Ags$, $w \in W$, and $v \in \overline{w}$, $v \approx_\alpha u$ for every $u \in \mathtt{Choice}^{\overline{w}}_\alpha(v)$.
  - $(\mathtt{Unif-H})_\mathtt{K}$ For all $\alpha \in Ags$, if $v,u \in W$ are such that $v \approx_\alpha u$, then for all $v' \in \overline{v}$ there exists $u' \in \overline{u}$ such that $v' \approx_\alpha u'$.

  For $\alpha \in Ags$ and $w \in W$, α's ex ante *information set at* w is defined as $\pi^\Box_\alpha[w] := \{v; w \approx_\alpha \circ R_\Box v\}$, which by frame condition $(\mathtt{Unif-H})_\mathtt{K}$ coincides with the set $\{v; wR_\Box \circ \approx_\alpha v\}$. To clarify, $(\mathtt{Unif-H})_\mathtt{K}$ implies that $R_\Box \circ \approx_\alpha = \approx_\alpha \circ R_\Box$. Thus, $\approx_\alpha \circ R_\Box$ is an equivalence relation such that $\pi^\Box_\alpha[w] = \pi^\Box_\alpha[v]$ for every $w, v \in W$ such that $w \approx_\alpha \circ R_\Box v$.

- For $\alpha \in Ags$, R^I_α is a serial, transitive, and euclidean relation on W such that $R^I_\alpha \subseteq \approx_\alpha \circ R_\Box$ and such that the following condition is satisfied:
  - $(\mathtt{Den})_{\mathtt{K}}$ For all $v, u \in W$ such that $v \approx_\alpha \circ R_\Box u$, there exists $z \in W$ such that $vR^I_\alpha z$ and $uR^I_\alpha z$.

  For $\alpha \in Ags$, R^{I+}_α denotes the reflexive closure of R^I_α. For $w \in W$, $w \uparrow_{R^{I+}_\alpha}$ denotes the set $\left\{v \in W; wR^{I+}_\alpha v\right\}$.

  For $w, v \in W$, I write $\overline{w} \approx_\alpha \overline{v}$ iff there exist $w' \in \overline{w}$ and $v' \in \overline{v}$ such that $w' \approx_\alpha v'$. For $w, v \in W$ such that $\overline{w} \approx_\alpha \overline{v}$ and $L \in \mathtt{Choice}^{\overline{w}}_\alpha$, L's epistemic cluster at $\overline{v}$ is the set $[\![L]\!]^{\overline{v}}_\alpha := \{u \in \overline{v}; \text{ there is } o \in L \text{ such that } o \approx_\alpha u\}$.

- $\mathtt{Value}_{\mathscr{O}}$ and $\mathtt{Value}_{\mathscr{S}}$ are functions that independently assign to each world $w \in W$ a real number.

  These functions are used to define an objective ordering $\preceq$ and a subjective ordering $\preceq_s$ of choices. Formally, for $\alpha \in Ags$ and $w \in W$, one first defines two general orderings $\leq$ and $\leq_s$ on 2^W by the rules: $X \leq Y$ iff $\mathtt{Value}_{\mathscr{O}}(w) \leq \mathtt{Value}_{\mathscr{O}}(w')$ for all $w \in X$ and $w' \in Y$; and $X \leq_s Y$ iff $\mathtt{Value}_{\mathscr{S}}(w) \leq \mathtt{Value}_{\mathscr{S}}(w')$ for all $w \in X$ and $w' \in Y$. An objective dominance ordering $\preceq$ is then defined on $\mathtt{Choice}^{\overline{w}}_\alpha$ by the rule: $L \preceq L'$ iff $L \cap S \leq L' \cap S$ for every $S \in \mathtt{State}^{\overline{w}}_\alpha$. In turn, a subjective dominance ordering $\preceq_s$ is then defined on $\mathtt{Choice}^{\overline{w}}_\alpha$ by the rule: $L \preceq_s L'$ iff $[\![L]\!]^{\overline{v}}_\alpha \cap S \leq_s [\![L']\!]^{\overline{v}}_\alpha \cap S$ for every v such that $w \approx_\alpha v$ and every $S \in \mathtt{State}^{\overline{v}}_\alpha$. I write $L \prec L'$ iff $L \preceq L'$ and $L' \npreceq L$, and I write $L \prec_s L'$ iff $L \preceq_s L'$ and $L' \npreceq_s L$, so that $\mathtt{Optimal}^{\overline{w}}_\alpha := \left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha; \text{ there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec L'\right\}$ and $\mathtt{SOptimal}^{\overline{w}}_\alpha := \left\{L \in \mathtt{Choice}^{\overline{w}}_\alpha; \text{ there is no } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec_s L'\right\}$.

A Kripke-*kios*-model $\mathscr{M}$ consists of the tuple that results from adding a valuation function $\mathscr{V}$ to a Kripke-*kios*-frame, where $\mathscr{V} : P \to 2^W$ assigns to each atomic proposition a set of worlds (recall that P is the set of propositions in $\mathscr{L}_{\mathsf{R}}$).

Kripke-*kios*-models allow us to evaluate the formulas of $\mathscr{L}_{\mathsf{R}}$ with semantics that are analogous to the ones provided for *kiobt*-models:

Definition A.3 (Evaluation rules on Kripke models)**.** Let $\mathscr{M}$ be a Kripke-*kios*-model, the semantics on $\mathscr{M}$ for the formulas of $\mathscr{L}_{\mathsf{KO}}$ are defined recursively by the following truth conditions, evaluated at world w:

$$
\begin{array}{lll}
\mathscr{M}, w \models p & \text{iff} & w \in \mathscr{V}(p) \\
\mathscr{M}, w \models \neg\phi & \text{iff} & \mathscr{M}, w \not\models \phi \\
\mathscr{M}, w \models \phi \wedge \psi & \text{iff} & \mathscr{M}, w \models \phi \text{ and } \mathscr{M}, w \models \psi \\
\mathscr{M}, w \models \Box\phi & \text{iff} & \text{for each } v \in \overline{w}, \mathscr{M}, v \models \phi \\
\mathscr{M}, w \models [\alpha]\phi & \text{iff} & \text{for each } v \in \mathtt{Choice}^{\overline{w}}_\alpha(w), \mathscr{M}, v \models \phi \\
\mathscr{M}, w \models K_\alpha\phi & \text{iff} & \text{for each } v \text{ s. t. } w \approx_\alpha v, \mathscr{M}, v \models \phi \\
\mathscr{M}, w \models I_\alpha\phi & \text{iff} & \text{there exists } x \in \pi^\Box_\alpha[w] \text{ s. t. } x \uparrow_{R^{I+}_\alpha} \subseteq |\phi| \\
\mathscr{M}, w \models \odot_\alpha\varphi & \text{iff} & \text{for all } L \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } \mathscr{M}, v \not\models \varphi \text{ for some } v \in L, \text{ there is} \\
 & & L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec L' \text{ and, if } L'' = L' \text{ or } L' \preceq_s L'', \\
 & & \text{then } \mathscr{M}, w' \models \varphi \text{ for every } w' \in L''_\alpha \\
\mathscr{M}, w \models \odot^{\mathscr{S}}_\alpha\varphi & \text{iff} & \text{for all } L \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } \mathscr{M}, v \not\models \varphi \text{ for some } w' \text{ s. t. } w \approx_\alpha w' \\
 & & \text{and some } v \in [\![L]\!]^{w'}_\alpha, \text{ there is } L' \in \mathtt{Choice}^{\overline{w}}_\alpha \text{ s. t. } L \prec_s L' \\
 & & \text{and, if } L'' = L' \text{ or } L' \preceq_s L'', \text{then } \mathscr{M}, w''' \models \varphi \text{ for every } w'' \\
 & & \text{s. t. } \overline{w} \approx_\alpha \overline{w''} \text{ and every } w''' \in [\![L'']\!]^{w''}_\alpha,
\end{array}
$$

where I write $|\phi|$ to refer to the set $\{w \in W; \mathscr{M}, w \models \phi\}$. Satisfiability, validity on a frame, and general validity are defined as usual.

A truth-preserving correspondence between Kripke-*kios*-models and *kiobt*-models is shown as follows:

Definition A.4 (Associated *kiobt*-frame)**.** Let

$$\mathscr{F} = \left\langle W, Ags, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R^I_\alpha\right\}_{\alpha\in Ags}, \texttt{Value}_{\mathscr{O}}, \texttt{Value}_{\mathscr{S}}\right\rangle$$

be a bi-valued Kripke-*kios*-frame.

Then $\mathscr{F}^T := \left\langle M_W, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha\in Ags}, \tau, \mathbf{Value}_{\mathscr{O}}, \mathbf{Value}_{\mathscr{S}}\right\rangle$ is called the bi-valued *kiobt*-frame associated with $\mathscr{F}$ iff

- $M_W, \sqsubset, \mathbf{Choice}$, $\{\sim_\alpha\}_{\alpha\in Ags}$, and τ are defined just as in Definition 11 in [3].
- $\mathbf{Value}_{\mathscr{O}}$ and $\mathbf{Value}_{\mathscr{S}}$ are defined by the following rules: for $h_v \in H$, $\mathbf{Value}_{\mathscr{O}}(h_v) = \texttt{Value}_{\mathscr{O}}(v)$, and $\mathbf{Value}_{\mathscr{S}}(h_v) = \texttt{Value}_{\mathscr{S}}(v)$.

Proposition A.5. *Let $\mathscr{F}$ be a bi-valued Kripke-*kios*-frame. Then $\mathscr{F}^T$ is a bi-valued* kiobt*-frame, indeed.*

Proof. Follows from Proposition 2 in [3] and Definition A.4. □

Lemma A.6. *Let $\mathscr{M}$ be a bi-valued Kripke-*kios*-model, and let $\mathscr{M}^T$ be its associated bi-valued* kiobt*-model. For all $\alpha \in Ags$, $w \in W$, and $L, N \in \texttt{Choice}^{\overline{w}}_\alpha$, the following conditions hold:*

(a) $L \preceq N$ *iff* $L^T \preceq N^T$ *and* $L \prec N$ *iff* $L^T \prec N^T$.

(b) $L \preceq_s N$ *iff* $L^T \preceq_s N^T$ *and* $L \prec_s N$ *iff* $L^T \prec_s N^T$.

(c) $L \in \texttt{Optimal}^{\overline{w}}_\alpha$ *iff* $L^T \in \mathbf{Optimal}^{\overline{w}}_\alpha$.

(d) $L \in \texttt{S}-\texttt{Optimal}^{\overline{w}}_\alpha$ *iff* $L^T \in \mathbf{S}-\mathbf{Optimal}^{\overline{w}}_\alpha$.

Proof. The reader is referred to the proof of Proposition 4 in `https://doi.org/10.48550/arXiv.1903.10577` for a proof. □

Proposition A.7 (Truth-preserving correspondence)**.** *Let $\mathscr{M}$ be a bi-valued Kripke-*kios*-model, and let $\mathscr{M}^T$ be its associated bi-valued* kiobt*-model. For all ϕ of $\mathscr{L}_R$ and $w \in W$, $\mathscr{M}, w \models \phi$ iff $\mathscr{M}^T, \langle \overline{w}, h_w\rangle \models \phi$.*

Proof. We proceed by induction on the complexity of ϕ. For the base case, the cases of Boolean connectives, and the cases of all modal operators except I_α, the proofs are exactly the same as their analogs' in Proposition 4 in `https://doi.org/10.48550/arXiv.1903.10577`. For the case of I_α, the proof is the same as its analog in Proposition 3 in [3].

□

Thus, completeness with respect to bi-valued *kiobt*-models is proved with Propositions A.8 and A.9 below.

Proposition A.8 (Completeness w.r.t. bi-valued Kripke-*kios*-models)**.** *The proof system Λ_R' is complete with respect to the class of bi-valued Kripke-*kios*-models.*

Proof. Completeness with respect to bi-valued Kripke-*kios*-models is shown via canonical models. To be precise, one defines a structure $\mathscr{M} := \left\langle W^{\Lambda'_R}, R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}, \left\{R^I_\alpha\right\}_{\alpha\in Ags} \texttt{Value}_{\mathscr{O}}, \texttt{Value}_{\mathscr{S}}, \mathscr{V}\right\rangle$, where $W^{\Lambda'_R} = \{w; w \text{ is a } \Lambda'_R\text{-MCS}\}$, where $R_\Box, \texttt{Choice}, \{\approx_\alpha\}_{\alpha\in Ags}$, $\left\{R^I_\alpha\right\}_{\alpha\in Ags}$, and $\mathscr{V}$ are defined just as in Definition 12 in [3], and where $\mathtt{Value}_{\mathscr{O}}$ and $\mathtt{Value}_{\mathscr{S}}$ are defined as follows: for $\alpha \in Ags$ and $w \in W^{\Lambda}$, one first defines $\Sigma^w_\alpha := \{[\alpha]\phi; \odot[\alpha]\phi \in w\}$ and $\Gamma^w_\alpha := \{K_\alpha\phi; \odot_{\mathscr{S}}[\alpha]\phi \in w\}$. Then, taking $\Sigma^w = \bigcup_{\alpha\in Ags}\Sigma^w_\alpha$ and $\Gamma^w = \bigcup_{\alpha\in Ags}$, the deontic functions are given by

$$\begin{aligned} \mathtt{Value}_{\mathscr{O}}(w) &= \begin{cases} 1 \text{ iff } \Sigma^w \subseteq w, \\ 0 \text{ otherwise.} \end{cases} \\ \mathtt{Value}_{\mathscr{S}}(w) &= \begin{cases} 1 \text{ iff } \Gamma^w \subseteq w, \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

The canonical structure $\mathscr{M}$ is shown to be a bi-valued Kripke-*kios*-model just as in Proposition 4 in [3]. Then, the so-called *truth lemma* is shown by merging Lemma 2 in [3] and Lemma 4 in `https://doi.org/10.48550/arXiv.1903.10577`. This renders completeness with respect to bi-valued Kripke-*kios*-models. □

Proposition A.9 (Completeness w.r.t. bi-valued *kiobt*-models)**.** *The proof system Λ_R' is complete with respect to the class of bi-valued* kiobt-*models.*

Proof. Let ϕ be a Λ'_R-consistent formula of $\mathscr{L}_{\mathrm{R}}$. Proposition A.8 implies that there exists a bi-valued Kripke-*kios*-model $\mathscr{M}$ and a world w in its domain such that $\mathscr{M}, w \models \phi$. Proposition A.7 then ensures that the bi-valued *kiobt*-model $\mathscr{M}^T$ associated with $\mathscr{M}$ is such that $\mathscr{M}^T, \langle \overline{w}, h_w \rangle \models \phi$. □

Therefore, Proposition A.1 and Proposition A.9 imply that the following result, appearing in the main body of the paper, is true:

Theorem 4.4 (Soundness & Completeness of Λ'_R)**.** *Let Λ'_R be the proof system obtained from Λ_R by eliminating* (*ConSO*) *in Definition 4.1. Then Λ'_R is sound and complete with respect to the class of bi-valued* kiobt-*models.*

A Sufficient Condition for Gaining Belief in Byzantine Fault-Tolerant Distributed Systems*

Thomas Schlögl
TU Wien, Vienna, Austria
tschloegl@ecs.tuwien.ac.at

Ulrich Schmid
TU Wien, Vienna, Austria
s@ecs.tuwien.ac.at

Existing protocols for byzantine fault tolerant distributed systems usually rely on the correct agents' ability to detect faulty agents and/or to detect the occurrence of some event or action on some correct agent. In this paper, we provide sufficient conditions that allow an agent to infer the appropriate beliefs from its history, and a procedure that allows these conditions to be checked in finite time. Our results thus provide essential stepping stones for developing efficient protocols and proving them correct.

1 Introduction

At least since the ground-breaking work by Halpern and Moses [9], epistemic logic and interpreted runs and systems [6] are known as powerful tools for analyzing distributed systems. *Distributed systems* are multi-agent systems, where a set of $n \geq 2$ agents, each executing some protocol, exchange messages in order to achieve some common goal. In the *interpreted runs and systems* framework, the set of all possible runs R (executions) of the agents in a system determines a set of Kripke models, formed by the evolution of the global state $r(t)$ in all runs $r \in I$ over time $t \in \mathbb{N}$. Epistemic reasoning has been extended to *fault-tolerant* distributed systems right from the beginning, albeit restricted to benign faulty agents, i.e., agents that may only crash and/or drop messages [17, 18, 5, 9].

Actions performed by the agents when executing their protocol take place when they have accumulated specific epistemic knowledge. According to the pivotal *Knowledge of Preconditions Principle* [15], it is universally true that if φ is a necessary condition for an agent to take a certain action, then i may act only if $K_i\varphi$ is true. For example, in order for agent i to decide on 0 in a binary fault-tolerant consensus algorithm [13] (where correct agents must reach a common decision value based on local initial values), it must know that some process has started with initial value 0, i.e., $K_i\varphi \equiv K_i(\exists j : x_j = 0)$ holds true. Showing that agents act without having attained $K_i\varphi$ for some necessary knowledge φ is hence a very effective way for proving impossibilities. Conversely, optimal distributed algorithms can be provided by letting agents act as soon as $K_i\varphi$ for all the necessary knowledge has been established. One example are the crash-resilient *unbeatable* consensus protocols introduced in [2], which are not just worst-case optimal, but not even strictly dominated w.r.t. termination time by any other protocol in *any* execution.

Epistemic reasoning has recently been extended to the analysis of *byzantine* distributed systems [11, 10, 21, 8] as well, where agents may not just crash or lose messages, but where they may also misbehave arbitrarily [13]. Solving a distributed computing problem in such systems is much more difficult, and tighter constraints (e.g. on the maximum number f of faulty agents) are usually needed. For example, byzantine consensus can only be solved if $n \geq 3f+1$ [13], whereas $n > f$ is sufficient for agents that may crash only [2].

*Supported by Digital Modeling of Asynchronous Integrated Circuits (P32431-N30).

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 487–497, doi:10.4204/EPTCS.379.37

When inspecting existing byzantine fault-tolerant protocols, one identifies two basic tasks that usually need to be solved by every correct agent, in some way or other: (1) detecting faulty agents, and (2) detecting whether some correct agents are/were in a certain state. We mentioned already a "static" example for (2), namely, finding out whether some correct agent started with initial value 0 in unbeatable consensus protocols [2], which is also needed in byzantine-resilient protocols like [23]. For a more dynamic example, we note that several existing byzantine fault-tolerant protocols employ the fundamental *consistent broadcasting* (CB) primitive, introduced in [22]. In particular, CB is used in fault-tolerant clock synchronization [22, 24, 20], in byzantine synchronous consensus [23, 4], and (in a slightly extended form) in the simulation of crash-prone protocols in byzantine settings proposed in [14]. A variant of CB has been studied epistemically in [8], namely, *firing rebels with relay* (FRR), which is the problem of letting all correct agents execute an action FIRE in an all-or-nothing fashion when sufficiently many agents know of an external START event. It was shown that any correct protocol for implementing FRR (and, hence, CB) requires detecting whether START has occurred on some correct process.

Regarding an example for (1), we point out that it has been shown by Kuznets et. al. in [11] that it is impossible to reliably detect whether some process is *correct* in asynchronous byzantine distributed systems, due to the possibility of a brain-in-a-vat scenario, whereas it is sometimes possible to detect that an agent is *faulty*. And indeed, the ability to (sometimes) reliably diagnose an actually byzantine faulty agent, using approaches like [1], has enabled the design of *fault-detection, isolation and recovery* (FDIR) schemes [19] for high-reliability systems.

In this paper, we will provide sufficient conditions that allow a correct agent i to gain belief about a fact φ encoding (1) resp. (2) that is inherently local at some other correct agent j. Using the *belief modality* (also known as defeasible knowledge [16]) $B_i\varphi \equiv K_i(correct_i \to \varphi)$ that captures what is known by agent i if it is correct, and the *hope modality* $H_i\varphi \equiv correct_i \to B_i\varphi$ introduced in [11], this can be succinctly condensed into the following question: *Under which conditions and by means of which techniques can $B_iH_j\varphi$, which is the belief that i obtains by receiving a message from j that claims (possibly wrongly) that $B_j\varphi$ holds, be lifted to $B_i\varphi$ in asynchronous byzantine systems?* The crucial difference is that correct agent i can infer something about φ from the latter, but not from the former. Note that it has been established in [10, Thm. 15] that $B_i\varphi$ is indeed necessary for agent i to achieve this.

Detailed contributions: For an asynchronous byzantine system with weak communication assumptions,

(1) we provide an algorithm that allows an agent j to compute its belief about the faultiness of the agents, based on both directly received obviously faulty messages and on appropriate notifications from sufficiently many other agents recorded in its local history,

(2) we provide a sufficient condition for agent j to infer, also from its local history, the belief that some event or action has occurred at a correct agent.

Our conditions are sufficient in the sense that if they hold, then the appropriate belief can be obtained. Hence, the question about whether our conditions are also necessary might pop up. It is important to note, however, that the possibility of gaining belief depends heavily on the actual properties of the system. For example, it will turn out that the condition for (2) can be relaxed when a non-empty set of faulty agents is available, e.g. obtained via (1). The same is true if the actual system satisfies stronger communication assumptions. Consequently, we just focus on sufficient conditions for communication assumptions met by all asynchronous byzantine systems we are aware of.

Paper organization: In Section 2, we briefly introduce the cornerstones of the byzantine modeling framework of [11, 12] needed for proving our results. Section 3 provides our basic communication assumptions. Section 4 and Section 5 contains our results for detecting faulty agents (1) and the occurrence of events (2), respectively. Some conclusions and directions of future work are provided in Section 6. Due to lack of space, additional technical details and all the proofs have been relegated to an appendix.

2 The Basic Model

Since this paper uses the framework of [11], we restate the core terms and aspects needed for our results.

There is a finite set $\mathscr{A} = \{1, \dots, n\}$ (for $n \geq 2$) of **agents**, who do not have access to a global clock and execute a possibly non-deterministic joint **protocol**. In such a protocol, agents can perform **actions**, e.g., send **messages** $\mu \in \mathit{Msgs}$, and witness **events**, in particular, message deliveries: the action of sending a copy (numbered k) of a message $\mu \in \mathit{Msgs}$ to an agent $j \in \mathscr{A}$ in a protocol is denoted by $\mathit{send}(j, \mu_k)$, whereas a receipt of such a message from $i \in \mathscr{A}$ is recorded locally as $\mathit{recv}(i, \mu)$. The set of all **actions** (**events**) available to an agent $i \in \mathscr{A}$ is denoted by $\mathit{Actions}_i$ (Events_i), subsumed as **haps** $\mathit{Haps}_i := \mathit{Actions}_i \sqcup \mathit{Events}_i$, with $\mathit{Actions} := \bigcup_{i \in \mathscr{A}} \mathit{Actions}_i$, $\mathit{Events} := \bigcup_{i \in \mathscr{A}} \mathit{Events}_i$, and $\mathit{Haps} := \mathit{Actions} \sqcup \mathit{Events}$.

The other main player in [11] is the **environment** ε, which takes care of scheduling haps, failing agents, and resolving non-deterministic choices in the joint protocol. Since the notation above only describes the local view of agents, there is also a **global** syntactic representation of each hap, which is only available to the environment and contains additional information (regarding the time of a hap, a distinction whether a hap occurred in a correct or byzantine way, etc.). One distinguishes the sets of global events $\mathit{GEvents}_i := \overline{\mathit{GEvents}_i} \sqcup \mathit{BEvents}_i \sqcup \mathit{SysEvents}_i$ of agent i, for a correct agent (signified by the horizontal bar), a byzantine faulty agent, or a system event as explained below. Regarding global actions, one distinguishes correct actions $\overline{\mathit{GActions}_i}$ and faulty actions $\mathit{fake}\,(i, A \mapsto A')$, where the agent actually performs A but claims to have performed A'. Finally, $\mathit{GEvents} := \bigsqcup_{i \in \mathscr{A}} \mathit{GEvents}_i$, $\mathit{GHaps} := \mathit{GEvents} \sqcup \overline{\mathit{GActions}}$. Generally, horizontal bars signify phenomena that are correct, as contrasted by those that may be correct or byzantine.

The model is based on discrete time, of arbitrarily fine resolution, with time domain $t \in \mathbb{T} := \mathbb{N} = \{0, 1, \dots\}$. All haps taking place after a **timestamp** $t \in \mathbb{T}$ and no later than $t+1$ are grouped into a **round** denoted t½ and treated as happening simultaneously. In order to prevent agents from inferring the global time by counting rounds, agents are generally unaware of a round, unless they perceive an event or are prompted to act by the environment. The latter is accomplished by special system events $go(i)$, which are complemented by two more system events for faulty agents: $\mathit{sleep}\,(i)$ and $\mathit{hibernate}\,(i)$ signify a failure to activate the agent's protocol and differ in that the latter does not even wake up the agent. None of the **system events** $\mathit{SysEvents}_i := \{go(i), \mathit{sleep}\,(i), \mathit{hibernate}\,(i)\}$ is directly observable by agents.

Events and actions that can occur in each round, if enabled by $go(i)$, are determined by the protocols for agents and the environment, with non-deterministic choices resolved by the **adversary** that is considered part of the environment. A **run** r is a function mapping a point in time t to an $n+1$ tuple, consisting of the environment's history and local histories $r(t) = (r_\varepsilon(t), r_1(t), \dots, r_n(t))$ representing the state of the whole system (**global state**) at that time t. The set of all global states is denoted by $\mathscr{G}$. The **environment's history** $r_\varepsilon(t) \in \mathscr{L}_\varepsilon$ is a sequence of all haps that happened, in contrast to the local histories faithfully recorded in the global format. Accordingly, $r_\varepsilon(t+1) = X \circ r_\varepsilon(t)$ for the set $X \subseteq \mathit{GHaps}$ of all haps from round t½, where $\circ$ stands for concatenation. Agent i's local view of the system after round t½, i.e., its share of the global state $h = r(t) \in \mathscr{G}$, is recorded in i's **local state** $r_i(t+1) \in \mathscr{L}_i$, also called i's **local history**, sometimes denoted h_i. $r_i(0) \in \Omega_i$ are the **initial local states**, with $\mathscr{G}(0) := \prod_{i \in \mathscr{A}} \Omega_i$. If a round contains neither $go(i)$ nor any event to be recorded in i's local history, then the history $r_i(t+1) = r_i(t)$ remains unchanged, denying the agent knowledge that the round just passed. Otherwise, the agent performs actions from its protocol $P_i(r_i(t)) \subseteq 2^{\mathit{Actions}_i}$ and updates its history $r_i(t+1) = X \circ r_i(t)$, for the set $X \subseteq \mathit{Haps}_i$ of all actions and events perceived by i in round t½. The sets $\beta^t_{\varepsilon_i}(r)$, $\beta^t_i(r)$ denote the sets of events and the set of actions respectively happening in round t½ in global format. For some hap o we write $o \in r_i(t)$ if there exists a round $t' \leq t$, where o was appended to i's local history. Consequently, the local history $r_i(t) = h_i = (h_i(|h_i|), h_i(|h_i| - 1), \ldots, h_i(0))$ is the sequence of all haps $h_i(k)$ perceived by i in the k-th round it was **active** in.

The exact updating procedure is the result of a complex state transition consisting of several phases, described in detail in Appendix A.2, which are grouped into a **transition template** τ that yields a transition relation $\tau_{P_\varepsilon,P}$ for any joint and environment protocol P and P_ε. The set R of all **transitional runs** are all runs that can be generated from some set of initial states $\mathscr{G}(0)$ via some transition template τ.

Proving the correctness of a protocol for solving a certain distributed computing problem boils down to studying the set of runs that can be generated. As **liveness properties** cannot be ensured on a round-by-round basis, they are enforced by restricting the allowable set of runs via **admissibility conditions** Ψ, which are subsets of the set R of all transitional runs. A **context** $\gamma = (P_\varepsilon, \mathscr{G}(0), \tau, \Psi)$ consists of an environment's protocol P_ε, a set of global initial states $\mathscr{G}(0)$, a transition template τ, and an admissibility condition Ψ. For a joint protocol P, we call $\chi = (\gamma, P)$ an **agent context**. The set of all χ-consistent runs is denoted by R^χ that is the set of all transitional runs starting with initial states from $\mathscr{G}(0)$ and transitioning via $\tau_{P_\varepsilon,P}$ (both from χ). The set of all agent contexts we denote by $\mathscr{E}$, where $\mathscr{E}^B \subseteq \mathscr{E}$ consists of all byzantine asynchronous agent contexts, with transition template $\tau^B_{P_\varepsilon,P}$, and $\mathscr{E}^{B_f} \subseteq \mathscr{E}^B$ consists of all byzantine asynchronous agent contexts, where at most f agents can become faulty, with transition template $\tau^{B_f}_{P_\varepsilon,P}$.

Epistemics. [11] defines interpreted systems in this framework as Kripke models for multi-agent distributed environments [6]. The states in such a Kripke model are given by global histories $r(t') \in \mathscr{G}$ for runs $r \in R^\chi$ given some agent context χ and timestamps $t' \in \mathbb{T}$. A **valuation function** $\pi \colon Prop \to 2^{\mathscr{G}}$ determines states where an atomic proposition from $Prop$ is true. This determination is arbitrary except for a small set of **designated atomic propositions**: For $FEvents_i := BEvents_i \sqcup \{sleep\,(i), hibernate\,(i)\}$, $i \in \mathscr{A}$, $o \in Haps_i$, and $t \in \mathbb{T}$ such that $t \leq t'$,

- $correct_{(i,t)}$ is true at $r(t')$ iff no faulty event happened to i by timestamp t, i.e., no event from $FEvents_i$ appears in $r_\varepsilon(t)$,
- $correct_i$ is true at $r(t')$ iff no faulty event happened to i yet, i.e., no event from $FEvents_i$ appears in $r_\varepsilon(t')$,
- $faulty_i$ is true iff $\neg correct_i$ is and $faulty_{(i,t)}$ is true iff $\neg correct_{(i,t)}$ is,
- $fake_{(i,t)}\,(o)$ is true at $r(t')$ iff i has a **faulty** reason to believe that $o \in Haps_i$ occurred in round $(t-1)$½, i.e., $o \in r_i(t)$ because (at least in part) of some $O \in BEvents_i \cap \beta^{t-1}_{\varepsilon_i}\,(r)$ (see Appendix A.2),
- $\overline{occurred}_{(i,t)}(o)$ is true at $r(t')$ iff i has a **correct** reason to believe $o \in Haps_i$ occurred in round $(t-1)$½, i.e., $o \in r_i(t)$ because (at least in part) of $O \in (\overline{GEvents}_i \cap \beta^{t-1}_{\varepsilon_i}\,(r)) \sqcup \beta^{t-1}_i\,(r)$ (see Appendix A.2),
- $\overline{occurred}_i(o)$ is true at $r(t')$ iff at least one of $\overline{occurred}_{(i,m)}(o)$ for $1 \leq m \leq t'$ is; also $\overline{occurred}\,(o) := \bigvee_{i \in \mathscr{A}} \overline{occurred}_i(o)$,
- $occurred_i(o)$ is true at $r(t')$ iff either $\overline{occurred}_i(o)$ is or at least one of $fake_{(i,m)}\,(o)$ for $1 \leq m \leq t'$ is,
- $happened_i(a)$ is true at $r(t')$ for action $a \in Actions_i$ iff there exists a global action A, where $a \in local\,(A)$ s.t. $A \in r_\varepsilon(t'-1)$ or $fake\,(i, A \mapsto A') \in r_\varepsilon(t'-1)$,
- $fhappened_i(a)$ is true at $r(t')$ for action $a \in Actions_i$ iff there exists a global action A, where $a \in local\,(A)$ s.t. $fake\,(i, A \mapsto A') \in r_\varepsilon(t'-1)$,
- $\overline{init}_i(\lambda_0)$ is true at $r(t')$ for initial state $\lambda_0 \in \Omega_i$ (see Appendix A.2) iff $r_i(0) = \lambda_0$. Note that all agents are still correct in any initial state.

The following terms are used to categorize agent faults caused by the environment's protocol P_ε: agent $i \in \mathscr{A}$ is *fallible* if for any $X \in P_\varepsilon\,(t)$, $X \cup \{fail\,(i)\} \in P_\varepsilon\,(t)$; *correctable* if $X \in P_\varepsilon\,(t)$ implies that $X \setminus FEvents_i \in P_\varepsilon(t)$; *delayable* if $X \in P_\varepsilon(t)$ implies $X \setminus GEvents_i \in P_\varepsilon(t)$; *gullible* if $X \in P_\varepsilon(t)$ implies that, for any $Y \subseteq FEvents_i$, the set $Y \sqcup (X \setminus GEvents_i) \in P_\varepsilon(t)$ whenever it is t-coherent (consists of mutually compatible events only, see Definition A.1). Informally, fallible agents can be branded byzantine at any time; correctable agents can always be made correct for the round by removing all their byzantine events; delayable agents can always be forced to skip a round completely (which does not make them faulty); gullible agents can exhibit any faults in place of correct events. Common types of faults, e.g., crash or omission failures, can be obtained by restricting allowable sets Y in the definition of gullible agents.

An **interpreted system** is a pair $\mathscr{I} = (R^\chi, \pi)$. The following BNF defines the **epistemic language** $\mathfrak{L}_\mathfrak{g}$ considered throughout this paper, for $p \in Prop$ and $i \in \mathscr{A}$: $\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid \Box\varphi$ (other Boolean connectives are defined as usual). We also use belief $B_i\varphi := K_i(correct_i \to \varphi)$ and hope $H_i\varphi := correct_i \to B_i\varphi$ as introduced in [11] and axiomatized in [7, 3]. The interpreted systems semantics is defined as usual with global states $r(t)$ and $r'(t')$ indistinguishable for i iff $r_i(t) = r'_i(t')$. Semantics for temporal operator $\Box$ we define as $(\mathscr{I}, r, t) \models \Box\varphi \quad \text{iff} \quad (\forall t' \geq t)\ (\mathscr{I}, r, t') \models \varphi$.

3 Basics of Fault-Tolerant Communication

In this section, we introduce some basic notation and assumptions needed for fault-tolerant communication. Throughout this paper, we consider non-excluding agent contexts where all agents are fallible, gullible, correctable, and delayable [10], which models asynchronous byzantine distributed systems.

Definition 3.1. For the interpreted system $\mathscr{I} = (R^\chi, \pi)$, formula φ from $\mathfrak{L}_\mathfrak{g}$, and agents $i, j \in \mathscr{A}$, we define a set of "trustworthy" messages $Msgs_\varphi^{i \to j} \subseteq Msgs$ that an agent i sends to j only if i believes φ,

$$\mu \in Msgs_\varphi^{i\to j} \quad \Longleftrightarrow \quad \big((\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall D \in P_i(r_i(t)))\ send(j, \mu) \in D \ \Rightarrow\ (\mathscr{I}, r, t) \models B_i\varphi\big). \quad (1)$$

Definition 3.2. We call a formula φ *persistent* in the interpreted system $\mathscr{I} = (R^\chi, \pi)$ if, once true, φ never becomes false again:

$$(\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall t' > t)\ (\mathscr{I}, r, t) \models \varphi \ \Rightarrow\ (\mathscr{I}, r, t') \models \varphi. \quad (2)$$

Persistent formulas have several useful properties, which are easy to prove:[1]

Lemma 3.3. For any agent context $\chi \in \mathscr{E}^B$, agent $i \in \mathscr{A}$, natural number $k \in \mathbb{N}$, action or event $o \in Haps$ and $\lambda_0 \in \Omega_i$, the formulas $faulty_i$, $occurred_i(o)$, $\overline{occurred}_i(o)$ and $\overline{init}_i(\lambda_0)$ are persistent.

Lemma 3.4. If, for an agent context χ, formula φ is persistent, so is $correct_i \to \varphi$ for any agent $i \in \mathscr{A}$.

Lemma 3.5. If, for an agent context χ, a formula φ is persistent, so is $K_i\varphi$, $B_i\varphi$, $H_i\varphi$ for any agent $i \in \mathscr{A}$.

Lemma 3.6. If, for an agent context χ, formulas φ, ψ are persistent, so is $\varphi \wedge \psi$ and $\varphi \vee \psi$.

The following essential lemma states what an agent i can infer epistemically from receiving a trustworthy message referring to a persistent formula φ from agent j. As explained in [8, Remark 11], in the case of j being faulty, agent i need not share its reality with agent j. Consequently, agent i cannot infer $B_i\varphi$ just from receiving such a message.

Lemma 3.7. For agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, $t \in \mathbb{N}$, agents $i, j \in \mathscr{A}$, persistent formula φ, and trustworthy message $\mu \in Msgs_\varphi^{j \to i}$

$$recv(j, \mu) \in r_i(t) \quad \Rightarrow \quad (\mathscr{I}, r, t) \models B_iH_j\varphi. \quad (3)$$

In systems where not all agents can or want to communicate directly with all other agents, messages need to travel multiple hops before they reach a desired recipient. We therefore need to generalize Lemma 3.7 accordingly.

[1] Most of our proofs have been relegated to the appendix.

Definition 3.8. We define the set of all finite agent sequences (including the empty sequence ε), with repetitions allowed, as

$$AgSeq := \{(i_1,i_2,\dots,i_k) \mid i_1,i_2,\dots,i_k \in \mathscr{A}\}. \tag{4}$$

Definition 3.9. For agent sequence $\sigma \in AgSeq$ and formula φ, we define the *nested hope* $\overline{H}_\sigma\varphi$ for φ as

$$\overline{H}_\sigma\varphi := H_{\pi_1\sigma}H_{\pi_2\sigma}\dots H_{\pi_{|\sigma|}\sigma}\varphi, \tag{5}$$

where $\pi_k\sigma$ denotes the application of the kth projection function to σ, and $\overline{H}_\varepsilon\varphi = \varphi$.

Note that the nested hope caused by $\sigma = (i_1,i_2,\dots,i_k)$ goes from the origin i_k to i_1.

Definition 3.10. For a persistent formula φ in agent context χ, run $r \in R^\chi$, time $t \in \mathbb{N}$ and agent $i \in \mathscr{A}$, we define the set of agent sequences that lead to nested hope regarding φ at agent i as

$$\widehat{Recv}^{\,i}_{\varphi}(r_i(t)) := \{\sigma \in AgSeq \mid \mu \in Msgs^{j\to i}_{\overline{H}_{\overline{\sigma}}\varphi} \text{ and } \sigma = (j)\circ\overline{\sigma} \text{ and } recv(j,\mu) \in r_i(t)\}, \tag{6}$$

where $\circ$ denotes sequence concatenation, i.e., $(j)\circ(i_1,\dots,i_k) = (j,i_1,\dots,i_k)$. The agent sequences $\sigma \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t))$ will be called *hope chains*.

With these preparations, we can generalize Lemma 3.7 to arbitrary hope chains as follows:

Corollary 3.11. For agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi,\pi)$, run $r \in R^\chi$, $t \in \mathbb{N}$, agents $i,j \in \mathscr{A}$, and persistent formula φ,

$$(\forall\sigma \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t)))\ (\mathscr{I},r,t) \models B_i\overline{H}_\sigma\varphi. \tag{7}$$

The following two lemmas show that, for any hope chain in the set $\widehat{Recv}^{\,i}_{\varphi}(r_i(t))$ that contains a loop starting and ending in some agent $j \neq i$ in the actual communication chain, i.e., the agent sequence, the corresponding hope chain where the loop is replaced by a single instance of j exists in $\widehat{Recv}^{\,i}_{\varphi}(r_i(t))$.

Lemma 3.12. For persistent φ, $\chi \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi,\pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $\widehat{\sigma} \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t))$, $\sigma = (i)\circ\widehat{\sigma}$,

$$(\forall k \in [1,|\sigma|-1])\ (\mathscr{I},r,t) \models \bigwedge_{k'\in[1,k]} correct_{\pi_{k'}\sigma} \quad\Rightarrow\quad (\mathscr{I},r,t) \models \overline{H}_{\pi_{k+1}\sigma\circ\dots\circ\pi_{|\sigma|}\sigma}\varphi \tag{8}$$

Lemma 3.13. For persistent formula φ, $\chi \in \mathscr{E}^B$, $r \in R^\chi$, $t \in \mathbb{N}$, agent sequence $\sigma = \sigma_s\circ(i)\circ\sigma_p \in AgSeq$ where $\sigma_p \neq \varepsilon$,

$$\left((\mathscr{I},r,t) \models \bigwedge_{j\in\sigma_s\circ(i)} correct_j \text{ and } \sigma \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t))\right) \Rightarrow \sigma_p \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t)) \tag{9}$$

Since by Lemma 3.13 it is sufficient to consider only hope chains where no agent appears twice, we define the appropriate subset of the one in Definition 3.10.

Definition 3.14.

$$Recv^i_\varphi(r_i(t)) := \{\sigma \in \widehat{Recv}^{\,i}_{\varphi}(r_i(t)) \mid (\forall k,k' \neq k \in [1,|\sigma|])\ \pi_k\sigma \neq \pi_{k'}\sigma\} \tag{10}$$

Since this refined set is a subset from (6), we immediately get the following result.

Corollary 3.15. For $\chi \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi,\pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $i,j \in \mathscr{A}$, and persistent φ,

$$(\forall\sigma \in Recv^i_\varphi(r_i(t)))\ (\mathscr{I},r,t) \models B_i\overline{H}_\sigma\varphi. \tag{11}$$

Clearly, $Recv^i_\varphi(r_i(t))$ can be easily computed by agent i in finite time (as opposed to $\widehat{Recv}^{\,i}_{\varphi}(r_i(t))$), which may contain an infinite number of hope chains), by checking its local history for the appropriate message receptions for all agents involved in some hope chain.

Obviously, if just one agent involved in a hope chain is faulty, the receiving agent i cannot infer anything meaningful from it. In a byzantine asynchronous agent context $\chi \in \mathscr{E}^{B_f}$, where at most f agents can become faulty, it thus makes sense to use multiple hope chains leading to i for implementing *fault-tolerant communication*. In order to be effective, though, the agents appearing in different hope chains must be different.

We overload the set difference operator \ for sets of sequences and sets as follows:

Definition 3.16. Given a set of sequences $\Sigma = \{\sigma_1, \dots, \sigma_k\}$, where $\sigma_\ell = (i_{\ell_1}, \dots, i_{\ell_k})$, and set $S = \{j_1, \dots, j_m\}$, let

$$\Sigma \setminus S := \{\sigma \in \Sigma \mid (\forall i \in \text{[2]}\,\sigma)\; i \notin S\}. \tag{12}$$

Definition 3.17. Given a set of sequences $\Sigma = \{\sigma_1, \dots, \sigma_k\}$, we define the set of all sets of *disjoint sequences* in Σ, which are disjoint in the sense that no sequence contains an element of another sequence from the same set, as

$$DisjSS^{\Sigma} := \{\overline{\Sigma} \subseteq \Sigma \mid\; (\forall \sigma', \sigma'' \neq \sigma' \in \overline{\Sigma})(\forall k' \in \{1, \dots, |\sigma'|\})\; \pi_{k'}\sigma' \notin \sigma''\}. \tag{13}$$

The following Theorem 3.18 shows that fault-tolerant communication via multiple hope chains is effective for agent i to obtain $B_i\varphi$, provided there are sufficiently many disjoint ones:

Theorem 3.18. For agent context $\chi \in \mathscr{E}^{B_f}$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, agent $i \in \mathscr{A}$, persistent formula φ, and a set of agents $F \subseteq \mathscr{A}$ who i believes to be faulty,

$$(\exists \Sigma' \in DisjSS^{Recv^i_\varphi(r_i(t))\setminus F})\; |\Sigma'| \;>\; f - |F| \quad \Rightarrow \quad (\mathscr{I}, r, t) \models B_i\varphi. \tag{14}$$

Apart from being informed about φ via fault-tolerant communication, agent i can of course also obtain the belief[3] $B_i\varphi$ by observing φ directly in its local history, i.e., when φ is local at i:

Theorem 3.19. For agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, agent $i \in \mathscr{A}$, local state $\lambda_0 \in \Omega_i$, and action or event $o \in Haps_i$,

$$\begin{aligned} o \in r_i(t) \quad &\Rightarrow \quad (\mathscr{I}, r, t) \models K_i occurred_i(o) \\ \lambda_0 = r_i(0) \quad &\Rightarrow \quad (\mathscr{I}, r, t) \models K_i \overline{init}_i(\lambda_0) \end{aligned} \tag{15}$$

4 Belief Gain About Faultiness

In this section, we will address the question of how agent i can establish belief about some agent j being faulty. In line with Theorem 3.18 and Theorem 3.19, there are two ways of achieving this: by direct observation, namely, receiving an obviously faulty message from j, or by receiving trustworthy notifications about j's faultiness from sufficiently many other agents. We start with the former case.

Lemma 4.1 (Directly observing others' faults). For interpreted system $\mathscr{I} = (R^\chi, \pi)$ with agent context $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon, P}, \Psi), P)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$ and agents $i, j \in \mathscr{A}$, if $(\exists \mu \in Msgs)(\forall h_j \in \mathscr{L}_j)(\forall D \in P_j(h_j))\, send(i, \mu) \notin D \,\wedge\, recv(j, \mu) \in r_i(t)$, then $(\mathscr{I}, r, t) \models B_i faulty_j$.

Note carefully that the messages that allow direct fault detection in Lemma 4.1 must indeed be obviously faulty, in the sense that they must not occur in *any* correct run. This is the case for messages that report some local history of the sending agent that is inconsistent with the local history communicated earlier, which covers the fault detection requirements in [14], for example. However, messages that could not occur just in the *specific* run $r \in R^\chi$ cannot be used for direct fault detection. We capture this by the following characterization of directly detectable faults:

Definition 4.2. For some agent context $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon, P}, \Psi), P) \in \mathscr{E}^B$, agent i and local history $h_i \in \mathscr{L}_i$, the set of agents that i beliefs to be faulty, due to having received an obviously faulty message from them, is defined as

$$\begin{aligned} \mathsf{DirObBelFaultyAg}(h_i, i) := \{ j \in \mathscr{A} \mid &(\exists \mu \in Msgs)(\forall h_j \in \mathscr{L}_j)(\forall D \in P_j(h_j)) \\ &send(i, \mu) \notin D \wedge recv(j, \mu) \in h_i \}. \end{aligned} \tag{16}$$

[2] By slight abuse of notation we write $i \in \sigma$ iff i appears somewhere in the sequence σ.

[3] Be aware of the following validities: $K_i\varphi \;\Rightarrow\; B_i\varphi$ and $K_i occurred_i(o) \;\Rightarrow\; B_i\overline{occurred}_i(o)$ (following directly from the definition of the belief modality).

Corollary 4.3. For some $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon,P}, \Psi), P) \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi, \pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $i \in \mathscr{A}$, if

$$j \in \mathsf{DirObBelFaultyAg}(r_i(t), i) \;\Rightarrow\; (\mathscr{I}, r, t) \models B_i faulty_j. \tag{17}$$

Turning our attention to the case corresponding to Theorem 3.18, namely, detecting faultiness of an agent by receiving sufficiently many trustworthy notifications from other agents in an agent context $\chi \in \mathscr{E}^{B_f}$, it seems obvious to use the following characterization:

$$\mathscr{B}(h_i, i) := \{\ell \in \mathscr{A} \mid (\exists \Sigma \in DisjSS^{Recv^i_{faulty_\ell}(h_i) \setminus \mathscr{B}(h_i,i)})\ |\Sigma| \;>\; f - |\mathscr{B}(h_i, i)|\}. \tag{18}$$

Indeed, if there are more than f minus the currently believed faulty agents disjoint hope chains for $faulty_\ell$ leading to agent i, it can safely add agent ℓ to its set of currently believed faulty agents. Unfortunately, however, this definition is cyclic, as $\mathscr{B}(h_i, i)$ appears in its own definition (with the exact same parameters): Who an agent believes to be faulty depends on who an agent already believes to be faulty. We will get rid of this problem by a fixpoint formulation, which can be solved algorithmically.

We start out from direct notifications by a faulty agent, i.e., when an agent j that has somehow detected its own faultiness (see below) informs agent i about this fact. Note that this is different from Corollary 4.3, where agent i directly observes j's misbehavior (i received an obviously faulty message).

We capture this by the following characterization of directly notified faults:

Definition 4.4. For some agent context $\chi \in \mathscr{E}^{B_f}$, agent $i \in \mathscr{A}$ and local state $h_i \in \mathscr{L}_i$, we define the set of agents, who agent i believes to be faulty due having received a direct notification from exactly those agents, as

$$\mathsf{DirNotifBelFaultyAg}(h_i, i) := \{j \in \mathscr{A} \mid (j) \in Recv^i_{faulty_j}(h_i)\}. \tag{19}$$

Lemma 4.5. For some agent context $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon,P}, \Psi), P) \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, $t \in \mathbb{N}$ and agent $i \in \mathscr{A}$, if

$$j \in \mathsf{DirNotifBelFaultyAg}(r_i(t), i) \;\Rightarrow\; (\mathscr{I}, r, t) \models B_i faulty_j. \tag{20}$$

As it is possible for an agent i in an agent context $\chi \in \mathscr{E}^{B_f}$ to sometimes also detect its own faultiness [11], we need to consider this as well. The following function returns **true** if the agent i observes some erroneous behavior in its own history $h_i = (h_i(|h_i|), h_i(|h_i| - 1), \ldots, h_i(0))$ (recall that $h_i(k)$ is the set of all haps agent i perceived in the k-th round it was active in), which implies that i itself is faulty:

Definition 4.6. For any $\chi \in \mathscr{E}$, agent $i \in \mathscr{A}$, timestamp $t \in \mathbb{N}$, and local history $h_i = r_i(t)$, let

$$\mathsf{DirObMeKnowFaulty}(r_i(t), i) := \begin{cases} \textbf{true} & \text{if } (\exists a \in Actions_i)(\exists m \in [1, |h_i| - 1]) \\ & \quad (\forall D \in P_i(\pi_m h_i, \pi_{m+1} h_i, \ldots, \pi_{|h_i|} h_i) \\ & \quad a \notin D \;\wedge\; a \in \pi_{m+1} h_i \\ \textbf{false} & \text{otherwise.} \end{cases} \tag{21}$$

Corollary 4.7 (Observing one's own faulty history). For an agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, and agent $i \in \mathscr{A}$, if

$$\mathsf{DirObMeKnowFaulty}(r_i(t), i) = \textbf{true} \quad \Rightarrow \quad (\mathscr{I}, r, t) \models K_i faulty_i \tag{22}$$

With these preparations, we are now ready to present a procedure, given in in Algorithm 1, by which any agent i can compute its belief regarding the faultiness of agents (including itself). Rather than explicitly constructing the underlying Kripke model, it exploits the *a priori* knowledge of the sets resp. the function in Definition 3.1, 4.2, 4.4 resp. 4.6. Note carefully that they can indeed be pre-computed "offline' and supplied to the algorithm via the resulting look-up tables. In sharp contrast to constructing the Kripke model, our procedure is guaranteed to terminate in a bounded number of steps.

We start our correctness proof of Algorithm 1 by showing the following invariant of the set F:

Lemma 4.8. For Algorithm 1 called with parameters (χ, h_i, i, f), where $\chi \in \mathscr{E}^{B_f}$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $i \in \mathscr{A}$, and $h_i = r_i(t)$, the following invariant holds for the variable F during

Algorithm 1 Gain-belief algorithm for faulty agents in agent context $\chi \in \mathscr{E}^{B_f}$ for agent i with history h_i

```
 1: function BeliefWhoIsFaultyAlgorithm($\chi, h_i, i, f$)
 2:     $F := \mathsf{DirObBelFaultyAg}(h_i, i) \cup \mathsf{DirNotifBelFaultyAg}(h_i, i)$
 3:     if $\mathsf{DirObMeKnowFaulty}(h_i, i) = true$ then
 4:         $F := F \cup \{i\}$
 5:     repeat
 6:         $F_{Old} := F$
 7:         for all $\ell \in \mathscr{A} \setminus F$ do
 8:             for all $\Sigma \in DisjSS^{Recv^i_{faulty_\ell}(h_i) \setminus F}$ do
 9:                 if $|\Sigma| > f - |F|$ then
10:                     $F := F \cup \{\ell\}$
11:                     continue next iteration at line 7
12:     until $F = F_{Old}$
13:     return F
```

its iterations:

$$(\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall \ell \in F)\ (r_i(t) = h_i) \ \Rightarrow\ (\mathscr{I}, r, t) \models B_i faulty_\ell. \tag{23}$$

Corollary 4.9. For agent context $\chi \in \mathscr{E}^{B_f}$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, and agent $i \in \mathscr{A}$,

$$\ell \in \texttt{BeliefWhoIsFaultyAlgorithm}(\chi, r_i(t), i, f) \ \Rightarrow\ (\mathscr{I}, r, t) \models B_i faulty_\ell. \tag{24}$$

Since it follows immediately from the definition of $\mathscr{E}^{B_f}$ that the number of faulty agents in any run $r \in R^\chi \in \mathscr{E}^{B_f}$ is at most f, the result of Algorithm 1 respects f as well:

Lemma 4.10. For $\chi \in \mathscr{E}^{B_f}$, $r \in R^\chi$, $t \in \mathbb{N}$, correct agent $i \in \mathscr{A}$, and the set F returned by Algorithm 1 $\texttt{BeliefWhoIsFaultyAlgorithm}(\chi, h_i, i, f)$, it holds that $|F| \leq f$.

The following theorem finally proves that Algorithm 1 terminates after a bounded number of steps, provided the agent context $\chi \in \mathscr{E}^{B_f}$ ensures that agent i's history is finite at every point in time, meaning $(\forall t \in \mathbb{N})(\exists b \in \mathbb{N})(\forall t' : 0 < t' \leq t)\ |r_i(t')| < b$.

Theorem 4.11. For agent context $\chi \in \mathscr{E}^{B_f}$, $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, and agent $i \in \mathscr{A}$, if $\mathscr{A}$ is finite and i's history is finite at every point in time, then the call $\texttt{BeliefWhoIsFaultyAlgorithm}$ $(\chi, r_i(t), i, f)$ invoking Algorithm 1 terminates after a bounded number of steps.

5 Belief Gain about Occurrences of Haps

In this section, we turn our attention to a sufficient condition for an agent to establish belief that a group of reliable agents (a reliable agent will stay forever correct) has obtained belief about the correct occurrence of some event or action. It follows already from Theorem 3.18 that sufficiently many disjoint hope chains for $\varphi = \bigvee_{\substack{G \subseteq \mathscr{A},\\ |G| = k}} \bigwedge_{j \in G} B_j \Box correct\, j \wedge \overline{occurred}(o)$, with $k + f \leq n$, are enough for establishing $B_i \varphi$. Theorem 5.1 adds another condition, namely, that among the disjoint hope chains for formula $\overline{occurred}(o)$, at least k are non faulty and hence truthfully deliver the information that some correct agent believes in $\overline{occurred}(o)$. Note that the two conditions are related but, in general, not identical.

Theorem 5.1. For agent context $\chi \in \mathscr{E}^{B_f}$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, action or event $o \in Haps$, agent $i \in \mathscr{A}$, natural number $k \in \mathbb{N} \setminus \{0\}$ s.t. $k + f \leq n$, and set $F \subseteq \mathscr{A}$, which

i believes to be faulty, if

$$\left(\begin{array}{r} (\exists \Sigma' \in DisjSS^{Recv^{i}_{\overline{occurred}(o)}(r_i(t))\setminus F})\ |\Sigma'| \ \geq \ k+f-|F| \quad \text{or} \\ \left(\exists \Sigma'' \in DisjSS^{Recv^{i}_{\bigvee_{G\subseteq\mathscr{A},\ |G|=k} \bigwedge_{j\in G} \Box correct_j \wedge B_j\overline{occurred}(o)}(r_i(t))\setminus F}\right) |\Sigma''| \ > \ f-|F| \end{array}\right) \tag{25}$$

$$\Rightarrow \ (\mathscr{I},r,t) \models B_i \bigvee_{\substack{G'\subseteq\mathscr{A},\\ |G'|=k}} \bigwedge_{j\in G'} \Box correct_j \wedge B_j\overline{occurred}(o). \tag{26}$$

We conclude this section by noting that the first condition in Theorem 5.1 could be strengthened by agent i also considering a possible occurrence of $\overline{occurred}(o)$ in its own history, see Theorem 3.19, in which case k can be reduced by 1 if none of the hope chains in Σ' contains i.

6 Conclusions

We provided sufficient conditions for an agent to obtain belief of (1) the faultiness of (other) agents and (2) of the occurrence of an event or action happening at some correct agent(s). Our conditions work for any agent context where at most f agents may be byzantine. They do not require the agent to compute the underlying Kripke model, but can rather be checked locally by the agent in bounded time just based on its current history. Since protocols for byzantine fault-tolerant distributed systems typically require an agent to detect (1) and/or (2), our results are important stepping stones for the development of communication-efficient protocols and for proving them correct.

Acknowledgments. We are grateful to Giorgio Cignarale, Hans van Ditmarsch, Stephan Felber, Krisztina Fruzsa, Roman Kuznets, Rojo Randrianomentsoa and Hugo Rincon Galeana for many fruitful discussions.

References

[1] J.C. Adams & K.V.S. Ramarao (1989): *Distributed diagnosis of Byzantine processors and links*. In: *[1989] Proceedings. The 9th International Conference on Distributed Computing Systems*, pp. 562–569, doi:10.1109/ICDCS.1989.37989.

[2] Armando Castañeda, Yannai A. Gonczarowski & Yoram Moses (2022): *Unbeatable Consensus*. *Distrib. Comput.* 35(2), p. 123–143, doi:10.1007/s00446-021-00417-3.

[3] Hans van Ditmarsch, Krisztina Fruzsa & Roman Kuznets (2022): *A New Hope*. In David Fernández-Duque, Alessandra Palmigiano & Sophie Pinchinat, editors: *Advances in Modal Logic*, 14, College Publications, pp. 349–369. Available at `https://doi.org/10.34726/2821`.

[4] Cynthia Dwork, Nancy Lynch & Larry Stockmeyer (1988): *Consensus in the Presence of Partial Synchrony*. *Journal of the ACM* 35(2), pp. 288–323, doi:10.1145/42282.42283.

[5] Cynthia Dwork & Yoram Moses (1990): *Knowledge and Common Knowledge in a Byzantine Environment: Crash Failures*. *Information and Computation* 88, pp. 156–186, doi:10.1016/0890-5401(90)90014-9.

[6] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Y. Vardi (1995): *Reasoning About Knowledge*. MIT Press.

[7] Krisztina Fruzsa (2019): *Hope for Epistemic Reasoning with Faulty Agents!* In: *Proc. 31st European Summer School in Logic, Language and In formation (ESSLLI 2019) Student Session*, FOLLI, , Riga,

Latvia, pp. 169–180. Available at `http://esslli2019.folli.info/wp-content/uploads/2019/08/tentative_proceedings.pdf`.

[8] Krisztina Fruzsa, Roman Kuznets & Ulrich Schmid (2021): *Fire!* In Joseph Y. Halpern & Andrés Perea, editors: *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2021, Beijing, China, June 25-27, 2021*, *EPTCS* 335, pp. 139–153, doi:10.4204/EPTCS.335.13.

[9] Joseph Y. Halpern & Yoram Moses (1990): *Knowledge and Common Knowledge in a Distributed Environment*. *Journal of the ACM* 37, pp. 549–587, doi:10.1145/79147.79161.

[10] Roman Kuznets, Laurent Prosperi, Ulrich Schmid & Krisztina Fruzsa (2019): *Causality and Epistemic Reasoning in Byzantine Multi-Agent Systems*. In Lawrence S. Moss, editor: *TARK 2019*, *Electronic Proceedings in Theoretical Computer Science* 297, Open Publishing Association, pp. 293–312, doi:10.4204/EPTCS.297.19.

[11] Roman Kuznets, Laurent Prosperi, Ulrich Schmid & Krisztina Fruzsa (2019): *Epistemic Reasoning with Byzantine-Faulty Agents*. In Andreas Herzig & Andrei Popescu, editors: *FroCoS 2019*, *Lecture Notes in Artificial Intelligence* 11715, Springer, pp. 259–276, doi:10.1007/978-3-030-29007-8_15.

[12] Roman Kuznets, Laurent Prosperi, Ulrich Schmid, Krisztina Fruzsa & Lucas Gréaux (2019): *Knowledge in Byzantine Message-Passing Systems I: Framework and the Causal Cone*. Technical Report TUW-260549, TU Wien. Available at `https://publik.tuwien.ac.at/files/publik_260549.pdf`.

[13] Leslie Lamport, Robert Shostak & Marshall Pease (1982): *The Byzantine generals problem*. *ACM Transactions on Programming Languages and Systems* 4(3), pp. 382–401, doi:10.1145/357172.357176.

[14] Hammurabi Mendes, Christine Tasson & Maurice Herlihy (2014): *Distributed Computability in Byzantine Asynchronous Systems*. In: *STOC 2014: Symposium on Theory of Computing*, ACM, pp. 704–713, doi:10.1145/2591796.2591853.

[15] Yoram Moses (2015): *Relating Knowledge and Coordinated Action: The Knowledge of Preconditions Principle*. In: *TARK 2015*, pp. 231–245, doi:10.4204/EPTCS.215.17.

[16] Yoram Moses & Yoav Shoham (1993): *Belief as defeasible knowledge*. *Artificial Intelligence* 64, pp. 299–321, doi:10.1016/0004-3702(93)90107-M.

[17] Yoram Moses & Mark R. Tuttle (1986): *Programming Simultaneous Actions Using Common Knowledge: Preliminary Version*. In: *FOCS 1986*, IEEE, pp. 208–221, doi:10.1109/SFCS.1986.46.

[18] Yoram Moses & Mark R. Tuttle (1988): *Programming Simultaneous Actions Using Common Knowledge*. *Algorithmica* 3, pp. 121–169, doi:10.1007/BF01762112.

[19] D. Powell, J. Arlat, L. Beus-Dukic, A. Bondavalli, P. Coppola, A. Fantechi, E. Jenn, C. Rabejac & A. Wellings (1999): *GUARDS: a generic upgradable architecture for real-time dependable systems*. *IEEE Transactions on Parallel and Distributed Systems* 10(6), pp. 580–599, doi:10.1109/71.774908.

[20] Peter Robinson & Ulrich Schmid (2011): *The Asynchronous Bounded-Cycle model*. *Theoretical Computer Science* 412, pp. 5580–5601, doi:10.1016/j.tcs.2010.08.001.

[21] Thomas Schlögl, Ulrich Schmid & Roman Kuznets (2021): *The Persistence of False Memory: Brain in a Vat Despite Perfect Clocks*. In Takahiro Uchiya, Quan Bai & Iván Marsá Maestre, editors: *PRIMA 2020: Principles and Practice of Multi-Agent Systems*, Springer International Publishing, Cham, pp. 403–411, doi:10.1007/978-3-030-69322-0_30.

[22] T. K. Srikanth & Sam Toueg (1987): *Optimal Clock Synchronization*. *Journal of the ACM* 34, pp. 626–645, doi:10.1145/28869.28876.

[23] T. K. Srikanth & Sam Toueg (1987): *Simulating authenticated broadcasts to derive simple fault-tolerant algorithms*. *Distributed Computing* 2, pp. 80–94, doi:10.1007/BF01667080.

[24] Josef Widder & Ulrich Schmid (2009): *The Theta-Model: achieving synchrony without clocks*. *Distributed Computing* 22, pp. 29–47, doi:10.1007/s00446-009-0080-x.

A Additional details for Section 2 (The Basic Model)

A.1 Global haps and faults.

As already mentioned, there is a global version of every *Haps* that provides additional information that is only accessible to the environment. Among it is the timestamp t For correct action $a \in Actions_i$, as initiated by agent i in the local format, the one-to-one function $global(i,t,a)$ gives the global version. Timestamps are especially crucial for proper message processing with $global(i,t,send(j,\mu_k)) := gsend(i,j,\mu,id(i,j,\mu,k,t))$ for some one-to-one function $id\colon \mathscr{A} \times \mathscr{A} \times Msgs \times \mathbb{N} \times \mathbb{T} \to \mathbb{N}$ that assigns each sent message a unique **global message identifier** (GMI). These GMIs enable the direct linking of send actions to their corresponding delivery events, most importantly used to ensure that only sent messages can be delivered (causality).

Unlike correct actions, correct events witnessed by agent i are generated by the environment ε, hence are already produced in the global format $\overline{GEvents_i}$. For each correct event $E \in \overline{GEvents_i}$, we use a faulty counterpart $fake(i,E)$ and will make sure that agent i cannot distinguish between the two. An important type of correct global events is delivery $grecv(j,i,\mu,id) \in \overline{GEvents_i}$ of message μ with GMI $id \in \mathbb{N}$ sent from agent i to agent j. The GMI must be a part of the global format (especially for ensuring causality) but cannot be part of the local format because it contains information about the time of sending, which should not be accessible to agents. The stripping of this information before updating local histories is achieved by the function $local\colon \overline{GHaps} \longrightarrow Haps$ converting **correct** haps from the global into the local formats for the respective agents in such a way that $local$ reverses $global$, i.e., $local\big(global(i,t,a)\big) := a$, in particular, $local\big(grecv(i,j,\mu,id)\big) := recv(j,\mu)$.

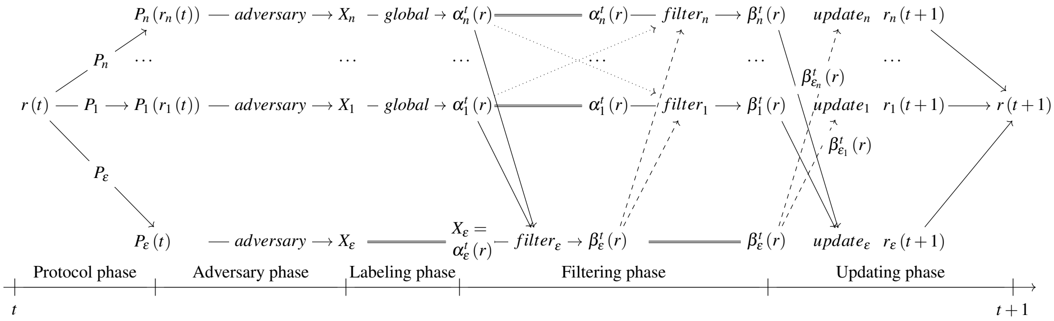
Faulty actions are modeled as byzantine events of the form $fake(i,A \mapsto A')$ where $A,A' \in \overline{GActions_i} \sqcup \{\mathbf{noop}\}$ for a special **non-action noop** in global format. These byzantine events are controlled by the environment and correspond to an agent violating its protocol by performing the action A, while recording in its local history that it either performs $a' = local(A') \in Actions_i$ if $A' \in \overline{GActions_i}$ or does nothing if $A' = \mathbf{noop}$.

A.2 Protocols, state transitions and runs.

The events and actions that occur in each round are determined by protocols (for agents and the environment) and non-determinism (adversary). Agent i's **protocol** $P_i\colon \mathscr{L}_i \to 2^{2^{Actions_i}} \setminus \{\varnothing\}$ provides a range $P_i(r_i(t))$ of sets of actions based on i's current local state $r_i(t) \in \mathscr{L}_i$ at time t in run r, from which the adversary non-deterministically picks one. Similarly the environment provides a range of (correct, byzantine, and system) events via its protocol $P_\varepsilon\colon \mathbb{T} \to 2^{2^{GEvents}} \setminus \{\varnothing\}$, which depends on a timestamp $t \in \mathbb{T}$ but **not** on the current state, in order to maintain its impartiality. It is required that all events of round $t½$ be mutually compatible at time t, called t-coherent according to Definition A.1. The set of all global states is denoted by $\mathscr{G}$.

Definition A.1 (Coherent events)**.** Let $t \in \mathbb{N}$ be a timestamp. A set $S \subset GEvents$ of events is called t**-coherent** if it satisfies the following conditions:

1. for any $fake(i,gsend(i,j,\mu,id) \mapsto A) \in S$, the GMI $id = id(i,j,\mu,k,t)$ for some $k \in \mathbb{N}$;
2. for any $i \in \mathscr{A}$ at most one of $go(i)$, $sleep(i)$, and $hibernate(i)$ is present in S;
3. for any $i \in \mathscr{A}$ and any $e \in Ext_i$ at most one of $ext(i,e)$ and $fake(i,ext(i,e))$ is present in S;
4. for any $grecv(i,j,\mu,id_1) \in S$, no event of the form $fake(i,grecv(i,j,\mu,id_2))$ belongs to S for any $id_2 \in \mathbb{N}$;



$P_n(r_n(t))$ — adversary ⟶ X_n – global → $\alpha_n^t(r)$ — $\alpha_n^t(r)$ — $filter_n$ ⟶ $\beta_n^t(r)$ $update_n$ $r_n(t+1)$
P_n
…
$r(t)$ — P_1 ⟶ $P_1(r_1(t))$ — adversary ⟶ X_1 – global → $\alpha_1^t(r)$ — $\alpha_1^t(r)$ — $filter_1$ ⟶ $\beta_1^t(r)$ $update_1$ $r_1(t+1)$ ⟶ $r(t+1)$
$\beta_{\varepsilon_n}^t(r)$
$\beta_{\varepsilon_1}^t(r)$
P_ε
$P_\varepsilon(t)$ — adversary ⟶ X_ε — $X_\varepsilon = \alpha_\varepsilon^t(r)$ — $filter_\varepsilon$ → $\beta_\varepsilon^t(r)$ — $\beta_\varepsilon^t(r)$ $update_\varepsilon$ $r_\varepsilon(t+1)$
Protocol phase
Adversary phase
Labeling phase
Filtering phase
Updating phase
t
$t+1$

Figure 1: The evolution of states in round $t.5$ (from timestamp $t \in \mathbb{N}$ to $t+1$) inside a run r constructed according to the transition function $\tau_{P_\varepsilon,P}$. Different communication models require changes to the filtering functions $filter_\varepsilon$ and $filter_i$.

5. for any $fake(i, grecv(i,j,\mu,id_1)) \in S$, no event of the form $grecv(i,j,\mu,id_2)$ belongs to S for any $id_2 \in \mathbb{N}$.

Given the **joint protocol** $P := (P_1,\dots,P_n)$ and the environment's protocol P_ε, we focus on $\tau_{P_\varepsilon,P}$-**transitional runs** r that result from following these protocols and are built according to a **transition relation** $\tau_{P_\varepsilon,P} \subseteq \mathscr{G} \times \mathscr{G}$. Each such transitional run begins in some initial global state $r(0) \in \mathscr{G}(0)$ and progresses, satisfying $(r(t), r(t+1)) \in \tau_{P_\varepsilon,P}$ for each timestamp $t \in \mathbb{T}$.

The transition relation $\tau_{P_\varepsilon,P}$ consisting of five consecutive phases is illustrated in Fig. 1 and works as follows:

In the **protocol phase**, a range $P_\varepsilon(t) \subset 2^{GEvents}$ of t-coherent sets of events is determined by the environment's protocol P_ε. Similarly for each $i \in \mathscr{A}$, a range $P_i(r_i(t)) \subseteq 2^{Actions_i}$ of sets of i's actions is determined by the joint protocol P.

In the **adversary phase**, the adversary non-deterministically chooses a set $X_\varepsilon \in P_\varepsilon(t)$ and one set $X_i \in P_i(r_i(t))$ for each $i \in \mathscr{A}$.

In the **labeling phase**, actions in the sets X_i are translated into the global format: $\alpha_i^t(r) := \{global(i,t,a) \mid a \in X_i\} \subseteq \overline{GActions_i}$.

In the **filtering phase**, filter functions remove all unwanted or impossible attempted events from $\alpha_\varepsilon^t(r) := X_\varepsilon$ and actions from $\alpha_i^t(r)$. This is done in two stages:
First, $filter_\varepsilon$ filters out "illegal" events. This filter will vary depending on the concrete system assumptions (in the byzantine asynchronous case, "illegal" constitutes receive events that violate causality). The resulting set of events to actually occur in round t½ is $\beta_\varepsilon^t(r) := filter_\varepsilon\left(r(t), \alpha_\varepsilon^t(r), \alpha_1^t(r), \dots, \alpha_n^t(r)\right)$. The byzantine asynchronous filter (ensuring causality) is denoted by $filter_\varepsilon^B(h, X_\varepsilon, X_1, \dots, X_n)$ and the byzantine asynchronous at-most-f-faulty-agents filter, which both ensures causality and removes all byzantine events if as a result the total number of faulty agents were to exceed f is denoted by $filter_\varepsilon^{B_f}(h, X_\varepsilon, X_1, \dots, X_n)$.

Definition A.2. The **standard action filter** $filter_i^B(X_1,\dots,X_n,X_\varepsilon)$ for $i \in \mathscr{A}$ either removes all actions from X_i when $go(i) \notin X_\varepsilon$ or else leaves X_i unchanged.

Second, $filter_i^B$ for each i returns the sets of actions to be actually performed by agents in round t½, i.e., $\beta_i^t(r) := filter_i^B\left(\alpha_1^t(r), \dots, \alpha_n^t(r), \beta_\varepsilon^t(r)\right)$. Note that $\beta_i^t(r) \subseteq \alpha_i^t(r) \subseteq \overline{GActions_i}$ and $\beta_\varepsilon^t(r) \subseteq \alpha_\varepsilon^t(r) \subset GEvents$.

In the **updating phase**, the events $\beta_\varepsilon^t(r)$ and actions $\beta_i^t(r)$ are appended to the global history $r(t)$. For each $i \in \mathscr{A}$, all non-system events from $\beta_{\varepsilon_i}^t(r) := \beta_\varepsilon^t(r) \cap GEvents_i$ and all actions $\beta_i^t(r)$ as **perceived** by the agent are appended in the local form to the local history $r_i(t)$. Note the local history may remain unchanged if no events trigger an update.

Definition A.3 (State update functions)**.** Given a global history $h = (h_\varepsilon, h_1, \ldots, h_n) \in \mathscr{G}$, a tuple of performed actions/events $X = (X_\varepsilon, X_1, \ldots, X_n) \in 2^{GEvents} \times 2^{\overline{GActions}_1} \times \ldots \times 2^{\overline{GActions}_n}$, we use the following abbreviation $X_{\varepsilon_i} = X_\varepsilon \cap GEvents_i$ for each $i \in \mathscr{A}$. Agents i's update function $update_i \colon \mathscr{L}_i \times 2^{\overline{GActions}_i} \times 2^{GEvents} \to \mathscr{L}_i$ outputs a new local history from $\mathscr{L}_i$ based on i's actions X_i and events X_ε as follows:

$$update_i\,(h_i, X_i, X_\varepsilon) := \begin{cases} h_i & \text{if } \sigma(X_{\varepsilon_i}) = \varnothing \text{ and } go(i) \notin X_\varepsilon \\ \left[\sigma\left(X_{\varepsilon_i} \sqcup X_i\right)\right] \circ h_i & \text{otherwise} \end{cases} \tag{27}$$

where $\sigma(X)$ removes all system events *SysEvents* from X and afterwards invokes *local* on the resulting set. Similarly, the environment's state update function $update_\varepsilon \colon \mathscr{L}_\varepsilon \times \left(2^{GEvents} \times 2^{\overline{GActions}_1} \times \ldots \times 2^{\overline{GActions}_n}\right) \to \mathscr{L}_\varepsilon$ outputs a new state of the environment based on X_ε:

$$update_\varepsilon\,(h_\varepsilon, X) := (X_\varepsilon \sqcup X_1 \sqcup \ldots \sqcup X_n) \colon h_\varepsilon \tag{28}$$

Thus, the global state is modified as follows:

$$update\,(h, X) := \left(update_\varepsilon\,(h_\varepsilon, X), update_1\,(h_1, X_1, X_\varepsilon), \ldots, update_n\,(h_n, X_n, X_\varepsilon)\right) \text{ and} \tag{29}$$

$$r_\varepsilon\,(t+1) := update_\varepsilon\left(r_\varepsilon\,(t), \quad \beta^t_\varepsilon\,(r), \quad \beta^t_1\,(r), \quad \ldots, \quad \beta^t_n\,(r)\right) \tag{30}$$

$$r_i\,(t+1) := update_i\left(r_i\,(t), \quad \beta^t_i\,(r), \quad \beta^t_\varepsilon\,(r)\right). \tag{31}$$

The operations in the phases 2–5 (adversary, labeling, filtering and updating phase) are grouped into a **transition template** τ that yields a transition relation $\tau_{P_\varepsilon, P}$ for any joint and environment protocol P and P_ε. Particularly, we denote as τ^B the transition template utilizing $filter^B_\varepsilon$ and $filter^B_i$ (for all $i \in \mathscr{A}$).

B Additional details for Section 3 (Basics of Fault-Tolerant Communication)

Lemma 3.7. For agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, $t \in \mathbb{N}$, agents $i, j \in \mathscr{A}$, persistent formula φ, and trustworthy message $\mu \in Msgs^{j \to i}_\varphi$

$$recv(j, \mu) \in r_i(t) \quad \Rightarrow \quad (\mathscr{I}, r, t) \models B_i H_j \varphi. \tag{3}$$

Proof.

1. $\mu \in Msgs^{j \to i}_\varphi$ (by assumption)
2. $recv(j, \mu) \in r_i(t)$ (by assumption)
3. $(\mathscr{I}, r, t) \not\models B_i H_j \varphi$ (by assumption)
4. $\widehat{r} \in R^\chi$ and $\widehat{t} \in \mathbb{N}$ and $r(t) \sim_i \widehat{r}(\widehat{t})$ and $(\mathscr{I}, \widehat{r}, \widehat{t}) \models correct_i \wedge \neg H_j \varphi$ (from line (3), by sem. of B_i, exist. inst.)
5. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models occurred_i(recv(j, \mu))$ (from lines (2), (4), by def. of $\sim_i$, sem. of $occurred()$, "and")
6. (a) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models fake_i\,(recv(j, \mu))$ (from line (5), by sem. of $occurred_i(), fake_i\,()$)
   (b) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models faulty_i$ (from line (6a), by sem. of $fake_i\,()$)
   (c) contradiction! (from lines (4), (6b), by sem. of "and")
7. (a) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models \overline{occurred}_i(recv(j, \mu))$ (from line (5), by sem. of $occurred_i(), \overline{occurred}_i()$)
   (b) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models happened_j(send(i, \mu))$ (from line (7a), by def. of $filter^B_\varepsilon$)
   (c) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models correct_i \wedge correct_j \wedge \neg B_j \varphi$ (from line (4), by sem. of $\neg H_j$)
   (d) i. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models fhappened_j(send(i, \mu))$ (from line (7b), by sem. of $happened_i(), fhappened_i()$)
   ii. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models faulty_j$ (from line (7(d)i), by sem. of $fhappened_j()$)
   iii. contradiction! (from lines (7c), (7(d)ii), by sem. of $\wedge$)
   (e) i. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models \overline{occurred}_j(send(i, \mu))$ (from line (7b), by sem. of $happened_j()$)

ii. $(\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall D \in P_j(r_j(t)))\ send(i,\mu) \in D \Rightarrow (\mathscr{I},r,t) \models B_j\varphi$ (from line (1), by Definition 3.1)

iii. $(\mathscr{I},\widehat{r},\widehat{t}) \models B_j\varphi$ (from lines (7(e)i), (7(e)ii), by univ. inst. sem. of $\overline{occurred}_j()$, $\Rightarrow$, persistence of φ and Lemma 3.5)

iv. contradiction! (from lines (7c), (7(e)iii), by sem. of $\wedge$)

□

Lemma 3.12. For persistent φ, $\chi \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi,\pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $\widehat{\sigma} \in \widehat{Recv}^{i}_{\varphi}(r_i(t))$, $\sigma = (i)\circ\widehat{\sigma}$,

$$(\forall k \in [1,|\sigma|-1])\ (\mathscr{I},r,t) \models \bigwedge_{k' \in [1,k]} correct_{\pi_{k'}\sigma} \quad \Rightarrow \quad (\mathscr{I},r,t) \models \overline{H}_{\pi_{k+1}\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi \tag{8}$$

Proof. by induction over $k \in [1,|\sigma|-1]$.

Ind. hyp: $(\forall k \in [1,|\sigma|-1])\ (\mathscr{I},r,t) \models \bigwedge_{k' \in [1,k]} correct_{\pi_{k'}\sigma} \quad \Rightarrow \quad (\mathscr{I},r,t) \models \overline{H}_{\pi_{k+1}\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$

Base case for $k=1$: by contradiction.

1. $(\mathscr{I},r,t) \not\models \overline{H}_{\pi_2\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (by contr. assumption)
2. $(\mathscr{I},r,t) \models \overline{occurred}_i(recv(\pi_1\sigma,\mu))$ and $\mu \in Msgs^{\pi_2\sigma\to i}_{\overline{H}_{\pi_3\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi}$ (since $\widehat{\sigma} \in \widehat{Recv}^{i}_{\varphi}(r(t))$, $(\mathscr{I},r,t) \models correct_i$, by Definition 3.10)
3. $(\mathscr{I},r,t) \models B_i\overline{H}_{\pi_2\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (from line (2), by Corollary 3.11)
4. $(\mathscr{I},r,t) \models \overline{H}_{\pi_2\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (from line (3), since $(\mathscr{I},r,t) \models correct_i$)
5. contradiction! (from lines (1), (4))

Ind. step $k-1 \to k$ (for $k \in [2,|\sigma|-1]$): by contradiction.

1. $\widehat{\sigma} \in \widehat{Recv}^{i}_{\varphi}(r(t))$ and $\sigma = (i)\circ\widehat{\sigma}$ (by contr. assumption)
2. $(\mathscr{I},r,t) \models \bigwedge_{k' \in [1,k]} correct_{\pi_{k'}\sigma}$ (by contr. assumption, sem. of $\Rightarrow$)
3. $(\mathscr{I},r,t) \not\models \overline{H}_{\pi_{k+1}\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (by contr. assumption)
4. $(\mathscr{I},r,t) \models \overline{H}_{\pi_k\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (by assumption of the ind. hyp. for $k-1$)
5. $(\mathscr{I},r,t) \models B_{\pi_k\sigma}\overline{H}_{\pi_{k+1}\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (from lines (2), (4), by sem. of $\bigwedge$, $B_{\pi_k\sigma}$)
6. $(\mathscr{I},r,t) \models \overline{H}_{\pi_{k+1}\sigma\circ\ldots\circ\pi_{|\sigma|}\sigma}\varphi$ (from lines (2), (5), by sem. of $B_{\pi_k\sigma}$, $\bigwedge$, reflexivity of $\sim_{\pi_k\sigma}$)
7. contradiction! (from lines (3), (6))

□

Lemma 3.13. For persistent formula φ, $\chi \in \mathscr{E}^B$, $r \in R^\chi$, $t \in \mathbb{N}$, agent sequence $\sigma = \sigma_s \circ (i) \circ \sigma_p \in AgSeq$ where $\sigma_p \neq \varepsilon$,

$$\left((\mathscr{I},r,t) \models \bigwedge_{j \in \sigma_s\circ(i)} correct_j \text{ and } \sigma \in \widehat{Recv}^{i}_{\varphi}(r_i(t)) \right) \Rightarrow \sigma_p \in \widehat{Recv}^{i}_{\varphi}(r_i(t)) \tag{9}$$

Proof. by contradiction.

1. $\sigma_p \neq \varepsilon$ (contr. assumption)
2. $\sigma_s \circ (i) \circ \sigma_p \in \widehat{Recv}^{i}_{\varphi}(r_i(t))$ (by contr. assumption)
3. $(\forall j \in \sigma_s \circ (i))\ (\mathscr{I},r,t) \models correct_j$ (by contr. assumption)
4. $\sigma_p \notin \widehat{Recv}^{i}_{\varphi}(r_i(t))$ (by contr. assumption)

5. $(\mathscr{I},r,t) \models \overline{H}_{(i)\circ\sigma_p}\varphi$ (from lines (2), (3), by Lemma 3.12 for $k = |\sigma_s|+1$, since in Lemma 3.12 the hope chain is extended by singleton sequence (i))
6. $(\mathscr{I},r,t) \models \overline{occurred}_i(o)$ and $\Big((\mathscr{I},r,t) \models \overline{occurred}_i(o) \Rightarrow \overline{H}_{(i)\circ\sigma_p}\varphi\Big)$ (from line (5), by [10, Theorem 13], by exist. inst. for o)
7. $o = recv(\pi_1\sigma_p,\mu)$ and $\mu \in Msgs^{\pi_1\sigma_p \to i}_{\overline{H}_{\pi_2\sigma_p\circ\ldots\circ\pi_{|\sigma_p|}\sigma_p}\varphi}$ (from lines (1), (6)[4], by Definition 3.1, by Definition 3.1, sem. of H, B)
8. $recv(\pi_1\sigma_p,\mu) \in r_i(t)$ (from lines (6), (7), by sem. of $\overline{occurred}_i(o)$)
9. $\sigma_p \in \widehat{Recv}^{i}_{\varphi}(r_i(t))$ (from lines (7), (8), by Definition 3.10)
10. contradiction! (from lines (4), (9))

□

Theorem 3.18. For agent context $\chi \in \mathscr{E}^{B_f}$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, agent $i \in \mathscr{A}$, persistent formula φ, and a set of agents $F \subseteq \mathscr{A}$ who i believes to be faulty,

$$(\exists \Sigma' \in DisjSS^{Recv^i_\varphi(r_i(t))\setminus F})\ |\Sigma'| \ > \ f - |F| \quad \Rightarrow \quad (\mathscr{I},r,t) \models B_i\varphi. \tag{14}$$

Proof.
1. $(\forall k \in F)\ (\mathscr{I},r,t) \models B_i faulty_k$ (by assumption)
2. $(\exists \Sigma' \in DisjSS^{Recv^i_\varphi(r_i(t))\setminus F})\ |\Sigma'| \ > \ f - |F|$ (by assumption)
3. $(\mathscr{I},r,t) \not\models B_i\varphi$ (by assumption)
4. $\Sigma \in DisjSS^{Recv^i_\varphi(r_i(t))\setminus F}$ (simultaneous with line (5), from line (2) , by exist. inst.)
5. $|\Sigma| \ > \ f - |F|$ (simultaneous with line (4), from line (2), by exist. inst.)
6. $(\forall \sigma' \in \Sigma)\ (\mathscr{I},r,t) \models B_i\overline{H}_{\sigma'}\varphi$ (from line (4), by Corollary 3.15 and Definition 3.17)
7. $\widetilde{r} \in R^\chi$ and $\widetilde{t} \in \mathbb{N}$ and $r(t) \sim_i \widetilde{r}(\widetilde{t})$ and $(\mathscr{I},\widetilde{r},\widetilde{t}) \models correct_i \wedge \neg\varphi$ (from line (3), by sem. of B_i, $\not\models$, exist. inst.)
8. $(\forall k \in F)\ (\mathscr{I},\widetilde{r},\widetilde{t}) \models correct_i \to faulty_k$ (from lines (1), (7), by sem. of B_i, $\Rightarrow$, $\wedge$, $\sim_i$, univ. inst.)
9. $(\forall \sigma' \in \Sigma)\ (\mathscr{I},\widetilde{r},\widetilde{t}) \models correct_i \to \overline{H}_{\sigma'}\varphi$ (from lines (6), (7), by definition of $\sim_i$, sem. of B_i, $\Rightarrow$, $\wedge$, univ. inst.)
10. $(\forall k \in F)\ (\mathscr{I},\widetilde{r},\widetilde{t}) \models faulty_k$ (from lines (7), (8), by sem. of $\to$, "and", $\wedge$)
11. $(\forall \sigma' \in \Sigma)\ (\mathscr{I},\widetilde{r},\widetilde{t}) \models \overline{H}_{\sigma'}\varphi$ (from lines (7), (9), by sem. of $\to$, "and", $\wedge$)
12. $\sigma \in \Sigma$ and $(\forall l \in \{1,\ldots,|\sigma|\})\ (\mathscr{I},\widetilde{r},\widetilde{t}) \models correct_{\pi_l\sigma}$ ((4), (5), (7), (10), by def. of $correct_k$, $faulty_k$, $filter^{B_f}_\varepsilon$, $\forall$, $\sim_i$, exist. inst., via pigeonhole argument)
13. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models \varphi$ (from lines (11), (12), by sem. of "and", $\forall$, $\to$, H, Definition 3.9, univ. inst. and reflexivity of $\sim_{\widetilde{j}}$)
14. contradiction! (from lines (7), (13), by sem. of "and", $\wedge$)

□

Theorem 3.19. For agent context $\chi \in \mathscr{E}^B$, interpreted system $\mathscr{I} = (R^\chi,\pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, agent $i \in \mathscr{A}$, local state $\lambda_0 \in \Omega_i$, and action or event $o \in Haps_i$,

$$\begin{aligned} o \in r_i(t) \quad &\Rightarrow \quad (\mathscr{I},r,t) \models K_i occurred_i(o) \\ \lambda_0 = r_i(0) \quad &\Rightarrow \quad (\mathscr{I},r,t) \models K_i\overline{init}_i(\lambda_0) \end{aligned} \tag{15}$$

Proof. Regarding the first line:
1. $o \in r_i(t)$ (by contr. assumption)
2. $(\mathscr{I},r,t) \not\models K_i occurred_i(o)$ (by contr. assumption)
3. $\widetilde{r} \in R^\chi$ and $\widetilde{t} \in \mathbb{N}$ and $r(t) \sim_i \widetilde{r}(\widetilde{t})$ and $(\mathscr{I},\widetilde{r},\widetilde{t}) \not\models occurred_i(o)$ (from line (2), by sem. of $\not\models$, K_i, exist. inst.)
4. $r_i(t) = \widetilde{r}_i(\widetilde{t})$ (from line (3), by sem. of $\sim_i$)
5. $o \in \widetilde{r}(\widetilde{t})$ (from lines (1), (4), by sem. of $=$)
6. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models occurred_i(o)$ (from line (5), by sem. of $occurred_i(o)$)

[4] Note that even if the information about $\overline{H}_{(i)\circ\sigma_p}\varphi$ propagated to i via a local edge in the reliable causal cone [10] in σ, this information must have reached agent i initially via some message if $\sigma_p \neq \varepsilon$.

7. contradiction! (from lines (3), (6))

Regarding the second line:

1. $\lambda_0 = r_i(0)$ (by contr. assumption)
2. $(\mathscr{I}, r, t) \not\models K_i\overline{init}_i(\lambda_0)$ (by contr. assumption)
3. $\widetilde{r} \in R^\chi$ and $\widetilde{t} \in \mathbb{N}$ and $r(t) \sim_i \widetilde{r}(\widetilde{t})$ and $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \not\models \overline{init}_i(\lambda_0)$ (from line (2), by sem. of $\not\models$, K_i, exist. inst.)
4. $r_i(t) = \widetilde{r}_i(\widetilde{t})$ (from line (3), by sem. of $\sim_i$)
5. $\lambda_0 = \widetilde{r}(\widetilde{0})$ (from lines (1), (4), by sem. of $=$)
6. $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \overline{init}_i(\lambda_0)$ (from line (5), by sem. of $\overline{init}_i(\lambda_0)$)
7. contradiction! (from lines (3), (6))

□

C Additional details for Section 4 (Belief Gain About Faultiness)

Lemma 4.1 (Directly observing others' faults). For interpreted system $\mathscr{I} = (R^\chi, \pi)$ with agent context $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon, P}, \Psi), P)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$ and agents $i, j \in \mathscr{A}$, if $(\exists \mu \in Msgs)(\forall h_j \in \mathscr{L}_j)(\forall D \in P_j(h_j))\, send(i, \mu) \notin D \wedge recv(j, \mu) \in r_i(t)$, then $(\mathscr{I}, r, t) \models B_i faulty_j$.

Proof. 1. $(\exists \mu \in Msgs)(\forall h_j \in \mathscr{L}_j)(\forall D \in P_j(h_j))\ send(i, \mu) \notin D$ (by contr. assumption)

2. $recv(j, \mu) \in r_i(t)$ (by contr. assumption)
3. $(\mathscr{I}, r, t) \not\models B_i faulty_j$ (by contr. assumption)
4. $(\forall \widetilde{r} \in R^\chi)(\forall \widetilde{t} \in \mathbb{N})\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \not\models \overline{occurred}_j(send(i, \mu))$ (from line (1), by def. of $\overline{occurred}(j)$)
5. $\widehat{r} \in R^\chi$ and $\widehat{t} \in \mathbb{N}$ and $r(t) \sim_i \widehat{r}(\widehat{t})$ and
   $(\mathscr{I}, \widehat{r}, \widehat{t}) \models correct_i \wedge correct_j$ (from line (3), by sem. of B_i, $\not\models$ and exist. inst.)
6. $r_i(t) = \widehat{r}_i(\widehat{t})$ (from line (5), by def. of $\sim_i$)
7. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models correct_i$ (from line (5), by sem. of $\wedge$)
8. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models correct_j$ (from line (5), by sem. of $\wedge$)
9. $(\mathscr{I}, r, t) \models occurred_i(recv(j, \mu))$ (from line (2), by def. of $occurred_i()$)
10. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models occurred_i(recv(j, \mu))$ (from line (6), (9), by sem. of $=$, $occurred_i()$)
11. $(\mathscr{I}, \widehat{r}, \widehat{t}) \models \overline{occurred}_i(recv(j, \mu)) \vee fake_i(recv(j, \mu))$ (from line (10), by def. of $occurred_i()$)
12. (a) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models \overline{occurred}_i(recv(j, \mu))$ (from line (11), by sem. of $\vee$)
    (b) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models fhappened_j(send(i, \mu))$ (from lines (4), (12a) by def. of $filter^B_\varepsilon$)
    (c) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models faulty_j$ (from line (12b), by def. of $fhappened_j()$)
    (d) contradiction! (from lines (8), (12c))
13. (a) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models fake_i(recv(j, \mu))$ (from line (11), by sem. of $\vee$)
    (b) $(\mathscr{I}, \widehat{r}, \widehat{t}) \models faulty_i$ (from line (13a), by def. of $fake_i()$)
    (c) contradiction! (from lines (7), (13b))

□

Lemma 4.5. For some agent context $\chi = ((P_\varepsilon, \mathscr{G}(0), \tau^B_{P_\varepsilon, P}, \Psi), P) \in \mathscr{E}^B$, $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, $t \in \mathbb{N}$ and agent $i \in \mathscr{A}$, if

$$j \in \mathsf{DirNotifBelFaultyAg}(r_i(t), i) \Rightarrow (\mathscr{I}, r, t) \models B_i faulty_j. \tag{20}$$

Proof. 1. $(j) \in Recv^i_{faulty_j}(r_i(t))$ (by contr. assumption, Definition 4.4)

2. $(\mathscr{I}, r, t) \not\models B_i faulty_j$ (by contr. assumption, Definition 4.4)
3. $\widetilde{r} \in R^\chi$ and $\widetilde{t} \in \mathbb{N}$ and $r(t) \sim_i \widetilde{r}(\widetilde{t})$ (from line (2), by sem. of B_i and exist. inst.)
4. $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \models correct_i \wedge correct_j$ (from line (2), by sem. of B_i and exist. inst.)

5. $\mu \in Msgs^{j \to i}_{faulty_j}$ (from line (1), by Definition 3.14 and exist. inst. for μ)
6. $recv(j,\mu) \in r_i(t)$ (from line (1), by Definition 3.14 and exist. inst. for μ)
7. $recv(j,\mu) \in \widetilde{r}_i(\widetilde{t})$ (from lines (3), (6), by def. of $\sim_i$)
8. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models \overline{occurred}_i(recv(j,\mu))$ (from lines (4), (7), by sem. of $correct_i$, $\wedge$)
9. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models \overline{occurred}_j(send(i,\mu))$ (from lines (4), (8), by def. of $filter^B_\varepsilon$, sem. of $correct_j$ and $\wedge$)
10. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models B_j faulty_j$ (from lines (5), (9), by Definition 3.1)
11. $(\mathscr{I},\widetilde{r},\widetilde{t}) \models faulty_j$ (from lines (4), (10), by sem. of B_j, $\wedge$, $\to$, reflexivity of $\sim_j$)
12. contradiction! (from lines (4), (11), sem. of $\wedge$)

□

Lemma 4.8. For Algorithm 1 called with parameters (χ,h_i,i,f), where $\chi \in \mathscr{E}^{B_f}$, interpreted system $\mathscr{I} = (R^\chi,\pi)$, $r \in R^\chi$, $t \in \mathbb{N}$, $i \in \mathscr{A}$, and $h_i = r_i(t)$, the following invariant holds for the variable F during its iterations:

$$(\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall \ell \in F)\ (r_i(t) = h_i) \ \Rightarrow\ (\mathscr{I},r,t) \models B_i faulty_\ell. \tag{23}$$

Proof. By induction over the size of set F, $l = |F|$.
Ind. Hyp.: $(\forall r \in R^\chi)(\forall t \in \mathbb{N})(\forall \ell \in F)\ (r_i(t) = h_i) \ \Rightarrow\ ((R^\chi,\pi),r,t) \models B_i faulty_\ell$
Base case for $l = |\mathsf{DirObBelFaultyAg}(h_i,i) \cup \mathsf{DirNotifBelFaultyAg}(h_i,i)|$: (by code line 2)
The induction hypothesis follows from Corollary 4.3 and 4.5. If the condition on code line 3 is true, then the statement additionally follows from Corollary 4.7.
Ind. Step: Suppose the induction hypothesis holds for $l = |F|$. The only line at which F is modified in the main loop (starting at line 5) is line 10. From line 8 and 9 in the code we get that there exists a $\Sigma \in DisjSS^{Recv^i_{faulty_\ell}(h_i) \setminus F}$ s.t. $|\Sigma| > f - |F|$ before the execution of line 10. Hence the induction hypothesis still remains satisfied by Theorem 3.18 after line 10 has been executed. □

Lemma 4.10. For $\chi \in \mathscr{E}^{B_f}$, $r \in R^\chi$, $t \in \mathbb{N}$, correct agent $i \in \mathscr{A}$, and the set F returned by Algorithm 1 `BeliefWhoIsFaultyAlgorithm`(χ,h_i,i,f), it holds that $|F| \leq f$.

Proof.
1. $|F| > f$ (by contr. assumption)
2. $F =$ `BeliefWhoIsFaultyAlgorithm`(χ,h_i,i,f) (by contr. assumption)
3. $\chi \in \mathscr{E}^{B_f}$ and $r \in R^\chi$ and $t \in \mathbb{N}$ (by contr. assumption)
4. $(\mathscr{I},r,t) \models correct_i$ (by contr. assumption)
5. $(\forall \ell \in F)\ (\mathscr{I},r,t) \models B_i faulty_\ell$ (from lines (2), (3), by Theorem 4.9)
6. #faulty agents in r is at most f (from line (3), by definition of $\mathscr{E}^{B_f}$)
7. $j \in F$ and $(\mathscr{I},r,t) \not\models faulty_j$ (from lines (1), (6), by sem. of $|\ldots|$, $>$, exist. inst.)
8. $(\mathscr{I},r,t) \models B_i faulty_j$ (from lines (5), (7), by univ. inst.)
9. $(\forall r' \in R^\chi)(\forall t' \in \mathbb{N})\ r(t) \sim_i r'(t') \ \Rightarrow\ (\mathscr{I},r',t') \models correct_i \to faulty_j$ (from line (8), by sem. of B_i)
10. $r(t) \sim_i r(t)$ (from line (3), by reflexivity of $\sim_i$)
11. $(\mathscr{I},r,t) \models faulty_j$ (from lines (4), (9), (10), by sem. of $\Rightarrow$, $\to$ univ. inst.)
12. contradiction! (from lines (7), (11))

□

Theorem 4.11. For agent context $\chi \in \mathscr{E}^{B_f}$, $\mathscr{I} = (R^\chi,\pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, and agent $i \in \mathscr{A}$, if $\mathscr{A}$ is finite and i's history is finite at every point in time, then the call `BeliefWhoIsFaultyAlgorithm` $(\chi,r_i(t),i,f)$ invoking Algorithm 1 terminates after a bounded number of steps.

Proof. The fact that variable F in Algorithm 1 is monotonically increasing follows from the code lines 2, 4 and 10 as these are the only lines that modify F. Hence since the set of agents is finite, the outer loop spanning across lines 5. - 12. cannot run forever, as it is bounded by $|\mathscr{A}|$, since we only add agents from the set $\mathscr{A} \setminus F$ and in the worst case only one new agent is added per iteration. Thus at the latest the loop must terminate, when finally $F = \mathscr{A}$. If during some iteration F doesn't change, the loop terminates early.

The same argument goes for the loop spanning across lines 7. - 11.

Regarding the loop spanning lines 8. - 11., since we assumed that agent i's local history is bounded for every local timestamp by some $b \in \mathbb{N}$, agent i could at the most have received messages about b different agent sequences (message chains). Since, by Definition 3.17, $DisjSS^{Recv^i_{faulty_\ell}(h_i) \setminus F}$ is a subset of the power set of $Recv^i_{faulty_\ell}(h_i) \setminus F$ this loop is thus bounded by 2^b.

This covers all the loops in the algorithm. Since all of them are bounded, so is the algorithm as a whole, as it contains no blocking statements. □

D Additional details for Section 5 (Belief Gain about Occurrences of Haps)

Theorem 5.1. For agent context $\chi \in \mathscr{E}^{B_f}$, interpreted system $\mathscr{I} = (R^\chi, \pi)$, run $r \in R^\chi$, timestamp $t \in \mathbb{N}$, action or event $o \in Haps$, agent $i \in \mathscr{A}$, natural number $k \in \mathbb{N} \setminus \{0\}$ s.t. $k + f \leq n$, and set $F \subseteq \mathscr{A}$, which i believes to be faulty, if

$$\left(\begin{array}{c} \left(\exists \Sigma' \in DisjSS^{Recv^i_{\overline{occurred}(o)}(r_i(t)) \setminus F}\right) |\Sigma'| \;\geq\; k + f - |F| \quad \text{or} \\ \left(\exists \Sigma'' \in DisjSS^{Recv^i_{\bigvee_{G \subseteq \mathscr{A},\, |G| = k} \bigwedge_{j \in G} \Box correct_j \wedge B_j \overline{occurred}(o)}(r_i(t)) \setminus F}\right) |\Sigma''| \;>\; f - |F| \end{array} \right) \tag{25}$$

$$\Rightarrow\; (\mathscr{I}, r, t) \models B_i \bigvee_{\substack{G' \subseteq \mathscr{A},\\ |G'| = k}} \bigwedge_{j \in G'} \Box correct_j \wedge B_j \overline{occurred}(o). \tag{26}$$

Proof. The second line of the disjunction follows immediately from Theorem 3.18.

We prove the first line by contradiction.

1. $\left(\exists \Sigma' \in DisjSS^{Recv^i_{\overline{occurred}(o)}(r_i(t)) \setminus F}\right) |\Sigma'| \;\geq\; k + f - |F|$ (by contr. assumption)
2. $(\mathscr{I}, r, t) \not\models B_i \bigvee_{\substack{G' \subseteq \mathscr{A},\\ |G'| = k}} \bigwedge_{j \in G'} \Box correct_j \wedge B_j \overline{occurred}(o)$ (by contr. assumption)
3. $\widetilde{r} \in R^\chi$ and $\widetilde{t} \in \mathbb{N}$ and $r(t) \sim_i \widetilde{r}(\widetilde{t})$ and $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \models correct_i \wedge \bigwedge_{\substack{G' \subseteq \mathscr{A},\\ |G'| = k}} \neg \bigwedge_{j \in G'} \Box correct_j \wedge$ $B_j \overline{occurred}(o)$ (from line (2), by sem. of $\not\models$, $\bigvee$, $\bigwedge$, $\neg$, $\neg B_i$, exist. inst.)
4. $|F| \leq f$ (from line (3), by Lemma 4.10)
5. $(\forall m \in F)\ (\mathscr{I}, r, t) \models B_i faulty_m$ (by assumption)
6. $(\forall m \in F)\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models faulty_m$ (from lines (3), (5), by sem. of B_i and univ. inst.)
7. $\Sigma \in DisjSS^{Recv^i_{\overline{occurred}(o)}(\widetilde{r}_i(\widetilde{t})) \setminus F}$ and $|\Sigma| \;\geq\; k + f - |F|$ (from lines (1), (3), by sem. of $\sim_i$ and exist. inst.)
8. $f + k \leq n$ (by assumption)
9. $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \bigvee_{\substack{G' \subseteq \mathscr{A},\\ |G'| = k}} \bigwedge_{j \in G'} \Box correct_j$ (from line (8), by at most f faulty agents in runs of $\mathscr{E}^{B_f}$)

10. $\Sigma \setminus F = \Sigma$ and $\overline{\Sigma} \subseteq \Sigma$ and $|\overline{\Sigma}| \geq k$ and $(\forall \overline{\sigma} \in \overline{\Sigma})(\forall j \in \{1, \dots, |\overline{\sigma}|\})\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \Box correct_{\pi_j \overline{\sigma}}$ (from lines (3), (4), (6), (7), (9), by at most f faulty agents in runs of $\mathscr{E}^{B_f}$, sem. of $\leq$, $\geq$, $\bigvee$, $\bigwedge$, exist. inst., using a pigeonhole argument)
11. $(\forall \overline{\sigma} \in \overline{\Sigma})(\exists \mu \in Msgs^{\pi_1 \overline{\sigma} \to i}_{\overline{H}_{\overline{\sigma}} \overline{occurred}(o)})\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \overline{occurred}_i(recv(\pi_1 \overline{\sigma}, \mu))$ (from lines (7), (10), by sem. of $\wedge$, $correct_i$, $\overline{occurred}_i(recv(\pi_1 \overline{\sigma}, \mu))$ and Definition 3.14 and 3.16)
12. $(\forall \overline{\sigma} \in \overline{\Sigma})(\exists \mu \in Msgs^{\pi_1 \overline{\sigma} \to i}_{\overline{H}_{\overline{\sigma}} \overline{occurred}(o)})\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \overline{occurred}_{\pi_1 \overline{\sigma}}(send(i, \mu))$ (from lines (10), (11), by sem. of $\overline{occurred}_{\pi_1 \overline{\sigma}}(send(i, \mu))$, Definition of $filter_{\varepsilon}^{B_f}$)
13. $(\forall \overline{\sigma} \in \overline{\Sigma})\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models B_{\pi_1 \overline{\sigma}} \overline{H}_{\pi_2 \overline{\sigma} \circ \dots \circ \pi_{|\sigma|} \overline{\sigma}} \overline{occurred}(o)$ (from line (12), by Definition 3.1, Lemma 3.3, Corollary 3.15, sem. of $\overline{occurred}_j(send(i, \mu))$)
14. let $\widetilde{G} := \{ j \in \mathscr{A} \mid \overline{\sigma} \in \overline{\Sigma} \quad \text{and} \quad j = \pi_{|\overline{\sigma}|} \overline{\sigma} \}$ (exist. inst.)
15. $|\widetilde{G}| \geq k$ (from lines (7), (10), by sem. of $\subseteq$, Definition 3.17)
16. let $\widetilde{G}' \subseteq \widetilde{G}$ and $|\widetilde{G}'| = k$ (from line (15), by sem. of $\subseteq$, exist. inst.)
17. $(\forall j \in \widetilde{G}')\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \Box correct_j$ (from lines (10), (14), (16))
18. $(\forall j \in \widetilde{G}')\ (\mathscr{I}, \widetilde{r}, \widetilde{t}) \models B_j \overline{occurred}(o)$ (by lines (13), (17), by sem. of H, $\Box$, $correct_j$, $\rightarrow$, reflexivity of $\sim_j$)
19. $(\mathscr{I}, \widetilde{r}, \widetilde{t}) \models \bigwedge_{j \in \widetilde{G}'} \Box correct_j \wedge B_j \overline{occurred}(o)$ (from line (13), by sem. of $\forall$)
20. contradiction! (from lines (3), (16), by sem. of $\wedge$)

□

Mining for Unknown Unknowns

Bernard Sinclair-Desgagné
SKEMA Business School
Sophia Antipolis, France
GREDEG, Université Côte d'Azur[*]
Sophia Antipolis, France
bsd@skema.edu

Unknown unknowns are future relevant contingencies that lack an ex ante description. While there are numerous retrospective accounts showing that significant gains or losses might have been achieved or avoided had such contingencies been previously uncovered, getting hold of unknown unknowns still remains elusive, both in practice and conceptually. Using Formal Concept Analysis (FCA) - a subfield of lattice theory which is increasingly applied for mining and organizing data - this paper introduces a simple framework to systematically think out of the box and direct the search for unknown unknowns.

> There are only two kinds of campaign plans, good ones and bad ones.
> The good ones almost always fail through unforeseen circumstances
> that often make the bad ones succeed.
>
> - Napoleon Bonaparte -

1 Introduction

As the recent Covid-19 pandemic reminded us, life is filled with unknown unknowns – i.e. contingencies one cannot be aware of ex ante, much less fit into standard risk analysis. In addition to a wealth of examples coming from history and politics, unknown unknowns are now well-documented, and their importance is acknowledged, in many areas of economics and management such as public policy [23], business strategy [5, 11], entrepreneurship [12], contracts and the theory of the firm [40], and security [32].

To be sure, getting hold of such contingencies might allow to achieve significant payoffs or avoid major losses. Substantial research efforts have thus been expended, and notable advances been made, in this direction. To get a rigorous conceptual grasp at the notion of unknown unknowns, one may now draw, notably, from the literatures on Knightian uncertainty (e.g., [4]), undescribable events (e.g., [24]), unforeseen contingencies (e.g., [7, 20]), unawareness (e.g., [35, 34]), and surprises [39, 26]. Yet, for someone who would primarily want to uncover ahead of time the concrete unknown unknowns she might be facing, the task would remain elusive.

This paper will now seek to meet this demand. Similar endeavors have already been tried, and results obtained, in areas where unknown unknowns occur frequently: like C-K Theory [17], TRIZ [19], and

[*] I am grateful to the participants at several seminars for constructive questions and remarks. I also owe special thanks to Thierry Bréchet, Thierry Burger-Helchem, Michel Desmarais, Soumittra Dutta, Ludovic Dibiaggio, Armand Hatchuel, Emile Grenier-Robillard, Henry Tulkens, and Benoit Weil for their encouragements, comments or suggestions. Three diligent and competent reviewers contributed to substantially clarify and improve the initial draft of this paper.

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 507–517, doi:10.4204/EPTCS.379.38

creativity support systems [13, 44] in innovation management; knowledge spaces [10] in learning; and elicitation methods [21, 31, 38] in engineering. As it turns out, Formal Concept Analysis (FCA), which I will be using here, provides an appropriate language, and especially structure, to build a framework which is both rigorous (it is grounded in lattice theory) and operational (its implementation requires only spreadsheets).

The suggested scheme is sketched in the following Section 2. It is next developed rigorously and with more generality in Section 3. A fourth section contains concluding remarks.

2 An informal account

Consider a 3x3 matrix with horizontal coordinates A, B, C, and vertical coordinates α β, γ. For concreteness, the former might refer to different objects, items or events and the latter to various attributes, characteristics, properties or features. Table 1 shows the features respectively held by each specific item: object A, for instance, possesses attributes α, β.

Attributes Objects	α	β	γ
A	×	×	
B			×
C		×	×

Table 1: The existing context

In Formal Concept Analysis (FCA), such a matrix showing relationships between 'objects' (items, events, etc.) and 'attributes' (properties, features, etc.) is called a context.

In practice, FCA users would of course face much more complicated types of contexts, with tables comprising dozens of rows and columns, mitigated relationships between objects and attributes, and (what is crucial for decision-making) value-weighted properties. But this simple example will suffice to convey our main points.

The upshot of a discovery, experiment or invention would be the expanded matrix displayed in Table 2. Two new items – D and E – and an extra characteristic δ were found. The initial objects A and B now bear attributes δ and α respectively, while D possesses properties α and δ, E exhibits features β and γ. This matrix forms a context as well.

Attributes Objects	α	β	γ	δ
A	×	×		×
B	×		×	
C		×	×	
D	×			×
E		×	×	

Table 2: The new context

Back in time, the incremental component of Table 2 – i.e. the rows D, E, column δ and the small x's – might have been impossible to describe, much less to anticipate even as random outcomes. They were, so to speak, *unknown unknowns*. Altogether, they make the context displayed in Table 3.

Attributes Objects	α	β	γ	δ
A				×
B	×			
C				
D	×			×
E		×	×	

Table 3: The discovery context

Let us now see how someone could have a grasp at Table 3 using known knowns **only**, these known knowns being the data available from Table 1.

In FCA, the primary mode of organizing the data of a context is through the use of 'concepts'. A concept is defined as a list of objects and attributes such that the mentioned objects are precisely the ones that share the listed attributes, and the mentioned attributes are precisely the ones shared by all the listed objects. Examples of concepts in Table 3 are the objects B, D with their common attribute α, event E with features β, γ, items A, D with the shared property δ, and object D with attributes α, δ.

FCA calls an incompletely specified concept, i.e. a list that misses some object and attributes, a preconcept. The list (B;α) is a preconcept of the concept (B,D;α), for instance. Since its object B and attribute α could already be seen in the existing context of Table 1, I shall refer to such specific preconcept as a *seed*.

Now, the relationship between B and α is captured in Table 4, which is actually the negative picture of Table 1. This table constitutes a context as well, and *it is made only of data from Table 1*. Its concepts – which might be called 'anti-concepts', since they are the counterpart of the existing initial concepts – include (B,C;α), (A;γ) and (B;α,β).

Attributes Objects	α	β	γ
A			Z
B	Z	Z	
C	Z		

Table 4: The negative existing context

This paper's main result is that *a seed* - like (B;α) - *will always be the pre-concept of some anti-concept* – namely, here, (B,C;α) or (B;α,β). This fact has at least three ramifications.

First, although one cannot say anything about which objects or which attributes will be discovered, the structure of the existing context bears some implications for the structure among discovered objects and attributes.[1] This opens the door for establishing a systematic procedure to get some grasp at, and eventually uncover, unknown unknowns:

[1] I am grateful to an anonymous referee for this observation.

(i) Build the negative picture of the existing context;
(ii) Examine the preconcepts of each anticoncept;
(iii) If a seed is found, dig into it to uncover some concepts in the unknown unknown.

Second, the latter procedure might be seen as an instance of abduction, a mode of reasoning associated with creativity and the generation of ideas [22, 41]. Unlike deduction, which draws the logical ramifications of previously given assertions, or induction, which infers general laws from the observation of recurrent facts, abduction looks for the best justification (which is here a concept of the discovery context) after hitting a singular event (a seed).

Third, as the upcoming section will show, it allows for the deployment of a potentially powerful tool for data exploration and exploitation - namely, Galois connections.

3 Formal Developments

Let's now present the mathematics which underlie the scheme outlined above. Subsection 3.1 revisits the basics of Formal Concept Analysis (FCA). Subsection 3.2 next introduces the notion of 'revelation mappings'. Subsection 3.3, finally, develops the systematic procedure for thinking out of the box. The treatment is meant to be self-contained. Only set-theoretic arguments are used throughout.

3.1 Basic FCA notions

A *formal context* is referred to as a triplet $K = (G, M; R)$, where G is a set of *objects*, M a set of *attributes* these objects may have, and R is a *relation* between G and M, i.e. a subset of the Cartesian product $G \times M$ with the interpretation that $(g, m) \in R$, or gRm, if object g has attribute m.

Denote $\wp(G)$ and $\wp(M)$ the respective power sets (or sets of all subsets) of G and M. Set inclusion $\subseteq$ provides a partial order on the elements of these sets.[2] The following set-to-set functions I_R and E_R defined as

$$\begin{aligned} \text{for S} &\subseteq G,\ I_R(S) = \{m \in M : gRm \ \text{ for all } \ \text{g} \in S\} \\ \text{for T} &\subseteq M,\ E_R(T) = \{g \in G : gRm \ \text{ for all } \ \text{m} \in T\} \end{aligned}$$

are called the *Birkhoff Operators* for G and M respectively. For a set of objects S, $I_R(S)$ - the *intent* of S - gives all the attributes in T which these objects have in common. For a given set of attributes T, $E_R(T)$ - the *extent* of T - gives all the objects in S that share these attributes. In the context displayed in Table 1, $I_R(\text{A,C}) = \{\beta\}$ and $E_R(\alpha, \beta) = \{\text{A}\}$.

A well-known property of the Birkhoff Operators is that of *duality*: knowing $I_R(\cdot)$ completely determines $E_R(\cdot)$, and vice-versa, specifying $E_R(\cdot)$ also defines $I_R(\cdot)$.

A *formal concept* in the context $K = (G, M; R)$ is now a pair of sets $(Q; V)$, with $Q \subseteq G$ and $V \subseteq M$, such that $I_R(Q) = V$ and $E_R(V) = Q$. The extent of a concept $(Q; V)$ is thus Q, while its intent is V.[3]

[2] A set Q is a *partially ordered set* (or *poset*) if there is a relation $\leq$ on Q (called a *partial order*) such that: (i) for $q \in Q$, $q \leq q$ (reflexivity property); (ii) for $q_1, q_2 \in Q$, $q_1 \leq q_2$ and $q_2 \leq q_1$ implies $q_1 = q_2$ (antisymmetry); for $q_1, q_2, q_3 \in Q$, $q_1 \leq q_2$ and $q_2 \leq q_3$ implies $q_1 \leq q_3$ (transitivity).

[3] The way FCA defines a formal concept agrees with the International Standard Organization's ISO 704 definition: "In a concept, one distinguishes its 'intension' and 'extension'. The intension of a concept comprises all attributes thought with it, the extension comprises all objects for which the concept can be predicated. In general, the richer the intension of a concept is, the lesser is its extension, and vice versa."

A *preconcept* in K, finally, is a pair $(P;U)$, with $P \subseteq G$ and $U \subseteq M$, such that $P \subseteq E_R(U)$ or, equivalently, $U \subseteq I_R(P)$. Preconcepts can be ordered as follows [8]: $(P;U) \sqsubseteq (P';U')$, meaning that $(P;U)$ is *less extensive than* $(P';U')$, if $P \subseteq P'$ and $U \subseteq U'$.

3.2 Revelation mappings

From now on, $K = (G,M;R)$ will denote the existing context and $K^+ = (G^+,M^+;R^+)$ the new context after the previous unknown unknowns have been revealed.[4]

Let's then call *revelation mappings* the functions $\Phi : \wp(G^+) \to \wp(M^+)$, $\Psi : \wp(M^+) \to \wp(G^+)$ such that[5]

$$\begin{aligned} \text{for } \mathrm{S}^+ &\subseteq G^+,\ \Phi(S^+) = I_{R^+}(S^+) \setminus \bigcup_{g \in S^+} I_R(g) \\ \text{for } \mathrm{T}^+ &\subseteq M^+,\ \Psi(T^+) = E_{R^+}(T^+) \setminus \bigcup_{m \in T^+} E_R(m) \end{aligned}$$

If one takes a set $S \subseteq G$ of objects from the existing context K, $\Phi(S)$ delivers the set of attributes (old or new) in M^+ which are newly associated with these objects. In Table 2, for instance, $\Phi(\mathrm{B,E}) = \varnothing$ and $\Phi(\mathrm{A}) = \{\delta\}$. Similarly, for a subset of initial attributes $T \subseteq M$, $\Psi(T)$ gives all (and only) the initial or new objects that now possess these attributes. For example, $\Psi(\gamma,\delta) = \varnothing$ but $\Psi(\beta,\gamma) = \{\mathrm{E}\}$.

As for the Birkhoff operators, there is a *duality property* between $\Phi(\cdot)$ and $\Psi(\cdot)$: each one uniquely characterizes the other. These mappings also hold additional features which are spelled out in the upcoming propositions.

First, say that a function $\pi : X \to Y$ between two sets X and Y, partially ordered by $\leq$ and $\preceq$ respectively, is *antitone* (or order-reversing) if, for $p_1, p_2 \in X$, $p_1 \leq p_2$ implies $\pi(p_2) \preceq \pi(p_1)$. A first statement is now at hand.

PROPOSITION 1: The revelation mappings Φ and Ψ are antitone.

PROOF:

First, consider Φ. Take two sets S_1^+, $S_2^+ \in \wp(G^+)$ such that $S_1^+ \subseteq S_2^+$; we must show that $\Phi(S_2^+) \subseteq \Phi(S_1^+)$. If $m \in \Phi(S_2^+)$, then $m \in I_{R^+}(S_2^+)$ so gR^+m for all $g \in S_2^+$. Since $S_1^+ \subseteq S_2^+$, we have that gR^+m for all $g \in S_1^+$, hence $m \in I_{R^+}(S_1^+)$. Now, if $m \notin M$, $m \notin I_R(g)$ for any $g \in S_1^+$; it follows that $m \in I_{R^+}(S_1^+) \setminus \bigcup_{g \in S_1^+} I_R(g) = \Phi(S_1^+)$. Suppose, alternatively, that $m \in M$. Since $m \in \Phi(S_2^+)$, it must be the case that $not(gRm)$ for all $g \in S_2^+$, hence $not(gRm)$ as well for all $g \in S_1^+$ since $S_1^+ \subseteq S_2^+$; it follows again that $m \in \Phi(S_1^+)$. This shows that $\Phi(S_2^+) \subseteq \Phi(S_1^+)$. The same line of reasoning works for Ψ (as can be expected from duality). ■

This property of revelation mappings means that, the more objects or attributes one starts with, the more demanding it is to find new relationships that fit them all. This intuitive result is also instrumental in deriving other important characteristics of revelation mappings.

A key notion to introduce at this point is that of a *Galois connection*.[6] Let X and Y be two sets

[4] Power sets, Birkhoff Operators, formal concepts, and preconcepts are similarly defined on their context of reference, be it K, K^+, or any other context.

[5] Let's agree that $I_R(g) = \varnothing$ when $g \notin G$, and $E_R(m) = \varnothing$ when $m \notin M$.

[6] Since at least Ore (1944)'s seminal article [25], Galois connections have been increasingly employed throughout mathematics and computer science. To go beyond the very short primer offered in this paper, the reader may look at [6, 10, 15], and some of their common references.

partially ordered by $\leq$ and $\preceq$ respectively. A (antitone) *Galois connection* (π, θ) on X and Y is a pair of functions $\pi : X \to Y$ and $\theta : Y \to X$ such that the following equivalent properties are satisfied.
(i) For each $p \in X$, $p \leq \theta\pi(p)$ and for each $q \in Y$, $q \preceq \pi\theta(q)$.
(ii) For $p \in X$ and $q \in Y$, $p \leq \theta(q)$ if and only if $q \preceq \pi(p)$.

It is well-known that the Birkhoff operators (I_R, E_R), (I_{R^+}, E_{R^+}) are antitone Galois connections on, respectively, the power sets $\wp(G)$, $\wp(M)$ and $\wp(G^+)$, $\wp(M^+)$ ordered by set inclusion (see, e.g., [15], p. 13-14). In this case, property (i) means that the attributes common to a given set of objects might be shared by more objects, while the objects that share a given set of attributes might have more attributes in common. Property (ii), on the other hand, says that some objects are among those sharing a given set of attributes if and only if these attributes are among those common to these objects.

As it turns out, the pair of revelation mappings (Φ, Ψ) forms a Galois connection.

PROPOSITION 2: The pair of revelation mappings (Φ, Ψ) is a Galois connection on the power sets $\wp(G^+)$ and $\wp(M^+)$ partially ordered by inclusion.

PROOF:
To see this, take two sets $S^+ \in \wp(G^+)$ and $T^+ \in \wp(M^+)$, and notice that

$$\begin{array}{rcl} S^+ & \subseteq & \Psi(T^+) \\ \text{if and only if } \forall g & \in & S^+, \forall m \in T^+\text{: } gR^+m \text{ and } not(gRm) \\ \text{if and only if } \forall m & \in & T^+, \forall g \in S^+\text{: } gR^+m \text{ and } not(gRm) \\ \text{if and only if } T^+ & \subseteq & \Phi(S^+) \end{array}$$

■

Proposition 2 underlies a central result. Like any Galois connection ([15], p. 14), (Φ, Ψ) establishes a relation, noted $R^+_{(\Phi,\Psi)}$, between the set of objects G^+ and the set of attributes M^+. This relation is defined as

$$\begin{array}{rcl} R^+_{(\Phi,\Psi)} & = & \big\{(g,m) \in G^+ \times M^+ \mid g \in \Psi(m)\big\} \\ & = & \big\{(g,m) \in G^+ \times M^+ \mid m \in \Phi(g)\big\} \end{array}$$

We can show that $R^+_{(\Phi,\Psi)}$ coincides with $R^+ \setminus R$, the set of all new relationships.

PROPOSITION 3: $R^+_{(\Phi,\Psi)} = R^+ \setminus R$.

PROOF: Observe that $(g,m) \in R^+_{(\Phi,\Psi)}$ if and only if gR^+m and $not(gRm)$, if and only if $(g,m) \in R^+ \setminus R$.

■

3.3 Thinking out of the box

From now on, let $R^+_{(\Phi,\Psi)} = R^+ \setminus R$ be referred to as R^*. The latter relation defines another formal context, the *discovery* context noted $K^* = (G^+, M^+; R^*)$, which is the context of the unknown unknowns. Can K^* be inferred from K, at least partly? We will now see that the answer actually errs on the yes side.

The ordered pair $(X;Y)$ with $X \neq \varnothing$, $Y \neq \varnothing$ is called a *seed* in K for K^* if it is a preconcept in K^* while $X \subseteq G$ and $Y \subseteq M$. As the next statement confirms, the existence of a seed is guaranteed when the existing context harbors at least one new relationship between the original objects and attributes.

PROPOSITION 4: If $R^* \cap (G \times M) \neq \varnothing$, then there is at least one seed in K for K^*.

PROOF:

The assumption implies that there is at least one concept $(Q;V)$ in K^* such that $Q \cap G \neq \varnothing$ and $V \cap M \neq \varnothing$. Since $Q \cap G \subseteq Q = I_{R^*}(V) \subseteq I_{R^*}(V \cap M)$ and $V \cap M \subseteq V = E_{R^*}(Q) \subseteq E_{R^*}(Q \cap G)$, the pair $(Q \cap G; V \cap M)$ is a preconcept in K^*. ■

As suggested in Section 2, looking for seeds might be a reasonable first step to uncover unknown unknowns. The major reason is that, as we will now demonstrate, *it is possible to characterize the location of seeds.*

First, according to the following proposition, a seed must combine objects and attributes which are a priori unrelated.

PROPOSITION 5: No preconcept (a fortiori concept) in the existing context K can be a seed for the discovery context K^*.

PROOF:

Let $(P;U)$ be a preconcept in K. By definition, $\Phi(P) = I_{R^+}(P) \setminus \bigcup_{g \in P} I_R(g)$. But $U \subseteq I_R(P) = \bigcap_{g \in P} I_R(g) \subseteq \bigcup_{g \in P} I_R(g)$. It follows that $U \nsubseteq \Phi(P)$, hence $(P;U)$ is not a preconcept in K^*.

This result tells us something about how not to look for novelties. A corollary is that a seed in K for K^* must be a pair $(P;U)$, with $P \subseteq G$ and $U \subseteq M$, such that $P \cap (\bigcup_{m \in U} E_R(m)) = \varnothing$ and $U \cap (\bigcup_{g \in P} I_R(g)) = \varnothing$. This suggests working with the negative of the existing context K, noted $\overline{K} = (G, M; \overline{R})$, where the relation $\overline{R} = G \times M \setminus R$ refers to the *reverse relation $g\overline{R}m$* which holds when object *g does not* have attribute m. The next (key, and somewhat surprising) proposition shows that $\overline{K}$ - which can be obtained using **only** the initial data - is the appropriate 'outbox' in which mining for unknown unknowns might begin.

PROPOSITION 6: A seed is a preconcept of the negative existing context $\overline{K}$.

PROOF:

Let $(P;U)$ be a seed for K^*. Then $U \subseteq \Phi(P) \cap M = M \cap I_{R^+}(P) \setminus \bigcup_{g \in P} I_R(g) \subseteq M \setminus \bigcup_{g \in P} I_R(g)$
$= \bigcap_{g \in P} (I_R(g))^c = I_{\overline{R}}(P)$. ■

Seeds for the discovery context K^* - which comprises a priori unknown relationships between objects in G and attributes in M - thus happen to point, not only at concepts in K^*, but also at the concepts of the negative existing context $\overline{K}$. This suggests the procedure already outlined in Section 2:

- Take the negative context $\overline{K}$ of K;
- Consider a concept in $\overline{K}$ (i.e. an anti-concept);
- Examine the latter's preconcepts;
- If one of these preconcepts brings out a new relationship between its objects and attributes, then a seed has been found which anticipates some concepts in the discovery context K^*.

Whether this scheme can be fruitful in practice remains to be seen. One hurdle could be computational complexity (see the concluding remarks).

Interestingly, however, Propositions 5 and 6 suggest that concepts in the negative existing context $\overline{K}$ can be constructed using the mappings $\widetilde{\Phi} : \wp(G) \rightarrow \wp(M)$ and $\widetilde{\Psi} : \wp(M) \rightarrow \wp(G)$ defined as

$$\begin{aligned} \text{for S} &\subseteq G,\ \widetilde{\Phi}(S) = M \setminus \bigcup_{g \in S} I_R(g) \\ \text{for T} &\subseteq M,\ \widetilde{\Psi}(T) = G \setminus \bigcup_{m \in T} E_R(m) \end{aligned}$$

respectively. Comparing the latter expressions with the ones corresponding to the above revelation mappings, the functions $\widetilde{\Phi}$ and $\widetilde{\Psi}$ can be seen as approximations for Φ and Ψ. Whether closer approximations (in a sense to be made precise) can be found, which would then provide a better grasp at unknown unknowns, would be a valuable research topic.

4 Concluding remarks

This paper submitted a new framework and approach to handle unknown unknowns. The scheme has rigorous foundations in lattice theory. It looks widely applicable, furthermore, since it can incorporate various kinds of data – quantitative and qualitative, objective and subjective, financial and non-financial. And it seems to be user-friendly, boiling down to using only spreadsheets.

At this stage, in addition to the extensions suggested at the end of the previous section, other ones could be the following:

First, on a technical note, listing all the concepts of a formal context is generally burdensome.[7] Yet, the search for seeds requires this exercise. Research and development on how to identify concepts in a given context is very much ongoing. Several algorithms and softwares already exist: many (mentioned in [37], for instance) are subject to a patent but others – GALICIA and JALABA, for example – can be freely downloaded. Two promising trends are to take full advantage of negative information (i.e. the information contained in the negative existing context $\overline{K}$), as in [33] or [27], and to assign weights to attributes, as in [3].

Second, the above derivation made minimal assumptions about the use of a priori knowledge, ignoring issues of landscape and timing, and forbidding the use of probabilities. In practice, however, one might be able to tap on probabilistic beliefs based on science, predictive models or sound experience, in order to figure out the plausibility of new relationships between objects and attributes. This endeavor will enhance the search for seeds, hence the prospecting for unknown-unknowns.

References

[1] Pauline Barrieu and Bernard Sinclair-Desgagné. "On Precautionary Policies". In: *Management Science* 52.8 (2006), pp. 1145–1154. ISSN: 0025-1909. DOI: 10.1287/MNSC.1060.0527.

[2] Radim Bělohlávek. *Introduction to formal concept analysis*. Tech. rep. Olomouc: Departement of Computer Science, Palacký University, 2008, p. 47.

[3] Radim Bělohlávek and Juraj Macko. "Selecting important concepts using weights". In: *Proceedings of the 9th International Conference on Formal Concept Analysis*. Berlin, Heidelberg: Springer, 2011, pp. 65–80. DOI: 10.1007/978-3-642-20514-9_7.

[7] An upper bound on the number of concepts in the context K = (G,M;R) is $\frac{3}{2}2^{\sqrt{|R|+1}} - 1$. See [15], p. 94.

[4] Truman F. Bewley. "Knightian uncertainty". In: *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures 1983–1997*. Ed. by Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien. Vol. 29. Cambridge University Press: Econometric Society Monographs, 1998, pp. 71–81. ISBN: 978-0521632225.

[5] Kevin Bryan, Michael Ryall, and Burkhard C. Schipper. "Value-Capture in the Face of Known and Unknown Unknowns". In: *Strategy Science* 7.3 (2021), pp. 157–189. DOI: 10.1287/stsc.2021.0126.

[6] Brian A. Davey and Hilary A. Priestley. *Introduction to lattices and order*. Cambridge University Press, 2002. ISBN: 0521784514.

[7] Eddie Dekel, Barton L. Lipman, and Aldo Rustichini. "Recent developments in modeling unforeseen contingencies". In: *European Economic Review* 42.3-5 (May 1998), pp. 523–542. ISSN: 0014-2921. DOI: 10.1016/S0014-2921(97)00114-1.

[8] Jeffrey T. Denniston, Austin Melton, and Stephen E. Rodabaugh. "Formal Contexts, Formal Concept Analysis, and Galois Connections". In: *Proceedings Festschrift for Dave Schmidt*. Vol. 129. Sept. 2013, pp. 105–120. DOI: 10.4204/EPTCS.129.8.

[9] Edwin Diday and Richard Emilion. "Maximal and stochastic Galois lattices". In: *Discrete Applied Mathematics* 127.2 (Apr. 2003), pp. 271–284. ISSN: 0166-218X. DOI: 10.1016/S0166-218X(02)00210-X.

[10] Jean-Paul Doignon and Jean-Claude Falmagne. *Knowledge spaces*. Berlin, Heidelberg: Springer, 1999, p. 334. ISBN: 978-3540645016.

[11] Timo Ehrig and Nicolai J. Foss. "Unknown Unknowns and the Treatment of Firm-Level Adaptation in Strategic Management Research". In: *Strategic Management Review* 3.1 (2022), pp. 1–24. ISSN: 2688-2612. DOI: 10.1561/111.00000035.

[12] Timo Ehrig and Nicolai J. Foss. "Why we need normative theories of entrepreneurial learning that go beyond Bayesianism". In: *Journal of Business Venturing Insights* 18 (Nov. 2022), pp. 1–5. ISSN: 2352-6734. DOI: 10.1016/J.JBVI.2022.E00335.

[13] A. Gabriel et al. "Creativity support systems: A systematic mapping study". In: *Thinking Skills and Creativity* 21 (Sept. 2016), pp. 109–122. ISSN: 1871-1871. DOI: 10.1016/J.TSC.2016.05.009.

[14] Bernhard Ganter and Sergei O. Kuznetsov. "Formalizing hypotheses with concepts". In: *Conceptual Structures: Logical, Linguistic and Computational Issues. ICCS 2000*. Ed. by Bernhard Ganter and G.W. Mineau. Vol. 1867. Berlin, Heidelberg: Springer, 2000, pp. 342–356. ISBN: 354067859X. DOI: 10.1007/10722280_24.

[15] Bernhard Ganter, Rudolf Wille, and C. Franzke. *Formal Concept Analysis: Mathematical Foundations*. Berlin, Heidelberg: Springer, 1999. ISBN: 978-3540627715.

[16] Robert Godin, Rokia Missaoui, and Hassan Alaoui. "Incremental concept formation algorithms based on Galois (concept) lattices". In: *Computational Intelligence* 11.2 (1995), pp. 246–267. DOI: 10.1111/j.1467-8640.1995.tb00031.x.

[17] Armand Hatchuel and Benoît Weil. "C-K design theory: An advanced formulation". In: *Research in Engineering Design* 19 (2009), pp. 181–192. DOI: 10.1007/s00163-008-0043-4.

[18] D. I. Ignatov. "Introduction to Formal Concept Analysis and its applications in Information Retrieval and related fields". In: *Information Retrieval*. Ed. by P. Braslavski et al. Springer, 2015. DOI: 10.1007/978-3-319-25485-2_3.

[19] Imoh M. Ilevbare, David Probert, and Robert Phaal. "A review of TRIZ, and its benefits and challenges in practice". In: *Technovation* 33.2-3 (Feb. 2013), pp. 30–37. ISSN: 0166-4972. DOI: 10.1016/J.TECHNOVATION.2012.11.003.

[20] Asen Kochov. "A behavioral definition of unforeseen contingencies". In: *Journal of Economic Theory* 175 (May 2018), pp. 265–290. ISSN: 0022-0531. DOI: 10.1016/J.JET.2018.01.018.

[21] H. Lakkaraju et al. "Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017. DOI: 10.1609/aaai.v31i1.10821.

[22] Ramzi Mabsout. "Abduction and economics: the contributions of Charles Peirce and Herbert Simon". In: *Journal of Economic Methodology* 22.4 (Aug. 2015), pp. 491–516. ISSN: 1350-178X. DOI: 10.1080/1350178X.2015.1024876.

[23] Bernardo Mueller. "Why public policies fail: Policymaking under complexity". In: *EconomiA* 21.2 (May 2020), pp. 311–323. ISSN: 1517-7580. DOI: 10.1016/J.ECON.2019.11.002.

[24] Nabil I. Al-Najjar, Luca Anderlini, and Leonardo Felli. "Undescribable Events". In: *The Review of Economic Studies* 73.4 (Oct. 2006), pp. 849–868. ISSN: 0034-6527. DOI: 10.1111/J.1467-937X.2006.00399.X.

[25] Oystein Ore. "Galois Connexions". In: *Transactions of the American Mathematical Society* 55.3 (1944), pp. 293–513. DOI: 10.2307/1990305.

[26] Günther Palm. *Novelty, Information and Surprise*. Berlin, Heidelberg: Springer, 2012, p. 293. ISBN: 9783662658758.

[27] Francisco Pérez-Gámez et al. "A new kind of implication to reason with Unkown information". In: *Formal Concept Analysis. ICFCA 2021*. Springer, 2021. DOI: 10.1007/978-3-030-77867-5_5.

[28] Jonas Poelmans et al. "Formal concept analysis in knowledge processing: A survey on applications". In: *Expert Systems with Applications* 40.16 (Nov. 2013), pp. 6538–6560. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2013.05.009.

[29] Jonas Poelmans et al. "Formal Concept Analysis in knowledge processing: A survey on models and techniques". In: *Expert Systems with Applications* 40.16 (Nov. 2013), pp. 6601–6623. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2013.05.007.

[30] Jonas Poelmans et al. "Fuzzy and rough formal concept analysis: a survey". In: *International Journal of General Systems* 43.2 (2014), pp. 105–134. DOI: 10.1080/03081079.2013.862377. URL: https://doi.org/10.1080/03081079.2013.862377.

[31] Ranga V. Ramasesh and Tyson R. Browning. "A conceptual framework for tackling knowable unknown unknowns in project management". In: *Journal of Operations Management* 32.4 (May 2014), pp. 190–204. ISSN: 0272-6963. DOI: 10.1016/J.JOM.2014.03.003.

[32] Awais Rashid et al. "Discovering unknown known security requirements". In: *Proceedings of the 38th International Conference on Software Engineering* (May 2016), pp. 866–876. ISSN: 02705257. DOI: 10.1145/2884781.2884785.

[33] J. M. Rodriguez-Jimenez et al. "Concept lattices with negative information: A characterization theorem". In: *Information Sciences* 369 (Nov. 2016), pp. 51–62. ISSN: 0020-0255. DOI: 10.1016/J.INS.2016.06.015.

[34] Burkhard C. Schipper. "Awareness". In: *Handbook of Logics for Knowledge and Belief*. Ed. by Hans van Ditmarsch et al. London, UK: College Publications, 2014. DOI: 10.48550/arXiv.1503.00806.

[35] Burkhard C. Schipper. "Unawareness—A gentle introduction to both the literature and the special issue". In: *Mathematical Social Sciences* 70 (July 2014), pp. 1–9. ISSN: 0165-4896. DOI: `10.1016/J.MATHSOCSCI.2014.03.002`.

[36] Bernard Sinclair-Desgagné. "Measuring innovation and innovativeness: a data-mining approach". In: *Quality and Quantity* 56.4 (Aug. 2022), pp. 2415–2434. ISSN: 15737845. DOI: `10.1007/S11135-021-01231-6/METRICS`.

[37] Prem Kumar Singh, Cherukuri Aswani Kumar, and Abdullah Gani. "A comprehensive survey on formal concept analysis, its research trends and applications". In: *International Journal of Applied Mathematics and Computer Science* 26.2 (June 2016), pp. 495–516. ISSN: 1641876X. DOI: `10.1515/AMCS-2016-0035`.

[38] Alistair Sutcliffe and Pete Sawyer. "Requirements elicitation: towards the unknown unknowns". In: *2013 21st IEEE International Requirements Engineering Conference, RE 2013 - Proceedings*. 2013, pp. 92–104. ISBN: 9781467357654. DOI: `10.1109/RE.2013.6636709`.

[39] Nassim Nicholas Taleb. *The Black Swan: The impact of the highly improbable*. Random House, 2007.

[40] Jean Tirole. "Incomplete Contracts: Where do We Stand?" In: *Econometrica* 67.4 (July 1999), pp. 741–781. ISSN: 1468-0262. DOI: `10.1111/1468-0262.00052`.

[41] Fernando Tohmé and Ricardo Crespo. "Abduction in economics: A conceptual framework and its model". In: *Synthese* 190 (May 2013), pp. 4215–4237. ISSN: 15730964. DOI: `10.1007/S11229-013-0268-2/METRICS`.

[42] Petko Valtchev, Rokia Missaoui, and Robert Godin. "Formal concept analysis for knowledge discovery and data mining: the new challenges". In: *Proceedings- of the International Conference on Formal Concept Analysis. ICFCA 2004*. Sydney, Australia: Springer, 2004.

[43] Francisco J. Valverde-Albacete et al. "Supporting scientific knowledge discovery with extended, generalized Formal Concept Analysis". In: *Expert Systems with Applications* 44 (Feb. 2016), pp. 198–216. ISSN: 0957-4174. DOI: `10.1016/J.ESWA.2015.09.022`.

[44] Kai Wang and Jeffrey V. Nickerson. "A literature review on individual creativity support systems". In: *Computers in Human Behavior* 74 (Sept. 2017), pp. 139–151. ISSN: 0747-5632. DOI: `10.1016/J.CHB.2017.04.035`.

[45] Rudolf Wille. "Restructuring lattice theory: an approach based on hierarchies of concepts". In: *Ordered Sets*. Ed. by Ivan Ridal. Dordrecht: Springer, 1982. DOI: `10.1007/978-94-009-7798-3_15`.

Aggregating Credences into Beliefs: Agenda Conditions for Impossibility Results (extended abstract)

Minkyung Wang
CONCEPT
Cologne, Germany
Department of Philosophy
University of Cologne
Cologne, Germany
minkyungwang@gmail.com

Chisu Kim
Independent Researcher
Cologne, Germany
tschuessu@gmail.com

Binarizing belief aggregation addresses how to rationally aggregate individual probabilistic beliefs into collective binary beliefs. Similar to the development of judgment aggregation theory, formulating axiomatic requirements, proving impossibility theorems, and identifying exact agenda conditions of impossibility theorems are natural and important research topics in binarizing belief aggregation. Building on our previous research on impossibility theorems, we use an agenda-theoretic approach to generalize the results and to determine the necessary and sufficient level of logical interconnection between the issues in an agenda for the impossibility theorems to arise. We demonstrate that (1) path-connectedness and even-negatability constitute the exact agenda condition for the oligarchy result stating that binarizing belief aggregation satisfying proposition-wise independence and deductive closure of collective beliefs yields the oligarchies under minor conditions; (2) negation-connectedness is the condition for the triviality result obtained by adding anonymity to the oligarchy result; and (3) blockedness is the condition for the impossibility result, which follows by adding completeness and consistency of collective beliefs. Moreover, we compare these novel findings with existing agenda-theoretic characterization theorems in judgment aggregation and belief binarization.

1 Introduction

The question of how to rationally aggregate individual beliefs into collective beliefs is important and ubiquitous in our society. In this regard, there has been abundant literature on collective decision theory, judgment aggregation, and probabilistic opinion pooling studies. One of the essential features of belief is that there are different types of beliefs. For example, some beliefs may be represented by traditional ”logical” languages—she believes that it is raining outside—while other types of beliefs might be modeled by ”probability functions”—she believes with 90 percent certainty that it is raining outside. Logical languages are similar to our natural languages and are therefore efficient for communicating with human agents, despite the fact that they sometimes suffer from significant information reduction, as in the case of the Lottery paradox. In contrast, probabilistic beliefs hold a fair amount of information to deal with uncertain environments, although people usually do not reach that level of precision. Considering these pros and cons of different types of beliefs, it is not surprising that different types of beliefs may be required at different stages of belief aggregation procedures depending on situations. If objective chances of issues in question can be given, it is epistemically preferable to report individual opinions in terms of degrees of belief. If the conclusion of an epistemic collective decision guides action (e.g., a jury verdict), it is practically better to report the collective opinion by means of plain logic. Therefore, rational belief

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 518–526, doi:10.4204/EPTCS.379.39

aggregation should be able to deal with different types of beliefs. One important topic in aggregating one type of belief into a different type of belief is aggregating probabilistic beliefs into collective binary beliefs (e.g., [11] [16]). We call this subject matter ”binarizing belief aggregation” [16]. We can observe these belief aggregation problems in expert panels, the scientific community, and political parties, whenever individuals' opinions can be encoded probabilistically, and the group's beliefs should be more decisive.

Similar to the development of judgment aggregation theory (e.g., [6] [15]), formulating axiomatic requirements, proving impossibility theorems, and identifying exact agenda conditions of impossibility theorems are natural and important research topics in binarizing belief aggregation. Building on our previous research on impossibility theorems, this paper uses an agenda-theoretic approach to determine which level of logical interconnection between the issues in an agenda is necessary and sufficient for the impossibility theorems to arise. Indeed, our previous paper assumed the agenda to be an algebra, which is the most typical when dealing with probabilistic beliefs. However, in practice, the agenda being an algebra might be quite demanding because we might not be interested in, for example, the conjunction of two propositions when making a collective decision on the two propositions. Besides the literature on judgment aggregation, agenda-theoretic approaches can be found in other fields as well. In probabilistic opinion pooling, general agendas were investigated to characterize linear pooling (e.g., [2] [3]). In the belief binarization problem, general agendas were studied to characterize impossibility theorems (e.g., [4] [5]).

In this study, we demonstrate that (1) path-connectedness and even-negatability constitute the exact agenda condition for the oligarchy result, which states that binarizing belief aggregation satisfying proposition-wise independence and deductive closure of collective beliefs yields the oligarchies under certain conditions; (2) negation-connectedness is the condition for the triviality result obtained by adding anonymity to the oligarchy result; and (3) blockedness is the condition for the impossibility result, which follows by adding completeness and consistency of collective beliefs. Moreover, we compare these novel findings with existing agenda-theoretic characterization theorems in judgment aggregation and belief binarization. All proofs of lemmas and theorems are provided in the full paper.

2 Binarizing Belief Aggregation and the impossibility results

We begin by introducing notations and definitions we will use throughout this paper. Let W be a finite non-empty set of possible worlds. An *agenda* $\mathscr{A}$ is a non-empty set of subsets of W that is closed under complement. Let $N := \{1,...,n\}(n \geq 2)$ be the set of individuals. For each $i \in N$, an individual i's *probabilistic belief* P_i is a function extendable to a probability function on the smallest algebra that includes $\mathscr{A}$. We denote by $\vec{P} := (P_1,...,P_n) = (P_i)_{i \in N}$ a profile of n individuals' probabilistic beliefs. Binarizing belief aggregation deals with individuals' probabilistic beliefs and the group's binary beliefs. Binary beliefs are represented by a function $Bel : \mathscr{A} \to \{0,1\}$. Sometimes, we abuse the notation and denote by Bel the *belief set* $\{A \in \mathscr{A} | \, Bel(A) = 1\}$, and $BelA$ is a shorthand for $A \in Bel$ or $Bel(A) = 1$. A binarizing aggregator (BA) F is a function that takes a profile $\vec{P}$ of n probabilistic beliefs in a given domain and returns a binary belief $F(\vec{P})$.

Now, let us define the axiomatic requirements on BA that are needed to formulate our impossibility results. First, we need the following rationality requirements on the domain and codomain of a BA.

- Universal Domain (UD): the domain of F is the set of all profiles $\vec{P}$ of n probabilistic beliefs
- Collective Deductive Closure (CDC)/Consistency (CCS)/Completeness (CCP): for all $\vec{P}$ in the domain, the resulting collective beliefs $F(\vec{P})$ is deductively closed/consistent/complete, respectively

Note that a binary belief Bel is deductively closed iff it holds that, if $Bel \vDash A(i.e., \bigcap Bel \subseteq A)$, then $BelA$ for all $A \in \mathscr{A}$. Moreover, Bel is consistent if $Bel \nvDash \emptyset$, and Bel is complete if $BelA$ or $Bel\overline{A}$ for all $A \in \mathscr{A}$ where $\overline{A}$ is the complement of A. Second, we enlist different rationality requirements on BAs themselves.

• Certainty Preservation (CP)/Zero Preservation (ZP): for all $A \in \mathscr{A}$, if $\vec{P}(A)(:= (P_1(A), \cdots, P_n(A))) = (1, ..., 1)/\vec{P}(A) = (0, ..., 0)$, then $F(\vec{P})(A) = 1/F(\vec{P})(A) = 0$, respectively, for all $\vec{P}$ in the domain of F.

• Anonymity (AN): $F((P_{\pi(i)})_{i \in N}) = F((P_i)_{i \in N})$ for all $\vec{P}$ in the domain of F and all permutation π on N.

• Independence (IND): for all $A \in \mathscr{A}$, there exists a function G_A such that $F(\vec{P})(A) = G_A(\vec{P}(A))$ for all $\vec{P}$ in the domain of F.

• Systematicity (SYS): there exists a function G such that $F(\vec{P})(A) = G(\vec{P}(A))$ for all $A \in \mathscr{A}$ and for all $\vec{P}$ in the domain of F.

Our previous paper [16] proved the following theorems under the assumption that $\mathscr{A}$ is an algebra with at least three elements besides the empty set and W, which we call a non-trivial algebra. We aim to relax this in this study.

1. (The Oligarchy Result) The only BAs satisfying UD, CP, ZP, IND, and CDC are the following oligarchies: there is a non-empty subset M of N such that

$$F(\vec{P})(A) = \begin{cases} 1 & \text{if } P_i(A) = 1 \text{ for all } i \in M \\ 0 & \text{otherwise} \end{cases}$$

for all $A \in \mathscr{A}$.

2. (The Triviality Result) The only BAs satisfying UD, CP, ZP, IND, CDC and AN are the oligarchy with $M = N$, which we call the trivial rule.

3. (The Impossibility Result) There is no BA satisfying UD, CP, IND, CCP, and CCS.

3 The Agenda Condition for the Oligarchy Result

This section presents and proves our first main result: the agenda condition for the oligarchy result. The following two agenda conditions have been extensively studied, as they characterize the most famous impossibility agendas in judgment aggregation.

Definition 1 (Path-connected and Even-negatable Agenda). *(1) For any $A, B \in \mathscr{A}$, we say that A conditionally entails B ($A \vDash^* B$) if there is a subset $\mathscr{Y} \subseteq \mathscr{A}$ that is consistent with A and $\overline{B}$*[1] *such that $\{A\} \cup \mathscr{Y} \vDash B$ (i.e., $\bigcap(\{A\} \cup \mathscr{Y}) \subseteq B$ and we write this as $A \vDash^*_{\mathscr{Y}} B$). An agenda $\mathscr{A}$ is path-connected (PC) if $A \vDash^{**} B$ for all contingent issues $A, B \in \mathscr{A}$, where $\vDash^{**}$ is the transitive closure of $\vDash^*$.*
(2) An agenda $\mathscr{A}$ is even-negatable (EN) iff there is a minimally inconsistent set $\mathscr{Y} \subseteq \mathscr{A}$ such that $\mathscr{Y}_{\neg\mathscr{X}} := (\mathscr{Y} \setminus \mathscr{X}) \cup \{\overline{A} | A \in \mathscr{X}\}$ is consistent for some subset $\mathscr{X} \subseteq \mathscr{Y}$ of even size.

Path-connectedness means that every two issues are connected by a path, i.e., a chain of conditional entailment relations. Regarding conditional entailment relation, let us mention a useful fact. If $A \vDash^*_{\mathscr{Y}} B$, it also holds that $\overline{B} \vDash^*_{\mathscr{Y}} \overline{A}$, and thus if $A \vDash^{**} B$, then $\overline{B} \vDash^{**} \overline{A}$. And even-negatability says that a minimally inconsistent subset of the agenda can be made consistent by negating some even number of its element. It is well-known that an agenda is even-negatable unless the propositions in the agenda are composed only with negation and biconditional from some logically independent propositions. Note that these two conditions are weaker than the agenda being a non-trivial algebra, which is the assumption on the agenda in [16].

[1] That is, $\mathscr{Y} \cup \{A\} \nvDash \emptyset$ and $\mathscr{Y} \cup \{\overline{B}\} \nvDash \emptyset$

Lemma 1. *Every non-trivial algebra is path-connected and even-negatable.*

From now on, we add one more assumption on $\mathscr{A}$ that $\emptyset \notin \mathscr{A}$(and thereby $W \notin \mathscr{A}$).[2] Thus, our agenda $\mathscr{A}$ is a complement-closed finite non-empty set of some contingent subsets of the underlying set W. The following lemma shows that path-connectedness is sufficient to obtain what is called the contagion lemma.

Lemma 2 (Agenda Condition for the Contagion Lemma)**.** *Let $\mathscr{A}$ be path-connected. If a BA F with UD satisfies CDC, CP, and IND, then it satisfies SYS.*

This lemma parallels the one in generalized opinion pooling of Dietrich & List (2017a): path-connectedness characterizes that if generalized OP satisfies CP and IND, then it satisfies SYS. In our lemma as well, its converse—if $\mathscr{A}$ is not path-connected, then there is a BA F on $\mathscr{A}$ satisfying CDC, CP, and IND but not SYS—also holds. The counterexample will be indicated in the proof of Theorem 1.

The following definition and lemma will be needed to prove our succeeding main theorem.

Definition 2 (Non-simple Agenda and Pair-negatable Agenda)**.** *(1) An agenda $\mathscr{A}$ is non-simple(NS) iff there is a minimally inconsistent subset $\mathscr{Y} \subseteq \mathscr{A}$ with $|\mathscr{Y}| \geq 3$.*

(2) An agenda $\mathscr{A}$ is pair-negatable iff there is a minimally inconsistent set $\mathscr{Y} \subseteq \mathscr{A}$ such that $\mathscr{Y}_{\neg Z}$ is consistent for some subset $\mathscr{Z} \subseteq \mathscr{Y}$ with $|\mathscr{Z}| = 2$.

Non-simple agendas can be used as a criterion for determining whether a given agenda has minimal complexity. Pair-negatable agendas are a special case of even-negatable agendas. The following lemma shows that a pair-negatable agenda is sufficient to be an even-negatable agenda, and a path-connected agenda already has a fairly complex structure.

Lemma 3. *(1) An agenda $\mathscr{A}$ is even-negatable iff $\mathscr{A}$ is pair-negatable.*

(2) If an agenda $\mathscr{A}$ is path-connected, then it is non-simple.

Now we prove that the agenda being path-connected and even-negatable is the sufficient and necessary condition for the oligarchy result.

Theorem 1 (Agenda Condition for the Oligarchy Result)**.** *Let $|N| \geq 3$. An agenda $\mathscr{A}$ is path-connected and even-negatable iff the only BAs on $\mathscr{A}$ satisfying UD, ZP, CP, IND, and CDC are the oligarchies.*

The only-if direction of the theorem generalizes the oligarchy result and shows that even if an agenda satisfies a weaker condition—path-connectedness and even-negatability—than a non-trivial algebra, the oligarchy result holds. If we examine the proof of the oligarchy result in [16] in detail, we can observe that the agenda condition was used solely to establish the following two facts:
(Fact 1) if $\vec{a} \leq \vec{b}$ and if $G(\vec{a}) = 1$, then $G(\vec{b}) = 1$ where G is a function satisfying $F(\vec{P})(A) = G(\vec{P}(A))$.
(Fact 2) if $\vec{a} + \vec{b} - \vec{1} \geq \vec{0}$ and if $G(\vec{a}) = 1$ and $G(\vec{b}) = 1$, then $G(\vec{a} + \vec{b} - \vec{1}) = 1$.
Therefore, to prove the only-if direction, it is enough to derive (Fact 1) from even-negatability and (Fact 2) from path-connectedness. The agenda conditions are only relevant to (Fact 1) and (Fact 2), and once we see that they hold then we can apply the proof of the oligarchy result in [16].

Our proof also reveals that if we assume the stronger property of SYS instead of IND, then Lemma 1 is not needed, and non-simplicity (NS) is sufficient to obtain the oligarchy result. This observation indicates that stronger properties of a BA lead to weaker agenda conditions for achieving the oligarchy result. To provide additional agenda conditions for the oligarchy result, let us introduce the concept of monotonicity (MON) for a BA as follows:

[2] In the following, especially in Theorem 2 and Theorem 3, we will use some results of Nehring & Puppe (2010), where the agenda consists of contingent issues. To describe our proof more simply, we adopt that assumption.

$$\text{(MON) If } \vec{P}(A) \leq \vec{P}'(A) \text{ and } F(\vec{P})(A) = 1, \text{ then } F(\vec{P}')(A) = 1$$

where $\leq$ is applied to each component of two vectors. If we assume MON, we can bypass the need to prove (Fact 1), thereby eliminating the requirement for the agenda to be even-negatable (EN). This is because (Fact 1) is already implied by SYS and MON. The following table illustrates the agenda conditions that are sufficient to achieve the oligarchy result based on different properties of BA:

	IND	SYS
without MON	PC, EN	NS, EN
with MON	PC	NS

It is noteworthy that the agenda condition required for our oligarchy result is the same as the one for the dictatorship and oligarchy results in judgment aggregation (e.g., [6] [1]). In our proof of the if-direction, we extend their counterexamples to our domain in a manner that satisfies UD, ZP, CP, IND, CDC, and CCS: the counterexample for a non-path-connected agenda is a minimal extension satisfying MON, as we do not exclude even-negatablility; the one for a non-even-negatable agenda is an extension satisfying not MON but SYS, as we do not exclude path-connectedness. So the proof follows a similar structure to those in judgment aggregation, but the ways of extension to construct counterexamples are not trivial—particularly the counterexample for not even-negatable agendas—, and so our proof includes novel ideas that are needed due to the difference between binary and probabilistic beliefs.

4 The Agenda Condition for the Triviality Result

This section presents and proves our second main result: the agenda condition for the triviality result. Stronger properties of a BA yield weak agenda conditions. Thus, one might ask whether the agenda condition for the oligarchy result can be weakened, if we add AN. We will demonstrate that the agendas that yield the triviality result can be characterized by negation-connectedness, which is also the agenda condition for an impossibility result of belief binarization methods as shown in [5].

Definition 3 (Negation-connected Agenda)**.** *An agenda $\mathscr{A}$ is negation-connected (NC) iff for every contingent issue $A \in \mathscr{A}$ it holds that $A \vDash^{**} \overline{A}$.*

So negation-connectedness means that every issue has a path to its complement. According to Proposition 1 in Dietrich & List (2021), the agenda being negation-connected is equivalent to the agenda being partitioned into subagendas each of which is path-connected, where a subagenda is a non-empty subset of the agenda that is closed under complementation.

The following lemma will be needed for the proof of the first part of the succeeding theorem. Part (1) allows us to consider the stronger condition, namely path-connectedness, than negation-connectedness to prove the triviality result. Part (2) will be used when the agenda is path-connected and not even-negatable.

Lemma 4. *(1) If the triviality result holds—i.e., the only BA on $\mathscr{A}$ satisfying UD, CDC, ZP, CP, IND, and AN is the trivial one—for any path-connected agenda $\mathscr{A}$, then the same holds for any negation-connected agenda.*

(2) If an agenda $\mathscr{A}$ is not even-negatable, then for any minimally inconsistent subset $\mathscr{Y} \subseteq \mathscr{A}$ and any even-sized subset $\mathscr{X} \subseteq \mathscr{Y}$ it holds that $\mathscr{Y}_{\neg\mathscr{X}}$ is also minimally inconsistent.

The following lemma will be needed for the proof of the second part of the succeeding theorem. This lemma looks technical but it is closely related to the notion of median point in the next section. Indeed, if $\mathscr{H}_0$ is the empty set, then $\bigcap \mathscr{M}$ is the set of all median points where $\mathscr{H}_0$ and $\mathscr{M}$ are defined in the following lemma.

Lemma 5. *Let $\mathscr{H}_0$ be the set $\{A \in \mathscr{A} \mid A \vDash^{**} \overline{A}$ and $\overline{A} \vDash^{**} A\}$. If $\mathscr{A}$ is not negation-connected, then there is a non empty subset $\mathscr{M} \subseteq \mathscr{A} \setminus \mathscr{H}_0$ such that for any minimally inconsistent set $\mathscr{Y} \subseteq \mathscr{A}$ it holds that $|\mathscr{Y} \cap \mathscr{M}| \leq 1$. Furthermore, for any minimally inconsistent set $\mathscr{Y} \subseteq \mathscr{A}$ intersecting $\mathscr{H}_0$ it holds that $|\mathscr{Y} \cap \mathscr{M}| = 0$. In addition, for $B \in \mathscr{A} \setminus \mathscr{H}_0$, it holds that $B \in \mathscr{M}$ iff $\overline{B} \notin \mathscr{M}$.*

Now let us prove the theorem that negation-connectedness is the sufficient and necessary condition for the triviality result.

Theorem 2 (Agenda Condition for the Triviality Result)**.** *An agenda $\mathscr{A}$ is negation-connected iff the only BA on $\mathscr{A}$ satisfying UD, ZP, CP, CDC, IND, and AN is the trivial one.*

The only-if direction of the theorem shows that the triviality result holds if the agenda is negation-connected, which is a generalization of the triviality result. The proof suggests further that, if we assume SYS, then non-simplicity (NS) becomes the sufficient condition to obtain the triviality result. In this case, neither EN nor MON is needed, unlike in the case of Theorem 1, as illustrated in the following table:

	IND	SYS
with or without MON	NC	NS

Compared to the case of the oligarchy result, when we add AN, we obtain the triviality result even under a weaker agenda condition: (i) instead of requiring path-connectedness (PC), negation-connectedness (NC) is sufficient, and (ii) the triviality result holds even when the agenda is not even-negatable (EN). The difference mentioned in (i) does not play a role in finding the sufficient condition according to Lemma 4. However, the necessary condition is not path-connectedness but negation-connectedness. In cases where the agenda is PC and EN, we can apply Theorem 1 since the oligarchy satisfying AN is the trivial one (i.e., the oligarchy with $M = N$). Thus, we only need to focus on the cases where the agenda is PC and not EN. When the agenda is assumed to be not EN, we encounter the following difficulty: to show the triviality result, we used (Fact 1), which could be proved if the agenda was assumed to be EN. Our strategy here is to prove a weaker claim than (Fact 1):
(Fact 1$''$) If $G(\vec{a}) = 1$, then $G(\vec{c}) = 1$ for all $\vec{c} \geq |2\vec{a} - \vec{1}|$.
The new claim (Fact 1$''$) is weaker than (Fact 1), as it only guarantees that vectors greater than $|2\vec{a} - \vec{1}|$ are mapped to 1, rather than all vectors greater than $\vec{a}$.

One might ask whether we can apply the proof presented in Dietrich & List (2021) to our theorem, or vice versa. However, there are differences between the two proofs. On the one hand, we cannot use their proof because, while they deal with probabilistic beliefs, we are dealing with profiles of probabilistic beliefs: in particular, for negation-connected agendas in our framework, we can only show (Fact 1$''$) instead of (Fact 1). On the other hand, since we have not relied on the assumption that $|N| \geq 2$, our proof can be applied to the context of belief binarization, where $|N| = 1$, and so we can recover their results.

The if-direction gives a counterexample of the triviality result when an agenda is not negation-connected, which implies the agenda being not path-connected. The counterexample presented in Theorem 1 is not applicable in this case because it does not satisfy AN. Moreover, there would be no counterexample if we only assumed an agenda to be not path-connected. This is the reason why we need to weaken path-connectedness to negation-connectedness, even though they fulfill the same role concerning the sufficient agenda condition for the triviality result.

Our counterexample for a non-negation-connected agenda is an extension of the belief binarization rule proposed in Dietrich & List (2021). We extend the rule while maintaining MON, but not minimally, which differs from the way of the extension in Theorem 1.

5 The Agenda Condition for the Impossibility Result

Now we will show that the agendas for the impossibility result can be characterized by blocked agendas.

Definition 4 (Blocked Agenda)**.** *An agenda $\mathscr{A}$ is blocked iff there is an issue $A \in \mathscr{A}$ such that $A \vDash^{**} \overline{A}$ and $\overline{A} \vDash^{**} A$.*

So a blocked agenda contains an issue that has a path to its complement. Recall that $\mathscr{H}_0$ is defined by the set $\{A \in \mathscr{A} \mid A \vDash^{**} \overline{A}$ and $\overline{A} \vDash^{**} A\}$. Then $\mathscr{A}$ is negation-connected iff $\mathscr{H}_0 = \mathscr{A}$, and $\mathscr{A}$ is blocked iff $\mathscr{H}_0 \neq \emptyset$. If $\mathscr{A}$ is negation-connected, then it is blocked. The following definition and lemma will be needed for the succeeding theorem.

Definition 5 (Median Point)**.** *Let $\mathscr{A}$ be an agenda on the set W of possible worlds. A possible world $m \in W$ is a median point iff for any minimally inconsistent subset $\mathscr{Y} \subseteq \mathscr{A}$, it holds that $|\{A \in \mathscr{Y} \mid m \in A\}| \leq 1$.*

So a median point is a possible world that is contained in at most one issue in every minimally inconsistent set. It is well-known in judgment aggregation that if a median point is guaranteed to exist, then we can easily construct an anonymous, complete, and consistent judgment aggregator where a median point is thought of as a default collective judgment unless everybody believes the issue being true/false at the median point to be false/true [14]. The following lemma states that the agenda not being blocked is the necessary and sufficient condition for the existence of a median point.

Lemma 6. *An agenda $\mathscr{A}$ is not blocked iff there is a median point.*

Now let us formulate and prove our last theorem.

Theorem 3 (Agenda Condition for the Impossibility Result)**.** *An agenda $\mathscr{A}$ is blocked iff there is no BA on $\mathscr{A}$ satisfying UD, CP, IND, CCP, and CCS.*

Indeed, CCS and CCP together are stronger assumptions than CDC. As a result, we obtain the impossibility result more easily, without assuming AN and non-dictatorship, and with a more relaxed agenda condition. The proof demonstrates that by adding SYS, the impossibility result still holds even without CP and even when no agenda condition is assumed—e.g., even when $\mathscr{A} = \{A, \overline{A}\}$.

The blocked agenda is also the agenda condition for the impossibility results on judgment aggregation with AN in [15] and belief binarization in [1]. Our counterexample for non-blocked agenda is an extension of the counterexample in Dietrich & List (2018). It is an extension that satisfies MON, but not minimally so. This is the same as the extension in Theorem 2, but different from the one in Theorem 1. Note that the median point m in the proof of this theorem plays the same role as $\mathscr{M}$ in the proof of Theorem 2. The only difference is that m is a possible world and $\mathscr{M}$ is a set of issues. This difference arises from assuming CDC versus assuming CCS and CCP.

6 Discussion

All the results in this paper are stated in Table 1: (1) path-connectedness and even-negatability constitute the exact agenda condition for the oligarchy result; (2) negation-connectedness is for the triviality result; and (3) blockedness is for the impossibility result. These new findings can be compared to the existing characterization theorems in judgment aggregation and belief binarization. Regarding (1), it has the same agenda condition as (1′) [1] and (3′) [6] in judgment aggregation. For (2), it is similar to (2″) [5] in belief binarization, with the difference being the use of ZP instead of CCS for for (2″). Since applying our proofs can weaken CCS to ZP, the agenda condition for (2′), which has not been discussed in the

There is no BA satisfying ...	Agenda Condition
(1) UD, ZP, CP and IND + CDC + Non-oligarchy	path-connected, even-negatable
(2) UD, ZP, CP and IND + CDC + AN + Non-triviality	negation-connected
(3) UD, CP and IND + CCS and CCP	blocked

There is no judgment aggregator satisfying ...	Agenda Condition
(1′) UD, ZP, CP and IND + CDC + Non-oligarchy	path-connected, even-negatable
(2′) UD, ZP, CP and IND + CDC + AN + Non-triviality	negation-connected
(3′) UD, CP and IND + CCS and CCP + non-dictatorship	path-connected, even-negatable
(4′) UD, CP and IND + CCS and CCP + AN	blocked

There is no belief binarization rule satisfying ...	Agenda Condition
(2″) UD, CCS , CP and IND + CDC + Non-triviality	negation-connected
(4″) UD, CCS, CP and IND + CCP	blocked

Table 1: Classification of Agenda Conditions for Impossibility Results

literature, is also negation-connected because an anonymous and independent judgment aggregator can be viewed as a belief binarization function. As for (3), it is similar to (4′) [15] in judgment aggregation and (4″) [4] in belief binarization.

Let us mention some further research topics. One might think that the rationality norms for collective binary beliefs could be weakened since adhering to deductive closure might be too demanding for group agents. Instead, we could focus on requiring group beliefs to respect consistency or pairwise consistency. By exploring these weaker norms, we can investigate stronger impossibility results. Furthermore, let us discuss how to obtain possibility results. For this purpose, it is advantageous that binarizing belief aggregation provides a framework that generalizes the problem of judgment aggregation or belief binarization. As in judgment aggregation, we can employ and study premise-based binarizing belief aggregation methods. Alternatively, we can combine an individual belief binarization procedure with judgment aggregation. If we assume that linear or geometric pooling methods are very natural given individual credences, we can apply belief binarization methods to the pooled group credence. Of course, we can also come up with new procedures that cannot be reduced to existing methods. Ultimately, we should keep in mind that binarizing belief aggregation is an *epistemic* collective decision problem. Therefore, we should be concerned about which methods accurately track the truth. One natural approach would be to investigate belief binarization methods that minimize the expected distance from the truth in light of the group's pooled credence. In conclusion, binarizing belief aggregation opens a new research area in which various procedures of belief aggregation, different studies on the relation between credences and beliefs, and epistemic decision theory can be combined and explored.

Acknowledgements We would like to express our sincere gratitude to Hannes Leitgeb and Christian List for their invaluable feedback and profound insights. The research conducted by the first author benefited from the generous support of the German Academic Scholarship Foundation and the Alexander von Humboldt Foundation.

References

[1] Franz Dietrich & Christian List (2008): *Judgment aggregation without full rationality*. Social Choice and Welfare 31, pp. 15–39, doi:10.1007/s00355-007-0260-1.

[2] Franz Dietrich & Christian List (2017): *Probabilistic opinion pooling generalized. Part one: general agendas*. Social Choice and Welfare 48, pp. 747–786, doi:10.1007/s00355-017-1034-z.

[3] Franz Dietrich & Christian List (2017): *Probabilistic opinion pooling generalized. Part two: the premise-based approach*. Social Choice and Welfare 48, pp. 787–814, doi:10.1007/s00355-017-1035-y.

[4] Franz Dietrich & Christian List (2018): *From degrees of belief to binary beliefs: Lessons from judgement-aggregation theory*. Journal of Philosophy 115(5), pp. 787–814, doi:10.5840/jphil2018115516.

[5] Franz Dietrich & Christian List (2021): *The relation between degrees of belief and binary beliefs: a general impossibility theorem*. In Igor Douven, editor: Lotteries, Knowledge, and Rational Belief. Essays on the Lottery Paradox, Cambridge University Press, pp. 224–254, doi:10.1017/9781108379755.012.

[6] Elad Dokow & Ron Holzman (2018): *Aggregation of binary evaluations*. Journal of Economic Theory 145, p. 495–511, doi:10.1016/j.jet.2007.10.004.

[7] Peter Gärdenfors (2006): *A representation theorem for voting with logical consequences*. Economics and Philosophy 22(2), pp. 181–190, doi:10.1017/S026626710600085X.

[8] Christian Genest & James V. Zidek (1986): *Combining probability distributions: a critique and annotated bibliography*. Statistical Science 1(1), pp. 114–135, doi:10.1214/ss/1177013831.

[9] Frederik Herzberg (2015): *Aggregating infinitely many probability measures*. Theory and Decision 78(2), pp. 319–337, doi:10.1007/s11238-014-9424-5.

[10] Frederik Herzberg (2017): *Respect for experts vs. respect for unanimity: The liberal paradox in probabilistic opinion pooling*. Economic Letters 151, pp. 44–47, doi:10.1016/j.econlet.2016.12.012.

[11] Magdalena Ivanovska & Marija Slavkovik (2019): *Aggregating Probabilistic Judgments*. In: Proceedings Seventeenth Conference on Theoretical Aspects of Rationality and Knowledge, EPCTS 297, pp. 273–292, doi:10.4204/EPTCS.297.18.

[12] Christian List & Philip Pettit (2002): *Aggregating Sets of Judgments: An Impossibility Result*. Economics and Philosophy 18(1), pp. 89–110, doi:10.1017/S0266267102001098.

[13] Kevin J. McConway (1981): *Marginalization and linear opinion pools*. Journal of the American Statistical Association 76, pp. 410–414, doi:10.1080/01621459.1981.10477661.

[14] Klaus Nehring & Clemens Puppe (2007): *The structure of strategy-proof social choice - Part I: General characterization and possibility results on median spaces*. Journal of Economic Theory 135, pp. 269–305, doi:10.1016/j.jet.2006.04.008.

[15] Klaus Nehring & Clemens Puppe (2010): *Abstract Arrovian aggregation*. Journal of Economic Theory 145, p. 467–494, doi:10.1016/j.jet.2010.01.010.

[16] Minkyung Wang: *Aggregating Individual Credences into Collective Binary Beliefs: An Impossibility Result*. Unpublished ms.

Knowledge-wh and False Belief Sensitivity: A Logical Study (An Extended Abstract)

Yuanzhe Yang
Department of Philosophy and Religious Studies
Peking University
Beijing, China
1900014924@pku.edu.cn

In epistemic logic, a way to deal with knowledge-wh is to interpret them as a kind of mention-some knowledge (MS-knowledge). But philosophers and linguists have challenged both the sufficiency and necessity of such an account: some argue that knowledge-wh has, in addition to MS-knowledge, also a sensitivity to false belief (FS); others argue that knowledge-wh might only imply mention-some true belief (MS-true belief). In this paper, we offer a logical study for all these different accounts. We apply the technique of bundled operators, and introduce four different bundled operators - $[\mathsf{tB}^{\mathtt{MS}}]^x\phi := \exists x([\mathsf{B}]\phi \wedge \phi)$, $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi := \exists x([\mathsf{B}]\phi \wedge \phi) \wedge \forall x([\mathsf{B}]\phi \to \phi)$, $[\mathsf{K}^{\mathtt{MS}}]^x\phi := \exists x[\mathsf{K}]\phi$ and $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi := \exists x[\mathsf{K}]\phi \wedge \forall x([\mathsf{B}]\phi \to \phi)$ -, which characterize the notions of MS-true belief, MS-true belief with FS, MS-knowledge and MS-knowledge with FS respectively. We axiomatize the four logics which take the above operators (as well as $[\mathsf{K}]$) as primitive modalities on the class of $S4.2$-constant-domain models, and compare the patterns of reasoning in the obtained logics, in order to show how the four accounts of knowledge-wh differ from each other, as well as what they have in common.

1 Introduction

In standard epistemic logic, for the most time, we deal with *propositional knowledge* (or knowledge-*that*): that is, an agent knows *that* ϕ, where ϕ is a certain proposition. However, this clearly does not exhaust our daily use of the notion of "knowledge". Besides knowledge-that, we also frequently talk about various kinds of *knowledge-wh*: for example, I know *how* to ride a bike, I know *who* proved the incompleteness theorems, I know *when* a certain meeting is held, I know *where* to buy a certain book, I know *what* is the password of my computer, I know *why* a certain event happens, etc.

Thus, besides standard propositional knowledge, knowledge-wh also seems to be an interesting subject for epistemic logic to study. There are already a number of logical studies of various kinds of knowledge-wh (e.g. know whether in [3], know why in [24], know how in [19],[4], [11], [12] and [22], just to name a few), and a more general framework for logics of knowledge-wh is also proposed in [20]. In this paper, following [20], we will also focus mainly on the general logical structures shared by various kinds of knowledge-wh.

As suggested in [20] (following the philosophical stance of the so-called "intellectualism" initiated in [17]), in many cases, knowledge-wh can be interpreted as a kind of *mention-some knowledge* (MS-knowledge for short): for example, I know how to prove a theorem, iff *there exists* some proof such that I know that this proof is a proof for the theorem; I know where to buy newspapers, iff *there exists* some place where I know I can buy newspapers, etc. Then, in such cases, it seems that the logical structure of knowledge-wh can be formally captured by the first-order modal formula $\exists x[\mathsf{K}]\phi(x)$.[1]

[1] However, as it is also noted in [20], in some other situations, it seems more natural to interpret knowledge-wh in terms of mention-*all*, rather than mention-*some*, knowledge. For example, when I say "I know who came to the meeting yesterday",

R. Verbrugge (Ed.): Theoretical Aspects of
Rationality and Knowledge 2023 (TARK 2023)
EPTCS 379, 2023, pp. 527–544, doi:10.4204/EPTCS.379.40

However, while it is quite clear that in many situations, knowledge-wh does involve some kind of mention-some structure, it is not as clear whether MS-knowledge really is the right account for knowledge-wh in these situations. In fact, both the sufficiency and necessity of such an account are challenged.

For example, as it is argued in [5], [14], [7] and [23], knowledge-wh may not only involve mention-some knowledge, but also involve *false belief sensitivity* (FS for short). Let's consider the following scenario, adapted from one offered in [5], to illustrate this point.

Example 1.1 *There are two stores, Newstopia and Paperworld. Newstopia sells newspapers, while Paperworld sells only stationery. Now, Alice* knows *that Newstopia sells newspapers, but also* believes *erroneously that Paperworld sells newspapers.*

In such a scenario, it is natural to judge that that Alice does not know where to buy newspapers (psychological experiments conducted in [14] also show that such an intuition is shared by many people): even though she has a MS-knowledge concerning where to buy newspapers, it seems that her false belief that Paperworld sells newspapers would corrupt her knowledge-where.

Hence, maybe knowledge-wh should be sensitive to false belief: that is, even under an MS-reading, maybe MS-knowledge should not be characterized by $\exists x[\mathsf{K}]\phi(x)$ alone, but should rather be characterized by $\exists x[\mathsf{K}]\phi(x) \wedge \forall x([\mathsf{B}]\phi(x) \to \phi(x))$.

On the other hand, the necessity of the MS-knowledge account is also doubted. For example, as it is argued in [1], it seems that knowledge-wh is subject to a kind of *epistemic luck* which is not consistent with propositional knowledge. Let's consider the following scenario, adapted from one offered in [1], to illustrate this point.

Example 1.2 *Suppose that Bob believes that w is a way to change light bulbs, and w is indeed a reliable way to do so. His belief is obtained by reading an instruction in a book. However, unknown to him, all other contents in the book are erroneous, and it is merely due to a very rare print error that the instruction he read is correct.*

In this case, Bob's true belief that w is a way to change light bulbs is too lucky to be counted as his *knowledge*; but nevertheless, it still seems natural to judge that Bob knows how to change light bulbs.

Then, it seems that sometimes a mention-some true belief (MS-true belief for short), i.e. $\exists x([\mathsf{B}]\phi(x) \wedge \phi(x))$, is enough for knowledge-wh. (In philosophical discussions, such a stance is sometimes called "revisionary intellectualism", which is first proposed in [2], in contrast to intellectualism.)

Of course, none of the arguments presented above is decisive. But they do reveal an enormous complexity in the question concerning the nature of knowledge-wh. Hitherto, no consensus concerning this question is reached in philosophical discussions, and nor will we offer a determinate answer here. On the contrary, in this paper, we will study *all* the accounts mentioned above in a formal way.

In order to do so, we apply the technique of "bundled operators"[2]. The general idea is that we pack a complex first-order modal formula (e.g. $\exists x[\mathsf{K}]\phi(x) \wedge \forall x([\mathsf{B}]\phi(x) \to \phi(x))$) into the semantics of a single operator, and study the logic which takes such an "bundled operator" as primitive modality. By working in such languages with limited expressivity, we can focus on the behavior of the epistemic notion in which

it may mean that I know *all* the people who came to that meeting, which should probably be formalized as, for example, $\forall x(\phi(x) \to [\mathsf{K}]\phi(x))$ or $\forall x([\mathsf{K}]\phi(x) \vee [\mathsf{K}]\neg\phi(x))$. We will not deal with the mention-all reading of knowledge-wh in this paper, since the behavior of mention-all knowledge is rather different from mention-some knowledge, and it thus seems better to study it independently elsewhere.

In fact, axiomatization of mention-all knowledge in terms of $\forall x([\mathsf{K}]\phi(x) \vee [\mathsf{K}]\neg\phi(x))$ has been studied in [25], an unpublished undergraduate thesis.

[2] For a detailed introduction of such an idea, see [20] and [21].

we are really interested, without being distracted by irrelevant notions which can also be expressed in a stronger language. Moreover, with the help of bundled operators, we can study the complex notions in a compact manner.

In this paper, then, we will study the following four different bundled operators:[3]

$$\begin{array}{lcl}
[\mathsf{tB}^{\mathtt{MS}}]^x\phi(x) & := & \exists x([\mathsf{B}]\phi(x)\wedge\phi(x))\\
[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi(x) & := & \exists x([\mathsf{B}]\phi(x)\wedge\phi(x))\wedge\forall x([\mathsf{B}]\phi(x)\to\phi(x))\\
[\mathsf{K}^{\mathtt{MS}}]^x\phi(x) & := & \exists x[\mathsf{K}]\phi(x)\\
[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi(x) & := & \exists x[\mathsf{K}]\phi(x)\wedge\forall x([\mathsf{B}]\phi(x)\to\phi(x)).
\end{array}$$

We will axiomatize the logics which take these operators plus an operator for propositional knowledge as primitive modalities on the class of $S4.2$-models, a class of models which characterizes knowledge, belief and their interactions in a reasonable way. Completeness results will also be presented. Moreover, we will compare the obtained logics, in order to show the differences and commonalities in the ways we reason about knowledge-wh, propositional knowledge and belief, which are logically implied by the different accounts of knowledge-wh.

2 First-order $S4.2$-models

First, we introduce the models we use to characterize knowledge and belief on the semantic level.

Since first-order quantifiers are involved in the notions of MS-knowledge, MS-true belief and FS, we will use *first-order Kripke models* as the semantic basis. We fix a set of predicates $\mathscr{P}$. A first-order Kripke model, then, is defined as follow:[4]

Definition 2.1 *A* first-order Kripke model *is a* 4*-tuple* $\mathscr{M}=(W,D,R,\rho)$*, where*

- $W\neq\emptyset$ *is the set of epistemically possible worlds of the model;*
- $D\neq\emptyset$ *is the domain of the model;*
- $R\subseteq W^2$ *is the accessibility relation among the possible worlds, which characterizes epistemic indistinguishability;*
- $\rho:\mathscr{P}\times W\to\wp(D^{<\omega})$ *assigns each n-ary predicate an n-ary relation on each possible world.*

(We may abbreviate the term "first-order Kripke model" simply as "model" in the following discussions.)

Note that such a model can be interpreted rather freely on the conceptual level, so that it can characterize various kinds of knowledge-wh. For example, if we want to characterize the knowledge-how of an agent, then the elements in D can be interpreted as different methods or devices available for the agent in question, and a predicate $P\in\mathscr{P}$ can be interpreted as a certain goal, while $a\in\rho(P,w)$ reads "at the epistemically possible world w, a is a way to achieve P". Similarly, if we want to characterize knowledge-where, then the elements in D can be interpreted as different locations accessible for the agent, while predicates in $\mathscr{P}$ are interpreted as properties of these locations. Of course, in a similar fashion, different models can also be used to characterize knowledge-who, knowledge-when or knowledge-what.

[3] The bundled operator $[\mathsf{K}^{\mathtt{MS}}]^x$ is first introduced in [20] (the notation used there is $\Box^x$, though); later, further study concerning its decidability and complexity is presented in [13], and axiomatization in [18]. The result presented in this paper concerning this operator (namely, the axiomatization on $S4.2$), however, is new.

On the other hand, $[\mathsf{tB}^{\mathtt{MS}}]$, $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]$ and $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]$ are all novel bundled operators that have not yet been studied in literature.

[4] In this paper, we will not introduce function symbols and constants to our language. Hence, we will also not consider functions and constants in the following definition.

Also note that we only consider *constant-domain* models here: all possible worlds in a model share the same domain. This is mainly in order to avoid technical and conceptual complexities, and we believe this is indeed a reasonable (though inevitably idealized) assumption.

Of course, since we use first-order Kripke models to characterize the epistemic states of an agent, the Kripkean part of such models should also possess certain frame properties.

It is a popular choice to use $S5$-models to characterize an agent's knowledge, but we will not use such models in this paper. This is mainly because we need to deal with both knowledge and belief, as well as the interactions between them (moreover, in our discussion, the notion of belief should be interpreted in a rather strong sense, so we would prefer interaction principles like $[\mathsf{B}]\phi \to [\mathsf{B}][\mathsf{K}]\phi$ to hold), and we must also allow the possibility for false belief, in order for the notion of FS to make any sense at all. This, however, seems to be a difficult task when knowledge is characterized by $S5$-models.

Hence, we will use $S4.2$-*models* instead - that is, models which are reflexive, transitive and strongly convergent.[5] The formal definition is as follow:

Definition 2.2 *A frame (W,R) is* strongly convergent*, iff for all $w \in W$, there is some $u \in W$ s.t. for all $v \in W$, if wRv, then vRu.*

A model based on a reflexive, transitive and strongly convergent frame is called an $S4.2$-model.

We find such models attractive, because the class of $S4.2$-models validates the logic of knowledge **S4.2**, in which belief can be reasonably *defined* in terms of knowledge by the definition $[\mathsf{B}]\phi := \langle\mathsf{K}\rangle[\mathsf{K}]\phi$ (as explained in [9], the underlying idea is that, if one knows that she does not know something, then she would not believe it; and if she does not believe something, then she would know by introspection that she does not know it). Moreover, the logic for the belief defined in this way is **KD45**, and we also have many intuitive interaction principles between knowledge and belief (e.g. $[\mathsf{K}]\phi \to [\mathsf{B}]\phi$, $[\mathsf{B}]\phi \to [\mathsf{K}][\mathsf{B}]\phi$, $\neg[\mathsf{B}]\phi \to [\mathsf{K}]\neg[\mathsf{B}]\phi$, $[\mathsf{B}]\phi \to [\mathsf{B}][\mathsf{K}]\phi$). (It is Lenzen who first proposed **S4.2** as a logic for knowledge in [9] and [10], from a syntatic perspective. Later, Stalnaker also studied **S4.2** from a more semantic perspective in [16].)

Moreover, as it is noted by Stalnaker in [16], in an $S4.2$-frame (W,R), we can define the following relation R_{B}, which corresponds to the notion of belief defined in terms of knowledge:

Definition 2.3 *Given a frame (W,R), $R_{\mathsf{B}} \subseteq W^2$ is the relation which satisfies that for all $w,u \in W$, $wR_{\mathsf{B}}u$ iff for all $v \in W$ s.t. wRv, vRu.*

It is not hard to check that if (W,R) is an $S4.2$-frame, then (W,R_{B}) is $KD45$. Moreover, after we formally introduce the languages and their semantics, it will be easy to check that R_{B} corresponds to $[\mathsf{B}]$ in exactly the way R corresponds to $[\mathsf{K}]$.

3 Languages and semantics

Now, we introduce the languages for the bundled operators, as well as their exact semantics.

We first fix a set of variables $\mathbf{X}$. Then, for any $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$, the corresponding language $\mathscr{L}([\mathsf{K}_{\mathtt{wh}}])$ (and also $\mathscr{L}_{\approx}([\mathsf{K}_{\mathtt{wh}}])$) is defined as follow:

[5] Here, we use the notion of *strong* convergence to define $S4.2$-models; but elsewhere, when defining $S4.2$-models, the notion of *weak* convergence might be used instead (A frame (W,R) is *weakly* convergent, iff for all $w,v,v' \in W$ s.t. wRv and wRv', there is some $u \in W$ s.t. vRu and $v'Ru$). Standard modal logic cannot distinguish these two kinds of models (as noted in [16]), but some of the languages studied in this paper are strong enough to distinguish them.

We choose the stronger notion of convergence here, because it seems more favorable both technically and conceptually. This is also in accordance with Stalnaker's note in [16].

Definition 3.1 *$\mathscr{L}([\mathsf{K}_{\mathtt{wh}}])$-formulas are defined recursively as follow:*

$$\phi ::= P(y_1, ..., y_n) \mid \neg\phi \mid \phi \wedge \phi \mid [\mathsf{K}]\phi \mid [\mathsf{K}_{\mathtt{wh}}]^x\phi$$

where $P \in \mathscr{P}$, $n \geq 0$ and $x, y_1, ..., y_n \in \mathbf{X}$.

$\neg[\mathsf{K}]\neg\phi$ is denoted as $\langle\mathsf{K}\rangle\phi$; $[\mathsf{B}]\phi$ is an abbreviation for $\langle\mathsf{K}\rangle[\mathsf{K}]\phi$.

$\vee$, $\rightarrow$ and $\leftrightarrow$ are defined in the usual way.

Moreover, let $\mathscr{L}_{\approx}([\mathsf{K}_{\mathtt{wh}}])$ be the language obtained by further adding an identity relation $\approx$ to $\mathscr{L}([\mathsf{K}_{\mathtt{wh}}])$.[6]

Corresponding to our definition of the bundled operators, we define the semantics for the above languages recursively as follow:

Definition 3.2 *Given a model $\mathscr{M} = (W, D, R, \rho)$, a $w \in W$ and an assignment σ from $\mathbf{X}$ to D:*

$\mathscr{M}, w, \sigma \vDash P(x_1, ..., x_n)$	$\Longleftrightarrow$	$(\sigma(x_1), ..., \sigma(x_n)) \in \rho(P, w)$
$\mathscr{M}, w, \sigma \vDash x \approx y$	$\Longleftrightarrow$	$\sigma(x) = \sigma(y)$
$\mathscr{M}, w, \sigma \vDash \neg\phi$	$\Longleftrightarrow$	$\mathscr{M}, w, \sigma \nvDash \phi$
$\mathscr{M}, w, \sigma \vDash \phi \wedge \psi$	$\Longleftrightarrow$	*$\mathscr{M}, w, \sigma \vDash \phi$ and $\mathscr{M}, w, \sigma \vDash \phi$*
$\mathscr{M}, w, \sigma \vDash [\mathsf{K}]\phi$	$\Longleftrightarrow$	*For all $v \in W$, if wRv, then $\mathscr{M}, v, \sigma \vDash \phi$*
$\mathscr{M}, w, \sigma \vDash [\mathsf{tB}^{\mathtt{MS}}]^x\phi$	$\Longleftrightarrow$	*There is some $a \in D$, s.t. $\mathscr{M}, w, \sigma[x \mapsto a] \vDash [\mathsf{B}]\phi \wedge \phi$*
$\mathscr{M}, w, \sigma \vDash [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$	$\Longleftrightarrow$	*(i) There is some $a \in D$, s.t. $\mathscr{M}, w, \sigma[x \mapsto a] \vDash [\mathsf{B}]\phi \wedge \phi$* *(ii) For all $b \in D$, $\mathscr{M}, w, \sigma[x \mapsto b] \vDash [\mathsf{B}]\phi \rightarrow \phi$*
$\mathscr{M}, w, \sigma \vDash [\mathsf{K}^{\mathtt{MS}}]^x\phi$	$\Longleftrightarrow$	*There is some $a \in D$, s.t. $\mathscr{M}, w, \sigma[x \mapsto a] \vDash [\mathsf{K}]\phi$*
$\mathscr{M}, w, \sigma \vDash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$	$\Longleftrightarrow$	*(i) There is some $a \in D$, s.t. $\mathscr{M}, w, \sigma[x \mapsto a] \vDash [\mathsf{K}]\phi$* *(ii) For all $b \in D$, $\mathscr{M}, w, \sigma[x \mapsto b] \vDash [\mathsf{B}]\phi \rightarrow \phi$*

where $\sigma[x \mapsto a]$ is the assignment which maps x to a, and agrees with σ on any other point.

Note that we need not introduce an independent operator for belief, since $[\mathsf{B}]\phi$ is already defined by $\langle\mathsf{K}\rangle[\mathsf{K}]\phi$ in the languages given above. It is also not hard to check that on $S4.2$-models, the semantics for $[\mathsf{B}]\phi$ defined this way is indeed the following one:

$$\boxed{\mathscr{M}, w, \sigma \vDash [\mathsf{B}]\phi \iff \text{For all } v \in W, \text{ if } wR_{\mathsf{B}}v, \text{ then } \mathscr{M}, v, \sigma \vDash \phi}$$

4 The logics

Then, we introduce the four logics, corresponding to our four accounts of knowledge-wh respectively. Their axiomatizations are all obtained in a similar fashion: generally speaking, we start from a standard **S4.2** system for $[\mathsf{K}]$, and then add axioms and rules to describe the behaviors of the bundled operators.

Below is a list of schemas of axioms and rules that will be used to offer the axiomatizations (in which the operator $[\mathsf{K}_{\mathtt{wh}}]$ should be substituted by $[\mathsf{tB}^{\mathtt{MS}}]$, $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]$, $[\mathsf{K}^{\mathtt{MS}}]$ or $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]$ in the corresponding logics):[7]

[6] In the following discussions, we will be working in the language $\mathscr{L}([\mathsf{K}_{\mathtt{wh}}])$ when we do not specifically mention the language in which we are working. We will make it clear whenever we switch our working language to $\mathscr{L}_{\approx}([\mathsf{K}_{\mathtt{wh}}])$.

[7] Note that when we use the notation $\phi[y/x]$ to denote the formula obtained by replacing every free occurrences of x in ϕ with y, we also implicitly assume that y is *admissible* for x in ϕ: that is, y does not appear in the scope of any operator of the form $[\mathsf{K}_{\mathtt{wh}}]^y$ in ϕ.

Axioms

$\mathtt{TBtoK_{wh}}$	$([\mathsf{B}]\phi \wedge \phi)[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi$	$\mathtt{KtoK_{wh}}$	$[\mathsf{K}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi$
$\mathtt{K_{wh}toFS}$	$[\mathsf{K}_{\mathtt{wh}}]^x\phi \to ([\mathsf{B}]\phi \to \phi)[y/x]$	$\mathtt{BtoBK_{wh}}$	$[\mathsf{B}]\phi[y/x] \to [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x\phi$

Rules

$\mathtt{K_{wh}toTB}$	$\dfrac{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to \neg([\mathsf{B}]\phi \wedge \phi))\cdots)}{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to \neg[\mathsf{K}_{\mathtt{wh}}]^x\phi)\cdots)}$
$\mathtt{K_{wh}toK}$	$\dfrac{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to \neg[\mathsf{K}]\phi)\cdots)}{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to \neg[\mathsf{K}_{\mathtt{wh}}]^x\phi)\cdots)}$
$\mathtt{FS\&BtoK_{wh}}$	$\dfrac{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi \to \phi))\cdots)}{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi))\cdots)}$
$\mathtt{FS\&KtoK_{wh}}$	$\dfrac{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi \to \phi))\cdots)}{\vdash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{K}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi))\cdots)}$

(In all the rules above, n can be any natural number, and we require that $x \notin \bigcup_{i \le n} FV(\psi_i)$)

By using rules like $\mathtt{K_{wh}toTB}$ or $\mathtt{FS\&KtoK_{wh}}$, we have sacrificed some intuitiveness for technical reasons, but the underlying idea is straightforward: for example, $\mathtt{K_{wh}toTB}$ essentially says $[\mathsf{K}_{\mathtt{wh}}]^x\phi \to \exists x([\mathsf{B}]\phi \wedge \phi)$, and $\mathtt{FS\&KtoK_{wh}}$ says $\forall x([\mathsf{B}]\phi \to \phi) \wedge [\mathsf{K}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi$, in languages where the existential and universal quantifiers are not available.

With the help of the above axioms and rules, then, we can give the following four logics:

$\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]}$	$\mathbf{S4.2}^{[\mathsf{K}]} \oplus \{\mathtt{TBtoK_{wh}}, \mathtt{K_{wh}toTB}\}$
$\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}$	$\mathbf{S4.2}^{[\mathsf{K}]} \oplus \{\mathtt{K_{wh}toFS}, \mathtt{BtoBK_{wh}}, \mathtt{K_{wh}toTB}, \mathtt{FS\&BtoK_{wh}}\}$
$\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]}$	$\mathbf{S4.2}^{[\mathsf{K}]} \oplus \{\mathtt{KtoK_{wh}}, \mathtt{K_{wh}toK}\}$
$\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$	$\mathbf{S4.2}^{[\mathsf{K}]} \oplus \{\mathtt{K_{wh}toFS}, \mathtt{BtoBK_{wh}}, \mathtt{K_{wh}toK}, \mathtt{FS\&KtoK_{wh}}\}$

Moreover, for any $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$, when we work in the language $\mathscr{L}_{\approx}([\mathsf{K}_{\mathtt{wh}}])$, let $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\approx}$ be the logic defined as follows:

$$\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\approx} \;\Big|\; \mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]} \oplus \{x \approx x,\ x \approx y \to (\phi[x/z] \to \phi[y/z]),\ x \not\approx y \to [\mathsf{K}](x \not\approx y)\}$$

Note that all the logics given here are non-normal, since they are all non-aggregative: that is, $[\mathsf{K}_{\mathtt{wh}}]^x\phi \wedge [\mathsf{K}_{\mathtt{wh}}]^x\psi \to [\mathsf{K}_{\mathtt{wh}}]^x(\phi \wedge \psi)$ is not an inner theorem of $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ (or $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\approx}$) for any $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}],$ $[\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$ (in fact, in all these logics, $[\mathsf{K}_{\mathtt{wh}}]^x Px \wedge [\mathsf{K}_{\mathtt{wh}}]^x \neg Px$ is consistent). Moreover, some of the logics are even non-monotone, as we will see below.

Then, we show the completeness theorem for these logics.

Since we are now dealing with bundled operators with more complex structures, the strategy to prove completeness theorems for the case of $[\mathsf{K}^{\mathtt{MS}}]$ in [20] and [18] cannot be directly applied here (moreover, axiomatization of the logic of $[\mathsf{K}^{\mathtt{MS}}]$ on $S4.2$ has also not yet been studied). Hence, we will develop a new strategy to prove completeness theorems for all the above logics in a uniform way.

Theorem 4.1 $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ *(as well as* $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\approx}$*) is sound and strongly complete w.r.t. the class of $S4.2$-constant-domain models, where* $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$.

PROOF. We only sketch the general idea of the proof here. A detailed proof for the case of $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ can be found in the appendix.

Generally, we use maximal consistent sets (MCS) of formulas which also contain certain witness formulas to construct the canonical model. The main difficulty is to ensure at the same time that (i) every MCS in the model contains all the witness formulas we need, (ii) every formula of the form $\langle \mathsf{K} \rangle \phi$ in an MCS has some accessible MCS containing ϕ as its witness, and (iii) the canonical model is an $S4.2$-constant-domain model.

In order to construct such a model, we use a step-by-step method. We start from a consistent set Γ_0, and extend consistent sets to MCS, add new formula sets as witnesses for formulas of the form $\langle \mathsf{K} \rangle \phi$, and add witness formulas to formula sets during the same process. The key is to ensure, at each step of the construction, that every formula set except Γ_0 is *finite*, and all the information contained in a set is recorded in its predecessor with a formula of the form $\langle \mathsf{K} \rangle \phi$. This ensures that we can always add witness formulas to formula sets using rules like $\mathtt{K_{wh}toTB}$ and $\mathtt{FS\&KtoK_{wh}}$.

Then, after countably many steps, we obtain a model which satisfies both (i) and (ii), and is also an $S4$-constant-domain model. Finally, we add another set of MCSs to the model to make it strongly convergent, so that we can obtain an $S4.2$-model. □

Remark 4.2 *The above logics also have some interesting technical aspects.*

For example, it is shown in [13] that the language $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}])$ cannot distinguish constant-domain and increasing-domain models in general. However, when we confine the models to $S4.2$-ones, $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}])$ can *distinguish constant-domain and increasing-domain models, and consequently, $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]}$ is not sound w.r.t. the class of $S4.2$-increasing-domain models (e.g. $\langle \mathsf{K} \rangle [\mathsf{K}^{\mathtt{MS}}]^x \phi \to [\mathsf{K}^{\mathtt{MS}}]^x \langle \mathsf{K} \rangle \phi$ is an inner theorem of $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]}$, but is not valid on $S4.2$-increasing-domain models). In fact, for all $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$, $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ is not sound w.r.t. $S4.2$-increasing-domain models.*

Another interesting fact is that $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}$ and $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ are able to distinguish $S4.2$-models (defined in terms of strong *convergence) and models which are reflexive, transitive but only weakly convergent. The axiom $\mathtt{BtoBK_{wh}}$: $[\mathsf{B}]\phi[y/x] \to [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x \phi$ does the trick. When $[\mathsf{K}_{\mathtt{wh}}] = [\mathsf{tB}^{\mathtt{MS}}]$ or $[\mathsf{K}^{\mathtt{MS}}]$, on the other hand, we also have $[\mathsf{B}]\phi[y/x] \to [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x \phi$ as an inner theorem of $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$, but in this case, the formula does not have the power to distinguish strong and weak convergence, and consequently, $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]}$ and $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]}$ are also sound w.r.t. the class of reflexive, transitive and weakly convergent models.*

5 Comparisons

Now, we have the formal ground to compare the different accounts of knowledge-wh.

5.1 Differences

An interesting difference among the different accounts of knowledge-wh concerns the ways these accounts interact with propositional knowledge.

For example, consider *positive introspection*. Since we take $\mathbf{S4.2}$ to be the underlying logic for propositional knowledge, which is stronger than $\mathbf{S4}$, it is clear that propositional knowledge satisfies positive introspection: $[\mathsf{K}]\phi \to [\mathsf{K}][\mathsf{K}]\phi$ is an inner theorem of $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ for any $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$. However, does knowledge-wh also have positive introspection? To put it more formally, is $[\mathsf{K}_{\mathtt{wh}}]^x \phi \to [\mathsf{K}][\mathsf{K}_{\mathtt{wh}}]^x \phi$ an inner theorem of $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$? The answer is as follow:

Proposition 5.1 $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]} \vdash [\mathsf{K}^{\mathtt{MS}}]^x \phi \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}]^x \phi$*, but* $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]} \nvdash [\mathsf{K}_{\mathtt{wh}}]^x \phi \to [\mathsf{K}][\mathsf{K}_{\mathtt{wh}}]^x \phi$ *when* $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$.

The underlying reason for the failure of positive introspection in $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]}$, $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}$ and $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ is similar. Essentially, this is because these accounts may involve true beliefs (the MS-true belief in $[\mathsf{tB}^{\mathtt{MS}}]^x\phi$ or $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$, or a true belief required by the FS condition in $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ or $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$), but positive introspection requires *knowledge* rather than mere true belief, while the latter in general does not imply the former in an $\mathbf{S4.2}$ system.

The following proposition helps us make this point clear on the formal level. Note that in the formulation of (a part of) the following proposition, we will also need the identity relation $\approx$ and the logic $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\approx}$ which involves the axioms for $\approx$.

Proposition 5.2 *We have the following identities between logics:*

$$\begin{array}{lcl}
\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]} \oplus [\mathsf{tB}^{\mathtt{MS}}]^x\phi \to [\mathsf{K}][\mathsf{tB}^{\mathtt{MS}}]^x\phi & = & \mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]} \oplus [\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi \\
\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]} \oplus [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}][\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi & = & \mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]} \oplus [\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi \\
\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi & = & \mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)
\end{array}$$

In other words, under our $\mathbf{S4.2}$ setting for propositional knowledge, requiring $[\mathsf{tB}^{\mathtt{MS}}]^x\phi$ and $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ to satisfy positive introspection is in effect the same as requiring true belief to imply knowledge. The case for $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$, on the other hand, is a bit more complex: when $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ satisfies positive introspection, either true belief implies knowledge, or there is at most one element in the domain (in which case the notion of FS is clearly trivialized).

A similar phenomenon also appears in the case of the formula $[\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{K}]\phi$. Intuitively, the formula says that knowledge-wh offers the agent a way to obtain propositional knowledge: for example, if we interpret $[\mathsf{K}_{\mathtt{wh}}]$ in terms of knowledge-how, then the formula says that if an agent knows how to achieve ϕ, then she also knows how to make herself know that ϕ. In fact, Proposition 5.1 and 5.2 still hold after we substitute every occurrences of $[\mathsf{K}][\mathsf{K}_{\mathtt{wh}}]^x\phi$ in these propositions with $[\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{K}]\phi$, since $[\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{K}]\phi \leftrightarrow [\mathsf{K}][\mathsf{K}_{\mathtt{wh}}]^x\phi$ is in fact an inner theorem in $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ for all $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$.

A more interesting difference among the different accounts of knowledge-wh concerns the *monotonicity* of knowledge-wh. We say our notion of knowledge-wh is *monotone* if the following rule is admissible in the corresponding logic:

$$\mathtt{MONO} \quad \frac{\vdash \phi \to \psi}{\vdash [\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}_{\mathtt{wh}}]^x\psi}$$

The rule says that if ψ follows logically from ϕ, then if an agent has knowledge-wh of ϕ, then she automatically also has knowledge-wh of ψ.

Note that in order for this to hold, we need to assume that the agent we consider is logically omniscient; and we have indeed assumed so in our underlying logic for propositional logic, $\mathbf{S4.2}^{[\mathsf{K}]}$. However, even such a logically omniscient agent still may *not* have a monotone notion of knowledge-wh, when FS is involved in our account of knowledge-wh.

The propositions below show how FS influences the monotonicity of knowledge-wh. (Note that we need the identity relation $\approx$ to formulate Proposition 5.4.)

Proposition 5.3 $\mathtt{MONO}$ *is admissible in* $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}]}$ *and* $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}]}$*, but inadmissible in* $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}$ *and* $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$.

Proposition 5.4 *The following equivalences holds:*

$$\begin{array}{lcl}
\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus \mathtt{MONO} & = & \mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \to \phi) \\
\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus \mathtt{MONO} & = & \mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \to \phi)
\end{array}$$

As we can see, FS corrupts the monotonicity of knowledge-wh. In fact, as it is shown in Proposition 5.4, if we force $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ and $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ to be monotone, then either the agent can have no false belief at all, or there is only one element in the domain of the model which characterizes her knowledge and belief - in both cases, the notion of FS is completely trivialized. In this sense, we may say that FS is incompatible with the monotonicity of knowledge-wh in quite an essential way: in order to retain the monotonicity of knowledge-wh, we have to give up FS completely.

5.2 Commonalities

As we have seen, different accounts of knowledge-wh behave rather differently when interacting with propositional knowledge. However, when interacting with *belief*, their behaviors are much more similar.

For example, the following proposition shows some inner theorems shared by all the logics presented above:[8]

Proposition 5.5 *For all* $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$*, the following are* $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$*-theorems:*

$$\begin{array}{llll} \textit{(i)} & [\mathsf{B}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{B}]\phi & \textit{(ii)} & \neg[\mathsf{B}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\neg[\mathsf{B}]\phi \\ \textit{(iii)} & [\mathsf{B}]\phi[y/x] \to [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x\phi & \textit{(iv)} & [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x\phi \vee [\mathsf{B}]\neg[\mathsf{K}_{\mathtt{wh}}]^x\phi \end{array}$$

If we interpret $[\mathsf{K}_{\mathtt{wh}}]^x\phi$ in terms of knowledge-how, then (i) and (ii) say that if an agent believes / does not believe that some certain y is a way to achieve ϕ, then she knows how to make herself believe / not believe that ϕ; (iii) says that if the agent believes that some y is a way to achieve ϕ, then she also believes that she knows how to achieve ϕ; and (iv) says that an agent is "confident" concerning her own epistemic state: for any ϕ, she either believes that she knows how to ϕ, or believes that she does not knows how to ϕ. Note that in $\mathbf{S4.2}^{[\mathsf{K}]}$, we also have the interaction principles $[\mathsf{B}]\phi \to [\mathsf{K}][\mathsf{B}]\phi$, $\neg[\mathsf{B}]\phi \to [\mathsf{K}]\neg[\mathsf{B}]\phi$ and $[\mathsf{B}]\phi \to [\mathsf{B}][\mathsf{K}]\phi$ and $[\mathsf{B}][\mathsf{K}]\phi \vee [\mathsf{B}]\neg[\mathsf{K}]\phi$ between propositional knowledge and belief; hence, we may say that when interacting with propositional belief (rather than knowledge), our accounts of knowledge-wh show more aspects that resemble propositional knowledge.

Also note that from (i) and (iii), we can deduce the following two formulas, respectively:

$$\text{(v)} \quad [\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{B}]\phi \qquad \text{(vi)} \quad [\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x\phi$$

As we can see, though $[\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}_{\mathtt{wh}}]^x[\mathsf{K}]\phi$ and $[\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}][\mathsf{K}_{\mathtt{wh}}]^x\phi$ cannot be deduced in $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ when $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$, when the operator $[\mathsf{K}]$ is relaxed to $[\mathsf{B}]$, we obtain (v) and (vi), which are inner theorems of $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}$ for all $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$.

Another interesting commonality shared by all our logics (which also has to do with the interaction between knowledge-wh and belief) concerns what logic of knowledge-wh our agent believes.

In section 5.1, we have already shown some complexities in the reasoning about knowledge-wh: for example, concerning positive introspection and monotonicity, different accounts yield different behaviors of knowledge-wh. These complexities, however, only appear when *we* reason about the knowledge-wh of an agent from an external perspective; when *the agent herself* reasons about her own knowledge-wh from within, all such complexities evaporate.

To put this point more rigidly, we introduce the following notion:

Definition 5.6 *For any logic* $\mathbf{L}$*, let* $\mathbf{L}_{\mathsf{B}} = \{\phi \mid [\mathsf{B}]\phi \in \mathbf{L}\}$.

Intuitively, for a logic $\mathbf{L}$, $\mathbf{L}_{\mathsf{B}}$ collects all the formulas which $\mathbf{L}$ says that an agent believes. In this sense, if $\mathbf{L}$ characterizes the epistemic states of an agent, then $\mathbf{L}_{\mathsf{B}}$ characterizes the epistemic logic believed by this agent.

Then, with the help of this new notation, we can formulate the following theorem:

[8] Note that (iii) in the proposition below is in fact an *axiom* in $\mathbf{S4.2}^{[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]}$ and $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$.

Proposition 5.7 *For all* $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$*,* $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\mathsf{B}}$ *can be axiomatized by the following system:*

$$\begin{array}{l|l}
\mathbf{S5}^{[\mathsf{K}]} & \textit{All axioms and rules of an } \mathbf{S5} \textit{ system for } [\mathsf{K}] \\
\mathtt{KtoK_{wh}} & [\mathsf{K}]\phi[y/x] \to [\mathsf{K}_{\mathtt{wh}}]^x\phi \\
\mathtt{K_{wh}toK}^0 & \dfrac{\vdash [\mathsf{K}]\phi \to \psi}{\vdash [\mathsf{K}]^x\phi \to \psi} \ \textit{(where } x \notin FV(\psi)\textit{)}
\end{array}$$

It is also not hard to check that this system is equivalent to the system SMLMSK presented in [20], a system in the language $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}])$ which is sound and strongly complete w.r.t. the class of $S5$-models.

Hence, conceptually, the above theorem says that no matter which account of knowledge-wh we choose, it makes no difference for our agent: the agent always believes that her knowledge-wh behaves in exactly the same way as MS-knowledge, and the logic for the underlying propositional knowledge is as strong as **S5**. In such a logic, of course knowledge-wh is monotone and satisfies positive introspection; moreover, it even satisfies *negative introspection*: $\neg[\mathsf{K}_{\mathtt{wh}}]^x\phi \to [\mathsf{K}]\neg[\mathsf{K}_{\mathtt{wh}}]^x\phi$ is in $\mathbf{S4.2}^{[\mathsf{K}_{\mathtt{wh}}]}_{\mathsf{B}}$ for all $[\mathsf{K}_{\mathtt{wh}}] \in \{[\mathsf{tB}^{\mathtt{MS}}], [\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}], [\mathsf{K}^{\mathtt{MS}}], [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]\}$. On the other hand, all the subtle differences among the different accounts of knowledge-wh, generated from the gap between mere true belief and knowledge, as well as the peculiar behavior of the FS condition, are all invisible for the agent in question.

6 Conclusion

In this paper, we studied four bundled operators: $[\mathsf{tB}^{\mathtt{MS}}]$, $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]$, $[\mathsf{K}^{\mathtt{MS}}]$ and $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]$, which correspond to the four different accounts of knowledge-wh. We axiomatized the logics which take them (as well as $[\mathsf{K}]$) as primitive modalities on the class of $S4.2$-constant-domain models, and compared the ways we reason about knowledge-wh in different logics.

There many potential future works that can be done based on our work.

For example, we can further study the four bundled operators introduced in this paper. We have only studied their behavior on $S4.2$-models, which characterize knowledge and belief in a highly idealized way; our study of the obtained logics is also far from complete. Hence, it seems interesting to study the logics obtained in this paper in greater detail, or to study the behavior of the bundled operators on other reasonable models for knowledge and belief (of course, we need not confine ourselves to Kripke models). This may offer us a deeper understanding of the different accounts of knowledge-wh, and may eventually help us decide which account is indeed the right one.

Moreover, the kind of step-by-step proof method applied in this paper can be generalized to study other complex epistemic notion. For example, there are cases where it is better to understand knowledge-wh in terms of mention-all knowledge, and there are also various competing accounts of these kinds of knowledge-wh, e.g. the *weakly exhaustive* reading (first proposed in [8]), the *strongly exhaustive* reading (first proposed in [6]), and the *intermediately exhaustive* reading (first raised, but soon rejected, in [6], and later proposed again in [15]), which can be formalized as $\forall x(\phi(x) \to [\mathsf{K}]\phi(x))$, $\forall x([\mathsf{K}]\phi(x) \vee [\mathsf{K}]\neg\phi(x))$ and $\forall x(\phi(x) \to [\mathsf{K}]\phi(x)) \wedge \forall x([\mathsf{B}]\phi(x) \to \phi(x))$, respectively. Using the technique developed in this paper, we can easily pack these complex notions into bundled operators, and study their behavior.

Speaking on a more general level, the step-by-step method used in this paper can at least be generalized to any logic equipped with a set of ordinary modal operators $\{\Box_a\}_{a\in\tau}$ plus a set of bundled operators of the form $\blacksquare^x\phi := \exists x\alpha[\phi/p] \wedge \forall x\beta[\phi/p]$, where α and β are propositional modal formulas containing only one propositional symbol p, boolean connectives and operators in $\{\Box_a\}_{a\in\tau}$. Our trick works no matter how complicated the structures of α and β are, so a great deal of complex first-order modal notions can be handled in this way.

References

[1] J. Adam Carter & Duncan Pritchard (2015): *Knowledge-How and Epistemic Luck*. Noûs 49(3), pp. 440–453, doi:10.1111/nous.12054.

[2] Yuri Cath (2015): *Revisionary Intellectualism and Gettier*. Philosophical Studies 172(1), pp. 7–27, doi:10.1007/s11098-013-0263-y.

[3] Jie Fan, Yanjing Wang & Hans van Ditmarsch (2015): *Contingency and Knowing Whether*. Rev. Symb. Log. 8(1), pp. 75–107, doi:10.1017/S1755020314000343.

[4] Raul Fervari, Andreas Herzig, Yanjun Li & Yanjing Wang (2017): *Strategically knowing how*. In Carles Sierra, editor: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, ijcai.org, pp. 1031–1038, doi:10.24963/ijcai.2017/143.

[5] B. R. George (2013): *Knowing-'Wh', Mention-Some Readings, and Non-Reducibility*. Thought: A Journal of Philosophy 2(2), pp. 166–177, doi:10.1002/tht3.88.

[6] Jeroen Groenendijk & Martin Stokhof (1982): *Semantic analysis of "wh"-complements*. Linguistics and Philosophy, pp. 175–233, doi:10.1007/BF00351052.

[7] Keith Harris (2019): *Knowledge-How and False Belief*. Synthese 198(2), pp. 1845–1861, doi:10.1007/s11229-019-02172-2.

[8] Lauri Karttunen (1977): *Syntax and Semantics of Questions*. Linguistics and Philosophy 1(1), pp. 3–44, doi:10.1007/bf00351935.

[9] Wolfgang Lenzen (1978): *Recent Work in Epistemic Logic*. Acta Philosophica Fennica 30, pp. 1–219.

[10] Wolfgang Lenzen (1979): *Epistemologische Betrachtungen zu [S4, S5]*. Erkenntnis 14(1), pp. 33–56, doi:10.1007/BF00205012.

[11] Yanjun Li & Yanjing Wang (2017): *Achieving While Maintaining: - A Logic of Knowing How with Intermediate Constraints*. In Sujata Ghosh & Sanjiva Prasad, editors: Logic and Its Applications - 7th Indian Conference, ICLA 2017, Kanpur, India, January 5-7, 2017, Proceedings, Lecture Notes in Computer Science 10119, Springer, pp. 154–167, doi:10.1007/978-3-662-54069-5_12.

[12] Pavel Naumov & Jia Tao (2018): *Together We Know How to Achieve: An Epistemic Logic of Know-How*. Artificial Intelligence 262(C), pp. 279–300, doi:10.1016/j.artint.2018.06.007.

[13] Anantha Padmanabha, R. Ramanujam & Yanjing Wang (2018): *Bundled Fragments of First-Order Modal Logic: (Un)Decidability*. In Sumit Ganguly & Paritosh K. Pandya, editors: 38th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2018, December 11-13, 2018, Ahmedabad, India, LIPIcs 122, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 43:1–43:20, doi:10.4230/LIPIcs.FSTTCS.2018.43.

[14] Jonathan Phillips & B. R. George (2018): *Knowledge Wh and False Beliefs: Experimental Investigations*. Journal of Semantics 35(3), pp. 467–494, doi:10.1093/semant/ffy004.

[15] Benjamin Spector (2005): *Exhaustive interpretations: What to say and what not to say*. In: Unpublished paper presented at the LSA workshop on Context and Content.

[16] Robert Stalnaker (2006): *On Logics of Knowledge and Belief*. Philosophical Studies 128(1), pp. 169–199, doi:10.1007/s11098-005-4062-y.

[17] Jason Stanley & Timothy Willlamson (2001): *Knowing How*. Journal of Philosophy 98(8), pp. 411–444, doi:10.2307/2678403.

[18] Xun Wang (2021): *Completeness Theorems for $\exists\Box$-Fragment of First-Order Modal Logic*. In Sujata Ghosh & Thomas Icard, editors: Logic, Rationality, and Interaction: 8th International Workshop, Lori 2021, Xi'an, China, October 16-18, 2021, Proceedings, Springer Verlag, pp. 246–258, doi:10.1007/978-3-030-88708-7_-20.

[19] Yanjing Wang (2015): *A logic of knowing how*. In: *Logic, Rationality, and Interaction: 5th International Workshop, LORI 2015, Taipei, Taiwan, October 28-30, 2015. Proceedings 5*, Springer, pp. 392–405, doi:10.1007/978-3-662-48561-3_32.

[20] Yanjing Wang (2017): *A New Modal Framework for Epistemic Logic*. In Jérôme Lang, editor: *Proceedings Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2017, Liverpool, UK, 24-26 July 2017*, *EPTCS* 251, pp. 515–534, doi:10.4204/EPTCS.251.38.

[21] Yanjing Wang (2018): *Beyond Knowing That: A New Generation of Epistemic Logics*. In Hans van Ditmarsch & Gabriel Sandu, editors: *Jaakko Hintikka on Knowledge and Game Theoretical Semantics*, Springer, pp. 499–533, doi:10.1007/978-3-319-62864-6_21.

[22] Yanjing Wang (2018): *A logic of goal-directed knowing how*. *Synth.* 195(10), pp. 4419–4439, doi:10.1007/s11229-016-1272-0.

[23] Yimei Xiang (2016): *Complete and true: A uniform analysis for mention some and mention all*. In: *Proceedings of Sinn und Bedeutung*, 20, pp. 815–832.

[24] Chao Xu, Yanjing Wang & Thomas Studer (2021): *A logic of knowing why*. *Synthese* 198(2), pp. 1259–1285, doi:10.1007/s11229-019-02104-0.

[25] Zhouhang Zhou (2018): *An epistemic logic of mention-all*. Undergraduate thesis, Peking University.

Acknowledgement

I owe the very idea of studying knowledge-wh with FS using bundled operators to Yanjing Wang, and I would also like to thank him for his insightful advice on this paper.

I would also like to thank Yimei Xiang for introducing works concerning MS-knowledge and FS in linguistics to me, and Xun Wang for discussing matters concerning the $\exists\Box$-bundled fragment with me. I should also thank three anonymous reviewers from TARK 2023 for their helpful advice on the writing of this paper, and for identifying the typos and grammar mistakes in the paper.

Finally, I would like to thank the support of NSSF grant 19BZX135.

A Appendix

In the appendix, we show how to prove theorem 4.1, Proposition 5.2 and Proposition 5.4.

First, we consider theorem 4.1. We only prove the case for $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$, since the other cases can be proved in a similar way. Moreover, for most of the time, we will be working in the language $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, since our proof can easily be generalized to the case of $\mathscr{L}_{\approx}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$ with the help of some slight modifications. We will demonstrate how to do so along the proof.

First, we check that the soundness result holds.

Proposition A.1 $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ *is sound w.r.t. the class of S4.2-constant-domain models.*

PROOF. We only prove that $\mathtt{BtoBK}_{\mathtt{wh}}$ is valid on the class of $S4.2$-constant-domain models, and $\mathtt{FS\&KtoK}_{\mathtt{wh}}$ preserves validity on such models.

For $\mathtt{BtoBK}_{\mathtt{wh}}$:

Let $\mathscr{M} = (W, R, D, \rho)$ be a $S4.2$-model, let $w \in W$ be arbitrary, and let σ be an arbitrary assignment. Assume that $\mathscr{M}, w, \sigma \vDash [\mathsf{B}]\phi[y/x]$. Then, for all $v \in W$ s.t. $wR_{\mathsf{B}}v$, $\mathscr{M}, v, \sigma \vDash \phi[y/x]$.

Then, let $v \in W$ be arbitrary, and assume that $wR_{\mathsf{B}}v$.

We first show that $\mathscr{M}, v, \sigma \vDash [\mathsf{K}]\phi[y/x]$. This is clear, since for all $u \in W$ s.t. vRu, it is easy to check that $wR_{\mathsf{B}}u$, and thus $\mathscr{M}, u, \sigma \vDash \phi[y/x]$.

Then, we show that for all $a \in D$, $\mathscr{M}, v, \sigma[x \mapsto a] \vDash [\mathsf{B}]\phi \to \phi$. This is also clear: since (W, R_{B}) is $KD45$, v is R_{B}-reflexive.

Hence, it is easy to see that $\mathscr{M}, v, \sigma \vDash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$, and thus $\mathscr{M}, w, \sigma \vDash [\mathsf{B}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$.

For $\mathtt{FS\&KtoK_{wh}}$:

Let $\mathscr{M} = (W, D, R, \rho)$ be an arbitrary $S4.2$-model, and assume that $\psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi \to \phi))\cdots)$ is valid on $\mathscr{M}, w$, where n is an arbitrary natural number; and let x be an variable s.t. $x \notin \bigcup_{i \le n} FV(\psi_i)$.

Then, let σ be an arbitrary assignment, and suppose (towards a contradiction) that $\mathscr{M}, w, \sigma \nvDash \psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{K}]\phi[y/x] \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi))\cdots)$. Then, there is some $w_0, w_1, ..., w_n \in W$, s.t. $w = w_0 R w_1 R \cdots R w_n$, $\mathscr{M}, w_i, \sigma \vDash \psi_i$ for all $i \le n$, and $\mathscr{M}, w_n, \sigma \nvDash [\mathsf{K}]\phi[y/x] \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$. By the validity of $\psi_0 \to [\mathsf{K}](\psi_1 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi \to \phi))\cdots)$, and since $x \notin \bigcup_{i \le n} FV(\psi_i)$, for all $a \in D$, $\mathscr{M}, w_n, \sigma[x \mapsto a] \vDash [\mathsf{B}]\phi \to \phi$. But then, since $\mathscr{M}, w_n, \sigma \vDash [\mathsf{K}]\phi[y/x]$, $\mathscr{M}, w_n, \sigma[x \mapsto \sigma(y)] \vDash [\mathsf{K}]\phi$, and thus it should follow that $\mathscr{M}, w_n, \sigma \vDash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$, causing a contradiction. □

It is also not hard to check that $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ has the following inner theorems, which will be used in our completeness proof.

$\mathtt{NBK_{wh}toBNK_{wh}}$	$\langle\mathsf{B}\rangle[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{B}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$
$\mathtt{BK_{wh}toK_{wh}B}$	$[\mathsf{B}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x[\mathsf{B}]\phi$
$\mathtt{R}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$	$[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \leftrightarrow [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y\phi[y/x]$ (where y does not appear in ϕ)

Now, we are ready to prove the completeness theorem.

As preparation, we first define the language $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, which is obtained by adding countably many new variables to $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$. We use $\mathbf{X}^+$ to denote the set of variables of $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$.

Then, we use a step-by-step method to prove the completeness theorem. We first define the notion of a *network*. Note that when constructing such networks, the states will all be taken from a set of states $\{w_i \mid i \in \omega\}$, which we fix in advance.

Definition A.2 *A* network *is a triple $\mathscr{N} = (W, R, \nu)$, where*

- *$\{w_0\} \subseteq W \subseteq \{w_i \mid i \in \omega\}$;*
- *$R \subseteq W^2$, and (W, R) forms a tree where w_0 is the root;*
- *ν assigns each element in W a set of $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$-formulas.*

We also define the following two properties for the formula sets in a network:

Definition A.3 *($\mathtt{MS}$-property) An $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$-formula set Δ has* $\mathtt{MS}$-property, *iff for all $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$ and $x \in \mathbf{X}^+$, if $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \in \Delta$, then there is some $y \in \mathbf{X}^+$ s.t. $[\mathsf{K}]\phi[y/x] \in \Delta$.*

Definition A.4 *($\mathtt{FS}$-property) An $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$-formula set Δ has* $\mathtt{FS}$-property, *iff for all $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$ and $x, y \in \mathbf{X}^+$, if $\neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi, [\mathsf{K}]\phi[y/x] \in \Delta$, then there is some $z \in \mathbf{X}^+$ s.t. $([\mathsf{B}]\phi \wedge \neg\phi)[z/x] \in \Delta$.*

Then, we define the notion of coherence and saturation for networks:

Definition A.5 *A network $\mathscr{N} = (W, R, \nu)$ is* coherent, *iff the following conditions are satisfied:*

(i) W is finite;

(ii) For all $w \in W$, $\nu(w)$ is $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$-consistent; and for all $w \in W \setminus \{w_0\}$, $\nu(w)$ is finite;

(iii) For all $w, v \in W$ s.t. wRv, there is some ψ s.t. $\vdash \psi \leftrightarrow \bigwedge \nu(v)$ and $\langle\mathsf{K}\rangle\psi \in \nu(w)$;

(iv) There are countably many variables in $\mathbf{X}^+$ which do not appear in $\nu(w)$ for any $w \in W$.

Definition A.6 *A network $\mathscr{N} = (W, R, \nu)$ is* saturated*, iff for all $w \in W$ and $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, the following holds:*

(i) $\nu(w)$ *is a MCS of $\mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$-formulas;*

(ii) If $[\mathsf{K}]\phi \in \nu(w)$, then for all $v \in W$ s.t. wRv, $\phi \in \nu(v)$;

(iii) If $\langle\mathsf{K}\rangle\phi \in \nu(w)$, then there is some $v \in W$ s.t. wRv and $\phi \in \nu(v)$;

(iv) $\nu(w)$ *has the* MS*-property;*

(v) $\nu(w)$ *has the* FS*-property.*

Then, corresponding to the requirements of saturation, we also introduce the following notion of *defects*:

Definition A.7 *The possible kinds of* defects *we may find on a state on a $w \in W$ in a network $\mathscr{N} = (W, R, \nu)$ are as follow:*

(d1) $\phi \notin \nu(w)$ *and* $\neg\phi \notin \nu(w)$

(d2) $[\mathsf{K}]\phi \in \nu(w)$*, but there is some $v \in W$ s.t. wRv and $\phi \notin \nu(v)$*

(d3) $\langle\mathsf{K}\rangle\phi \in \nu(w)$*, but there is no $v \in W$ s.t. wRv and $\phi \in \nu(v)$*

(d4) $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \in \nu(w)$*, but there is no $y \in \mathbf{X}^+$ s.t. $[\mathsf{K}]\phi[y/x] \in \nu(w)$*

(d5) $\neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi, [\mathsf{K}]\phi[y/x] \in \nu(w)$*, but there is no $z \in \mathbf{X}^+$ s.t. $([\mathsf{B}]\phi \wedge \neg\phi)[z/x] \in \nu(w)$*

where $w \in \{w_i \mid i \in \omega\}$, $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$ and $x \in \mathbf{X}^+$.

Then, we prove the *repair lemma*, which shows how to repair defects in a coherent network, while maintaining its coherence.

Lemma A.8 *(Repair lemma) For any coherent network $\mathscr{N} = (W, R, \nu)$ and any defect (d) of $\mathscr{N}$, then there is a coherent network $\mathscr{N}' = (W', R', \nu')$ s.t. $W \subseteq W'$, $R \subseteq R'$, $\nu(w) \subseteq \nu'(w)$ for all $w \in W$, and $\mathscr{N}'$ does not has (d).*

PROOF. Let $\mathscr{N} = (W, R, \nu)$ be a coherent network, and assume that $\mathscr{N}$ has a defect (d) for some $w_m \in W$ and $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$.

Since (W, R) forms a tree where w_0 is the root, there is a unique path $w_0 = v_0 R v_1 R \cdots R v_n = w_m$ in $\mathscr{N}$ for some $n \in \omega$. Then, since $\mathscr{N}$ is coherent, for all $1 \leq i \leq n$, let ψ_i stand for the formula s.t. $\vdash \psi_i \leftrightarrow \bigwedge \nu(v_i)$ and $\langle\mathsf{K}\rangle\psi_i \in \nu(v_{i-1})$. Then, it is easy to see that

$$\nu(v_0) \vdash \langle\mathsf{K}\rangle(\psi_1 \wedge \langle\mathsf{K}\rangle(\psi_2 \wedge \cdots \langle\mathsf{K}\rangle(\psi_{n-1} \wedge \langle\mathsf{K}\rangle\psi_n)\cdots))$$

We then consider five cases.

Case 1: (d) is of the kind $(d1)$. That is, $\phi \notin \nu(v_n)$ and $\neg\phi \notin \nu(v_n)$. Then, it is easy to check that

$$\nu(v_0) \vdash \langle\mathsf{K}\rangle(\psi_1 \wedge \langle\mathsf{K}\rangle(\psi_2 \wedge \cdots \langle\mathsf{K}\rangle(\psi_n \wedge \phi)\cdots)) \vee \langle\mathsf{K}\rangle(\psi_1 \wedge \langle\mathsf{K}\rangle(\psi_2 \wedge \cdots \langle\mathsf{K}\rangle(\psi_n \wedge \neg\phi)\cdots))$$

Then, at least one of the disjuncts is consistent with $\nu(v_0) = \nu(w_0)$. We only consider the case where the former is consistent with $\nu(w_0)$, since the other case is similar. In this case, let

$$\begin{aligned}\nu' = &\{(w, \nu(w)) \mid w \neq v_i \text{ for all } i \leq n\}\\ &\cup \{(v_n, \nu(v_n) \cup \{\phi\})\}\\ &\cup \{(v_i, \nu(v_i) \cup \{\langle\mathsf{K}\rangle(\psi_{i+1} \wedge \cdots \langle\mathsf{K}\rangle(\psi_n \wedge \phi)\cdots)\}) \mid i < n\}\end{aligned}$$

and let $\mathscr{N}' = \langle W, R, \nu' \rangle$. It is then easy to check that $\mathscr{N}'$ is coherent, and does not have the defect (d).

Case 2: (d) is of the kind $(d2)$. That is, $[\mathsf{K}]\phi \in \nu(v_n)$, but there is some $u \in W$ s.t. $v_n R u$ and $\phi \notin \nu(u)$. Since $\mathscr{N}$ is coherent, there is some ψ_u s.t. $\vdash \psi_u \leftrightarrow \bigwedge \nu(u)$ and $\langle \mathsf{K} \rangle \psi_u \in \nu(v_n)$. Hence, it is easy to check that

$$\nu(v_0) \vdash \langle \mathsf{K} \rangle(\psi_1 \wedge \langle \mathsf{K} \rangle(\psi_2 \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge \langle \mathsf{K} \rangle(\psi_u \wedge \phi)) \cdots))$$

Then, let

$$\begin{aligned} \nu' = &\{(w, \nu(w)) \mid w \neq v_i \text{ for all } i \leq n\} \\ &\cup \{(u, \nu(u) \cup \{\phi\})\} \\ &\cup \{(v_i, \nu(v_i) \cup \{\langle \mathsf{K} \rangle(\psi_{i+1} \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge \phi) \cdots)\}) \mid i \leq n\} \end{aligned}$$

It is easy to check that $\mathscr{N}' = (W, R, \nu')$ is still coherent, and does not have the defect (d).

Case 3: (d) is of the kind $(d3)$. That is, $\langle \mathsf{K} \rangle \phi \in \nu(v_n)$, but there is no $u \in W$ s.t. $v_n R u$ and $\phi \in \nu(u)$. Since W is finite, there is some $\{w_i \mid i \in \omega\} \setminus W \neq \emptyset$. Then, let w_k be the element in $\{w_i \mid i \in \omega\} \setminus W$ with the least index number, and let $W' = W \cup \{w_k\}$, $R' = R \cup \{(v_n, w_k)\}$, and $\nu' = \nu \cup \{(w_k, \{\phi\})\}$. It is easy to check that $\mathscr{N} = (W', R', \nu')$ is coherent, but does not have (d).

Case 4: (d) is of the kind $(d4)$. That is, $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x \phi \in \nu(v_n)$, but there is no $y \in \mathbf{X}^+$ s.t. $[\mathsf{K}]\phi[y/x] \in \nu(v_n)$. Then, let $y \in \mathbf{X}^+$ be a variable that does not appear in $\nu(w)$ for any $w \in W$, and suppose (towards a contradiction) that

$$\nu(v_0) \vdash [\mathsf{K}](\psi_1 \to [\mathsf{K}](\psi_2 \to \cdots [\mathsf{K}](\psi_n \to \neg[\mathsf{K}]\phi[y/x]) \cdots))$$

Then, by $\mathtt{K_{wh}toK}$ (and $\mathtt{R}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$), we have

$$\nu(v_0) \vdash [\mathsf{K}](\psi_1 \to [\mathsf{K}](\psi_2 \to \cdots [\mathsf{K}](\psi_n \to \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x \phi) \cdots))$$

which contradicts the fact that $\nu(v_0) = \nu(w_0)$ is consistent. Hence, $\langle \mathsf{K} \rangle(\psi_1 \wedge \langle \mathsf{K} \rangle(\psi_2 \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge [\mathsf{K}]\phi[y/x]) \cdots))$ is consistent with $\nu(v_0) = \nu(w_0)$. Hence, let

$$\begin{aligned} \nu' = &\{(w, \nu(w)) \mid w \neq v_i \text{ for all } i \leq n\} \\ &\cup \{(v_n, \nu(v_n) \cup \{[\mathsf{K}]\phi[y/x]\})\} \\ &\cup \{(v_i, \nu(v_i) \cup \{\langle \mathsf{K} \rangle(\psi_{i+1} \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge [\mathsf{K}]\phi[y/x]) \cdots)\}) \mid i < n\} \end{aligned}$$

It is easy to check that $\mathscr{N}' = (W, R, \nu)$ is still coherent, and does not have the defect (d).

Case 5: (d) is of the kind $(d5)$. That is, $\neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x \phi \in \nu(v_n)$ and $[\mathsf{K}]\phi[y/x] \in \nu(v_n)$, but there is no $z \in \mathbf{X}^+$ s.t. $([\mathsf{B}]\phi \wedge \neg\phi)[z/x] \in \nu(v_n)$. Then, let $z \in \mathbf{X}^+$ be a variable that does not appear in $\nu(w)$ for any $w \in W$, and suppose (towards a contradiction) that

$$\nu(v_0) \vdash [\mathsf{K}](\psi_1 \to [\mathsf{K}](\psi_2 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{B}]\phi[z/x] \to \phi[z/x])) \cdots))$$

Then, by $\mathtt{FS\&KtoK_{wh}}$ (and $\mathtt{R}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$), we have

$$\nu(v_0) \vdash [\mathsf{K}](\psi_1 \to [\mathsf{K}](\psi_2 \to \cdots [\mathsf{K}](\psi_n \to ([\mathsf{K}]\phi[y/x] \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x \phi)) \cdots))$$

which contradicts the fact that $\nu(v_0) = \nu(w_0)$ is consistent. Hence, $\langle \mathsf{K} \rangle(\psi_1 \wedge \langle \mathsf{K} \rangle(\psi_2 \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge ([\mathsf{B}]\phi \wedge \neg\phi)[z/x]) \cdots))$ is consistent with $\nu(v_0) = \nu(w_0)$. Hence, let

$$\begin{aligned}\nu' =& \{(w, \nu(w)) \mid w \neq v_i \text{ for all } i \leq n\} \\ & \cup \{(v_n, \nu(v_n) \cup \{([\mathsf{B}]\phi \wedge \neg\phi)[z/x]\})\} \\ & \cup \{(v_i, \nu(v_i) \cup \{\langle \mathsf{K} \rangle(\psi_{i+1} \wedge \cdots \langle \mathsf{K} \rangle(\psi_n \wedge ([\mathsf{B}]\phi \wedge \neg\phi)[z/x]) \cdots)\}) \mid i < n\}\end{aligned}$$

It is easy to check that $\mathcal{N}' = (W, R, \nu)$ is still coherent, and does not have the defect (d). □

Then, we can easily show that every coherent network can be extended into a saturated network.

Lemma A.9 *For any coherent network $\mathcal{N} = (W, R, \nu)$, there exists a saturated network $\mathcal{N}' = (W', R', \nu')$ s.t. $W \subseteq W'$, $R \subseteq R'$ and $\nu(w) \subseteq \nu'(w)$ for all $w \in W$.*

PROOF. Let $\mathcal{N} = (W, R, \nu)$ be a coherent network.

It is not hard to see that there are only countably many possible defects. Hence, we can enumerate them as $(d)_1$, $(d)_2$, $(d)_3$, ...

Then, we define a countable sequence of networks $\mathcal{N}_i = (W_i, R_i, \nu_i)$ $(i \in \omega)$ recursively as follow:

- $\mathcal{N}_0 = \mathcal{N}$;
- Given a coherent network $\mathcal{N}_k$, let $(d)_m$ be the defect of $\mathcal{N}_k$ with the least index number (note that according to our definition of coherence, $\mathcal{N}_k$ necessarily has defects), and let $\mathcal{N}_{k+1} = (W_{k+1}, R_{k+1}, \nu_{k+1})$ be a coherent network which does not has $(d)_m$, and also satisfies that $W_k \subseteq W_{k+1}$, $R_k \subseteq R_{k+1}$, $\nu_k(w) \subseteq \nu_{k+1}(w)$ for all $w \in W_k$. The existence of such a network is guaranteed by lemma A.8.

Then, let $\mathcal{N}' = (W', R', \nu')$, where

- $W' = \bigcup_{i \in \omega} W_i$;
- $R' = \bigcup_{i \in \omega} R_i$;
- For all $w \in W$, $\nu'(w) = \bigcup_{i \geq k} \nu_i(w)$, where k is the least number s.t. $w \in W_k$.

It is not hard to see that $\mathcal{N}'$ is a saturated network s.t. $W_0 \subseteq W$, $R_0 \subseteq R$ and $\nu_0(w) \subseteq \nu(w)$ for all $w \in W_0$. □

Then, we show how to induce a canonical model from a saturated network.

Definition A.10 *Given a saturated network $\mathcal{N} = (W, R, \nu)$, $\mathcal{M}^c_{\mathcal{N}} = (W^c_{\mathcal{N}}, R^c_{\mathcal{N}}, D^c_{\mathcal{N}}, \rho^c_{\mathcal{N}})$ is the model induced from $\mathcal{N}$, where*

- *$W^c_{\mathcal{N}} = \{\nu(w) \mid w \in W\} \cup FC$,*
  where $FC = \{\Theta \mid \Theta$ is a MCS in $\mathcal{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]), \{\phi \mid [\mathsf{B}]\phi \in \nu(w_0)\} \subseteq \Theta\}$;[9]
- *$D^c_{\mathcal{N}} = \mathbf{X}^+$;*
- *$R^c_{\mathcal{N}}$ satisfies that for all $\Delta, \Theta \in W^c$, $\Delta R^c \Theta$ iff for all $\phi \in \mathcal{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, if $[\mathsf{K}]\phi \in \Delta$, then $\phi \in \Theta$;*
- *$\rho^c_{\mathcal{N}}$ satisfies that for all $\Delta \in W^c_{\mathcal{N}}$, $\bar{x} \in (D^c_{\mathcal{N}})^{<\omega}$ and $P \in \mathscr{P}$, $\bar{x} \in \rho^c(P, \Delta)$ iff $P\bar{x} \in \Delta$.*

We may drop the subscript $\mathcal{N}$ when the context is clear.

[9] *FC* stands for *F*inal *C*luster. In fact, we can show that for all $\Delta \in W^c$ and $\Theta \in FC$, $\Delta R^c \Theta$, which justifies our naming.

Remark A.11 *If we are working in the language $\mathscr{L}_{\approx}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, then we let $D^c_{\mathscr{N}} = \{[x] \mid x \in \mathbf{X}^+\}$, where $[x] = \{y \in \mathbf{X}^+ \mid x \approx y \in \nu(w_0)\}$.*

We then show that a model induced from a saturated network is indeed $S4.2$, and also has all the properties we need.

Lemma A.12 *For any saturated network $\mathscr{N}$, $\mathscr{M}^c_{\mathscr{N}}$ satisfies the following:*

(i) $\mathscr{M}^c_{\mathscr{N}}$ is an $S4.2$-model;

(ii) For all $\Delta \in W^c$ and $\langle \mathsf{K} \rangle \phi \in \Delta$, there is some $\Delta' \in W^c$ s.t. $\Delta R^c \Delta'$ and $\phi \in \Delta'$;

(iii) For all $\Delta \in W^c$, Δ has the `MS`*-property and the* `FS`*-property.*

PROOF. Let $\mathscr{M}^c_{\mathscr{N}}$ be an arbitrary model induced from a saturated network $\mathscr{N} = (W, R, \nu)$.

For item (i): By the definition of R^c and the canonicity of $\mathtt{T}^{[\mathsf{K}]}$ and $4^{[\mathsf{K}]}$, it is easy to see that $\mathscr{M}^c_{\mathscr{N}}$ is reflexive and transitive.

We then show that $\mathscr{M}^c_{\mathscr{N}}$ is strongly convergent.

Clearly $FC \neq \emptyset$, since $\langle \mathsf{B} \rangle \top \in \nu(w_0)$.

Then, we show that for all $\Delta \in W^c$ and $\Theta \in FC$, $\Delta R^c \Theta$. Let $\Delta \in W^c$ and $\Theta \in FC$ be arbitrary. We consider two cases:

Case 1: there is some $w \in W$ s.t. $\Delta = \nu(w)$. Let $[\mathsf{K}]\phi \in \nu(w)$ be arbitrary. It is easy to see that $\nu(w_0) R^c \nu(w)$; hence, $\langle \mathsf{K} \rangle [\mathsf{K}] \phi \in \nu(w_0)$, i.e. $[\mathsf{B}]\phi \in \nu(w_0)$. Hence, by definition, $\phi \in \Theta$. Thus, $\nu(w) R^c \Theta$.

Case 2: $\Delta \in FC$. Let $[\mathsf{K}]\phi \in \Delta$ be arbitrary. Then, since $\Delta \in FC$, $\langle \mathsf{B} \rangle [\mathsf{K}] \phi \in \nu(w_0)$, i.e. $[\mathsf{K}]\langle \mathsf{K} \rangle [\mathsf{K}] \phi \in \nu(w_0)$. Then, by $\mathtt{T}^{[\mathsf{K}]}$, $\langle \mathsf{K} \rangle [\mathsf{K}] \phi \in \nu(w_0)$, i.e. $[\mathsf{B}]\phi \in \nu(w_0)$. Hence, $\phi \in \Theta$ and thus, $\Delta R^c \Theta$.

Therefore, $\mathscr{M}^c_{\mathscr{N}}$ is strongly convergent.

For item (ii): Since $\mathscr{N}$ is saturated, we only need prove that for all $\Theta \in FC$ and $\langle \mathsf{K} \rangle \phi \in \Theta$, there is some $\Theta' \in W^c$ s.t. $\Theta R^c \Theta'$ and $\phi \in \Theta'$.

Let $\Theta \in FC$, $\langle \mathsf{K} \rangle \phi \in \Theta$ be arbitrary. Then, since $\Theta \in FC$, $\langle \mathsf{B} \rangle \langle \mathsf{K} \rangle \phi \in \nu(w_0)$, i.e. $[\mathsf{K}]\langle \mathsf{K} \rangle \langle \mathsf{K} \rangle \phi \in \nu(w_0)$. Hence, by $4^{[\mathsf{K}]}$, $[\mathsf{K}]\langle \mathsf{K} \rangle \phi \in \nu(w_0)$, i.e. $\langle \mathsf{B} \rangle \phi \in \nu(w_0)$, and thus, there is some $\Theta' \in FC$ s.t. $\phi \in \Theta'$. Then, as we have already proved, $\Theta R^c \Theta'$.

For item (iii): Again, since $\mathscr{N}$ is saturated, we only need to prove that every $\Theta \in FC$ has the `MS`-property and the `FS`-property.

Let $\Theta \in FC$ and $\phi \in \mathscr{L}^+([\mathsf{K}_{\mathtt{wh}}])$ be arbitrary.

First, assume that $[\mathsf{K}_{\mathtt{wh}}]^x \phi \in \Theta$. Then, $\langle \mathsf{B} \rangle [\mathsf{K}_{\mathtt{wh}}]^x \phi \in \nu(w_0)$, and thus, by $\mathtt{NBK_{wh}toBNK_{wh}}$, $[\mathsf{B}][\mathsf{K}]^x \phi \in \nu(w_0)$. Then, by $\mathtt{BK_{wh}toK_{wh}B}$, $[\mathsf{K}]^x [\mathsf{B}] \phi \in \nu(w_0)$. Then, since $\mathscr{N}$ is saturated, $\nu(w_0)$ has the `MS`-property, and thus there is some $y \in \mathbf{X}^+$ s.t. $[\mathsf{K}][\mathsf{B}]\phi[y/x] \in \nu(w_0)$. Hence, $[\mathsf{B}]\phi[y/x] \in \nu(w_0)$, and thus $[\mathsf{B}][\mathsf{K}]\phi[y/x] \in \nu(w_0)$. Hence, $[\mathsf{K}]\phi[y/x] \in \Theta$.

Next, assume that $[\mathsf{K}_{\mathtt{wh}}]^x \phi \in \Theta$. Then, $\neg[\mathsf{B}][\mathsf{K}_{\mathtt{wh}}]^x \phi \in \nu(w_0)$, and thus for all $y \in \mathbf{X}^+$, $\neg[\mathsf{B}]\phi[y/x] \in \nu(w_0)$ by $\mathtt{BtoBK_{wh}}$. Hence, for all $y \in \mathbf{X}^+$, $[\mathsf{B}]\neg[\mathsf{B}]\phi[y/x] \in \nu(w_0)$, and thus $[\mathsf{B}]\phi[y/x] \notin \Theta$. □

Then, it is routine to prove the truth lemma:

Lemma A.13 *For all $\mathscr{M}^c_{\mathscr{N}}$ induced from a saturated network $\mathscr{N}$, for all $\Delta \in W^c$ and $\phi \in \mathscr{L}^+([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, $\mathscr{M}^c_{\mathscr{N}}, \Delta, \sigma^c \vDash \phi \iff \phi \in \Delta$, where σ^c is the assignment s.t. $\sigma^c(x) = x$ for all $x \in \mathbf{X}^+$.*

Remark A.14 *If we are working in the language $\mathscr{L}_{\approx}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$, then in the formulation of the above lemma, we let σ^c be the assignment s.t. $\sigma^c(x) = [x]$ for all $x \in \mathbf{X}^+$.*

Finally, notice that for any $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$-consistent set Γ of $\mathscr{L}([\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}])$-formulas Γ_0, $(\{w_0\}, \emptyset, \{(w_0, \Gamma_0)\})$ is a coherent network. Hence, it can be extended into a saturated network $\mathscr{N}'$, from which we can induce a canonical model $\mathscr{M}^c_{\mathscr{N}'}$, such that $\mathscr{M}^c_{\mathscr{N}'}, \nu'(w_0), \sigma^c \vDash \Gamma_0$.

Hence, we have the following completeness theorem:

Theorem A.15 $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ *(as well as* $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx}$*) is sound and strongly complete w.r.t. the class of $S4.2$-constant-domain models.*

Then, we consider Proposition 5.2. The cases for $[\mathsf{tB}^{\mathtt{MS}}]^x\phi$ and $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ are relatively easy, since it is easy to see that $x \notin FV(\phi)$, $[\mathsf{tB}^{\mathtt{MS}}]^x\phi$ and $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ are equivalent to $[\mathsf{B}]\phi \wedge \phi$. Hence, we only prove the following proposition here:

Proposition A.16 *The following equivalence holds:*

$$\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \quad = \quad \mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$$

PROOF. We first show that $x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$ can be deduced in $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$. It is easy to check that $\vdash \phi \wedge x \not\approx y \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^z(x \approx z \to \phi[z/x])$, where z is a fresh variable. Then, by positive introspection, $\vdash \phi \wedge x \not\approx y \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^z(x \approx z \to \phi[z/x])$, and by $\mathtt{K_{wh}toFS}$, $\vdash \phi \wedge x \not\approx y \to [\mathsf{K}]([\mathsf{B}](x \approx x \to \phi) \to (x \approx x \to \phi))$. Hence, $\vdash \phi \wedge x \not\approx y \to [\mathsf{K}]([\mathsf{B}]\phi \to \phi)$, and thus $\vdash \phi \wedge x \not\approx y \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$. Hence, $\vdash x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$.

Then, we show that $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}][\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi$ can be deduced in $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$. Equivalently, we show that $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x[\mathsf{K}]\phi$ can be deduced. Since $\vdash x \not\approx y \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$ for some fresh y, by $\mathtt{K_{wh}toK}$, we have $\vdash [\mathsf{K}_{\mathtt{wh}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$. Then, we first show that $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$. On the one hand, it is easy to check that we have $\vdash \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to z \approx x$ (where z is a fresh variable), and thus $\vdash [\mathsf{K}]\phi[z/x] \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$. Hence, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$ by $\mathtt{K_{wh}toK}$ (and $\mathtt{R}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$). On the other hand, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \wedge \phi \to [\mathsf{K}]\phi)$, and thus $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$ by $\mathtt{K_{wh}toFS}$. Hence, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to ([\mathsf{B}]\phi \to [\mathsf{K}]\phi)$. Then, by $\mathtt{FS\&KtoK_{wh}}$ and $4^{[\mathsf{K}]}$, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to ([\mathsf{K}]\phi \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x[\mathsf{K}]\phi)$, and thus $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x[\mathsf{K}]\phi$ by $\mathtt{K_{wh}toK}$. □

Finally, for Proposition 5.4, we only prove the case for $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]$, since the case for $[\mathsf{tB}^{\mathtt{MS}}_{\mathtt{FS}}]$ is similar. That is, we prove the following proposition:

Proposition A.17 *The following equivalence holds:*

$$\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus \mathtt{MONO} \quad = \quad \mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \to \phi)$$

PROOF. We first show that $x \not\approx y \to ([\mathsf{B}]\phi \to \phi)$ can be deduced in $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus \mathtt{MONO}$. Clearly $\vdash (x \approx y) \wedge (x \not\approx y) \to \phi$, i.e. $\vdash x \approx y \to (x \not\approx y \to \phi)$. Then, by $\mathtt{MONO}$, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \approx y) \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y \to \phi)$. It is also easy to check that $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \approx y)$. Hence, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y \to \phi)$. Then, by $\mathtt{K_{w}htoFS}$, $\vdash [\mathsf{B}](x \approx y \vee \phi) \to (x \not\approx y \to \phi)$. Hence, clearly $\vdash [\mathsf{B}]\phi \to (x \not\approx y \to \phi)$, i.e. $\vdash x \not\approx y \to ([\mathsf{B}]\phi \to \phi)$.

Then, we show that $\mathtt{MONO}$ is admissible in $\mathbf{S4.2}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}_{\approx} \oplus x \not\approx y \to ([\mathsf{B}]\phi \to \phi)$. Since we have $x \not\approx y \to ([\mathsf{B}]\phi \to \phi)$ for some fresh y, by $\mathtt{K_{wh}toK}$, $[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\phi \to \phi)$. Assume that $\vdash \phi \to \psi$. We first prove that $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to ([\mathsf{B}]\psi \to \psi)$. On the one hand, $\vdash \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to z \approx x$ (where z is a fresh variable), and thus $\vdash [\mathsf{K}]\phi[z/x] \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to \phi$. Then, since $\vdash \phi \to \psi$, $\vdash [\mathsf{K}]\phi[z/x] \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to \psi$, and thus $\vdash [\mathsf{K}]\phi[z/x] \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\psi \to \psi)$. Then, by $\mathtt{K_{wh}toK}$ and $\mathtt{R}^{[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]}$ $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge \neg[\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\psi \to \psi)$. On the other hand, clearly $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^y(x \not\approx y) \to ([\mathsf{B}]\psi \to \psi)$. Hence, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to ([\mathsf{B}]\psi \to \psi)$. And since $\vdash \phi \to \psi$, we also have $\vdash [\mathsf{K}]\phi \to [\mathsf{K}]\psi$. Hence, by $\mathtt{FS\&KtoK_{wh}}$, $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \wedge [\mathsf{K}]\phi \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\psi$, and thus $\vdash [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\phi \to [\mathsf{K}^{\mathtt{MS}}_{\mathtt{FS}}]^x\psi$ by $\mathtt{K_{wh}toK}$. □