EPTCS 437

Proceedings of the

Twentieth Conference on

Theoretical Aspects of Rationality and Knowledge

Düsseldorf, Germany, July 14-16, 2025

Edited by: Adam Bjorndahl

Published: 27th November 2025
DOI: 10.4204/EPTCS.437
ISSN: 2075-2180
Open Publishing Association

Table of Contents

Table of Contents i

Preface iv
Adam Bjorndahl

Mechanism Design under Unawareness - Extended Abstract 1
Kym Pram and Burkhard C. Schipper

More Reasoning, Less Outcomes: A Monotonicity Result for Reasoning in Dynamic Games 14
Andrés Perea

Solving Four Open Problems about Core Stability in Altruistic Hedonic Games 15
Jörg Rothe and Ildikó Schlotter

A Modular Equilibrium Theory for Admissibility in Games 30
Christian W. Bach and Stephan Jagau

Optimal Information Design in Sender-Receiver Cheap Talk Interactions 31
Itai Arieli, Ivan Geffner and Moshe Tennenholtz

Choquet Expected Utility and Never Best Choice 46
Christopher Kops, Hans Peters and Dries Vermeulen

Robust Price Discrimination 47
Itai Arieli, Yakov Babichenko, Omer Madmon and Moshe Tennenholtz

Hyperintensional Intention 48
Daniil Khaitovich and Aybüke Özgün

Coordinating Beliefs 65
Richard Roth

Common p-Belief with Plausibility Measures: Extended Abstract 66
Eric Pacuit and Leo Yang

Distributed Knowing How 80
Bin Liu and Yanjing Wang

Maximum Entropy and Bayesian Conditioning Under Extended Space 98
Boning Yu

On the Complexity of the Grounded Semantics for Infinite Argumentation Frameworks 112
Uri Andrews and Luca San Mauro

Conditionals Based on Selection Functions, Modal Operators and Probabilities 128
Tommaso Flaminio, Lluis Godo and Gluliano Rosella

Group Knowledge of Hypothetical Values 140
Alexandru Baltag and Sonja Smets

Causality Without Causal Models 160
Joseph Y. Halpern and Rafael Pass

Optimality of Simultaneous Consensus with Limited Information Exchange (Extended Abstract) . . . 175
Kaya Alpturer, Ron van der Meyden, Sushmita Ruj and Godfrey Wong

Common Knowledge, Sailboats, and Publicity 190
Sena Bozdag and Olivier Roy

Graded Distributed Belief 203
Emiliano Lorini and Dmitry Rozplokhas

Modal Logic for Reasoning About Uncertainty and Confusion 216
Marta Bílková, Thomas M. Ferguson and Daniil Kozhemiachenko

Comparing State-Representations for DEL Model Checking 233
Gregor Behnke, Malvin Gattinger, Avijeet Ghosh and Haitian Wang

Choosing What Game to Play without Selecting Equilibria: Inferring Safe (Pareto) Improvements in Binary Constraint Structures 251
Caspar Oesterheld and Vincent Conitzer

Skating System Unveiled: Exploring Preference Aggregation in Ballroom Tournaments 271
Laryssa Horn, Paul Nüsken, Jörg Rothe and Tessa Seeger

Logic of (Common or Distributed)* Knowledge 286
Chenwei Shi

Who is Afraid of Minimal Revision? 301
Edoardo Baccini, Zoé Christoff, Nina Gierasimczuk and Rineke Verbrugge

Are Large Random Graphs Always Safe to Hide? 321
Sourav Chakraborty, Sujata Ghosh and Smiha Samanta

Prudent Rationalizability and the Best Rationalization Principle 334
Nicodemo De Vito

Complexity of Łukasiewicz Modal Probabilistic Logics 350
Daniil Kozhemiachenko and Igor Sedlár

Bifurcation Logic: Separation Through Ordering 365
Didier Galmiche, Timo Lang, Daniel Méry and David Pym

Modal Logic for Simulation, Refinement, and Mutual Ignorance 379
Hans van Ditmarsch, Tim French, Rustam Galimullin and Louwe B. Kuijer

Impure Simplicial Complex and Term-Modal Logic with Assignment Operators 399
Yuanzhe Yang

AI-Generated Compromises for Coalition Formation: Modeling, Simulation, and a Textual Case Study 417
Eyal Briman, Ehud Shapiro and Nimrod Talmon

Muddy Waters 433
Hans van Ditmarsch

Preface

Adam Bjorndahl
abjorn@cmu.edu

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 were accepted for presentation at the Twentieth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2025), held between July 14 and July 16, 2025, at the Heinrich-Heine-Universität, Düsseldorf, Germany. The conference website can be found at `https://ccc.cs.uni-duesseldorf.de/tark-2025/`

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

- Kaushik Basu (Cornell University)
- Elizabeth Maggie Penn (Emory University)
- Marcus Pivato (Université Paris 1 Panthéon-Sorbonne)
- Tuomas Sandholm (Carnegie Mellon University)

The Programme Committee received 65 submissions. Of these, 33 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 33 accepted papers, 22 were presented as an oral lecture and 11 as a poster presentation accompanied by a flash talk. The best paper award was given to Andrés Perea for his submission, "More reasoning, less outcomes: A monotonicity result for reasoning in dynamic games".

As with previous proceedings of TARK, this volume exemplifies the interdisciplinary nature of research on theoretical aspects of rationality and knowledge: many papers engage seriously with multiple fields of study, crossing disciplinary boundaries, and all of them have been written so as to be accessible to readers from diverse backgrounds.

The members of the Programme Committee for this conference were:

- Christian Bach, University of Reading
- Adam Bjorndahl, Carnegie Mellon University (chair)
- Zoé Christoff, University of Groningen
- Franz Dietrich, Paris School of Economics & CNRS

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. iv–vi, doi:10.4204/EPTCS.437.0

- Hans van Ditmarsch, University of Toulouse, CNRS, IRIT
- Amanda Friedenberg, University of Michigan
- Malvin Gattinger, University of Amsterdam
- Sujata Ghosh, Indian Statistical Institute
- Nina Gierasimczuk, Technical University of Denmark
- Davide Grossi, University of Groningen
- Joseph Y. Halpern, Cornell University
- Aviad Heifetz, The Open University of Israel
- Willemien Kets, Utrecht University
- Jérôme Lang, CNRS, LAMSADE, Université Paris-Dauphine
- Larry Moss, Indiana University Bloomington
- Aybüke Özgün, University of Amsterdam
- Eric Pacuit, University of Maryland
- Andrés Perea, University Maastricht
- Olivier Roy, Universität Bayreuth
- Burkhard Schipper, University of California, Davis
- Sonja Smets, University of Amsterdam
- Elias Tsakas, Maastricht University
- Rineke Verbrugge, University of Groningen, Institute of Artificial Intelligence
- Francesca Zaffora Blando, Carnegie Mellon University

Many other people assisted with the reviewing process, including: Jeremy Avigad, Bart Bogaerts, Andrea Bracciali, Armando Castaneda, Balder ten Cate, Michele Crescenzi, Tommaso Flaminio, Hugo Rincon Galeana, Rustam Galimullin, Ivan Geffner, Avijeet Ghosh, Dušan Knop, Louwe B. Kuijer, David Lehnherr, Dazhu Li, Brett McLean, Krzysztof Mierzewski, Munyque Mittelmann, Anantha Padmanabha, Clayton Peterson, Soroush Rafiee Rad, R. Ramanujam, Christoph Schlom, Sunil Simon, Philip Sink, Shawn Standefer, Bruno Teheux, Paolo Turrini, Nicodemo De Vito, Haitian Wang, Yanjing Wang, Atefeh Keshavarzi Zafarghandi, and Gabriel Ziegler. A special, extra thanks is also owed to Rineke Verbrugge for taking on the PC Chair duties for those submissions I had a conflict of interest for. I also thank Rob van Glabbeek of EPTCS for his support along the way to creating the proceedings themselves.

I would like to extend my sincere gratitude to all the members of the Programme Committee and all other reviewers for the effort they put in to make TARK the world-class conference that it is. Through their expertise, insightful feedback, and fruitful discussions, authors who submit to TARK receive extremely helpful and constructive feedback, and the proceedings volume is of the highest quality. Of course I am also grateful to the authors themselves for submitting their novel and interesting work, without which there would be no conference at all!

Lastly, I would like to acknowledge and thank the tireless work of Jörg Rothe, the Organizing Chair, who made this edition of TARK at Heinrich-Heine-Universität in Düsseldorf possible. Without his dedication (and the efforts of Paul Nüsken to maintain and update the website), this conference could not have taken place in such a lovely venue nor would it have been so well organized.

One final mention of thanks to the TARK General Chair, Joe Halpern, who started this conference series in 1986 and who was always available and more than willing to help with anything that required advice, planning, or just another set of eyes. Joe was my PhD advisor and introduced me to TARK back in 2011 when I was a graduate student. It was the first real academic conference I presented my work at, and since then I have had a paper at every single meeting, so it is an enormous honor for me to have had the chance this year to chair the Programme Committee and give something back to this event that has given so much to me.

Adam Bjorndahl
Programme chair, TARK 2025
July 14th, 2025

Mechanism Design under Unawareness
- Extended Abstract -

Kym Pram
University of Nevada, Reno
kpram@unr.edu

Burkhard C. Schipper
University of California, Davis
bcschipper@ucdavis.edu

We study the design of mechanisms under asymmetric awareness and information. While the mechanism designer cannot necessarily commit to a particular social choice function in the face of unawareness, she can at least commit to properties of social choice functions such as efficiency given ex post awareness. Assuming quasi-linear utilities and private values, we show that we can implement in conditional dominant strategies a social choice function that is utilitarian ex post efficient under pooled awareness without the need of the social planner being fully aware ex ante. To this end, we develop novel dynamic versions of Vickrey-Clarke-Groves mechanisms in which true types are revealed and subsequently elaborated at endogenous higher awareness levels. We explore how asymmetric awareness affects budget balance and participation constraints. We show that ex ante unforeseen contingencies are no excuse for deficits. Finally, we propose a dynamic elaboration reverse second price auction for efficient procurement of complex incompletely specified projects with budget balance and participation constraints.

1 Introduction

Recent years witnessed the development of logic, game theory, and decision theory with unawareness. Many contributions to the literature have been presented at prior TARK conferences ([1, 7, 8, 10, 11, 12, 14, 15, 16, 22, 23, 24, 25, 28, 27, 29, 30, 32, 37, 40, 42, 43]). Models of unawareness have been applied to disclosure games, moral hazard, contract theory, screening, delegation, speculation, financial market microstructure, default in general equilibrium, electoral campaigning, business strategy, and conflict resolution; for a bibliography, see [41]. Yet, what is missing from the literature is the social engineering perspective: How to design mechanisms in the presence of unawareness? Mechanism design studies the design of institutions governing collective decisions such as markets, contracts, or political systems. Rather than defining a game and deriving its solutions like in game theory, mechanism design first identifies desirable outcomes and then designs a game such that players implement the outcome in a solution to the game. Typically, mechanisms apply to contexts with asymmetric information. However, agents may not just face asymmetric information but also asymmetric awareness. Unawareness refers to the lack of conception rather than the lack of information. Agents and the designer of mechanisms may be unaware of some events and actions affecting values and costs of complex private or public projects, allocations, and outcomes and may not even realize this fact. In this paper, we extend mechanism design to unawareness. Under the economically relevant assumption of quasi-linear preferences, we show how to design efficient mechanisms in the presence of asymmetric awareness. To this end, we introduce dynamic direct elaboration mechanisms in which not only are types communicated from agents but also awareness is raised among participants in back-and-forth communication between participants and transfers are inspired by Vickrey-Clarke-Groves (VCG) mechanisms.

Besides our theoretical motivation for extending mechanism design to unawareness, we are motivated by practical concerns about the usefulness of explicit mechanisms in reality. For instance, it has been

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 1–13, doi:10.4204/EPTCS.437.1

© Kym Pram & Burkhard C. Schipper
This work is licensed under the
Creative Commons Attribution License.

argued that auctions are inappropriate when projects are complex, incompletely designed, and custom-made such as the procurement of new fighter jets, buildings, consulting services, and IT projects. It is claimed that auctions may stifle communication between buyers and sellers, preventing buyers and sellers to use each other's expertise when designing projects ([17, 6, 5]). The Department of Defense (DoD), the General Service Administration (GSA), and the National Aeronautics and Space Administration (NASA) recently proposed to amend the Federal Acquisition Regulation (FAR) suggesting the prohibition of the use of reverse auctions for complex, specialized, or substantial design and construction services ([36]).[1] The received view is that for complex projects, negotiations outperform auctions because they can better take advantage of the expertise and know-how of contractors ([17, 46, 5]). We seek an extension of mechanism design to unawareness to make mechanisms also applicable to complex and incompletely designed projects whose efficient implementation require the pooling of expertise among participants. Our proposed mechanisms combine features of common business practices such as Request for Information (RFI) and Request for Proposals (RFP)[2] with standard VCG mechanisms such as second price auctions. That is, we combine features of negotiations and traditional mechanism design. While traditionally in economics, negotiations have been interpreted mostly as bargaining over the surplus and thus as a substitute to mechanisms, we view negotiations more as interactively defining the surplus using the expertise of participants, which is complementary to traditional mechanisms implementing and allocating the surplus.

Compared to typical TARK papers, this paper focuses on foundational aspects of mechanism design rather than epistemic foundations. It serves as a proof of concept that models of awareness can be fruitfully applied to fundamental problems of economic design. Moreover, our results are very general and robustly apply to no matter how beliefs and awareness are distributed among agents. From an epistemic perspective, we provide answers to questions about how awareness revelation can be incentivized among agents and how awareness can be pooled among agents via public announcements by a mediator.

This is just an extended abstract; the complete paper ([39]) with additional details, further results, and all proofs is available under `https://faculty.econ.ucdavis.edu/faculty/schipper/mechunaw.pdf` .

2 Model

2.1 Payoff Types with Unawareness

Let L be a finite lattice with order $\trianglerighteq$. Elements of the lattice represent awareness levels. The join of the lattice is denoted by $\bar{\ell} \in L$. Define the sublattice $L(\ell) := \{\ell' \in L : \ell' \trianglelefteq \ell\}$ for $\ell \in L$. The significance of $L(\ell)$ is that an agent with awareness level ℓ can only reason about awareness levels in $L(\ell)$.

Fix a nonempty finite set of agents I. For each agent $i \in I$, there a collection of nonempty disjoint payoff type spaces $\{T_i^\ell\}_{\ell \in L}$. Let $\mathscr{T}_i := \bigcup_{\ell \in L} T_i^\ell$. A payoff type $t_i \in \mathscr{T}_i$ is more than just a value for an object. If $t_i \in T_i^\ell$, then describing the payoff type also requires at least awareness level ℓ. We illustrate this feature with the following example.

Example 1 Consider the context of procurement. A principal may invite agents to bid on a complex project. Agents do their due diligence and identify relevant items that drive their costs. Suppose there

[1] When analyzing the use of auctions for procurement at the DoD, [2] state that auctions work best for well-specified and off-the-shelf commodities but nevertheless remark that the Navy was able to achieve substantial savings with procurement auctions for customized and complex contracts such as CVN camels, i.e., devices for safe separation of ships and piers.

[2] E.g., Federal Acquisitions Regulation 15.203.

are items $\{a,b,c\}$. Agents may be unaware of some items even after their due diligence. Agent 1 may only be aware of items $\ell_1 = \{a,b\}$ while agent 2 is only aware of items $\ell_2 = \{b,c\}$. In this case, L is isomorphic to the set of all subsets in $2^{\{a,b,c\}}$ and the natural lattice order is induced by set inclusion on $2^{\{a,b,c\}}$. Agent 1 may submit a bid given by the following table t_1 while agent 2 may submit a bid given by table t_2.

$t_1 =$

Item	Cost
a	23
b	41
Total	64

$t_2 =$

Item	Cost
b	38
c	29
Total	67

As the example demonstrates, payoff types can be multi-dimensional with varying dimensions. Needless to say, the lattice approach is more general than multi-dimensional payoff type spaces. □

For any agent $i \in I$ and awareness levels $\ell, k \in L$ with $k \trianglerighteq \ell$, we require a surjective projection $r_\ell^k : T_i^k \longrightarrow T_i^\ell$ such that for all $\ell, \ell', \ell'' \in L$, $\ell'' \trianglerighteq \ell' \trianglerighteq \ell$, $r_\ell^{\ell'}(r_{\ell'}^{\ell''}(t_i)) = r_\ell^{\ell''}(t_i)$ and $r_\ell^\ell(t_i) = t_i$. For brevity, we do not index r_ℓ^k by agents. The projection relates payoff types across awareness levels. Before we illustrate this notion, we also define extensions of payoff types to greater awareness levels. For any agent $i \in I$, awareness level $\ell \in L$, and payoff type $t_i \in T_i^\ell$, we let $t_i^\uparrow := \bigcup_{k \trianglerighteq \ell, k \in L} (r_\ell^k)^{-1}(t_i)$. That is, $t_i^\uparrow$ is the union of inverse images at (weakly) greater awareness levels of payoff type t_i. Similarly, for any subset of payoff types in a given payoff type space, we let superscript "$\uparrow$" indicate the union of inverse images in payoff type spaces corresponding to greater awareness levels. We illustrate these notions in our prior example.

Example 1 (Continuation) Continuing our prior example, consider payoff types t_1' and t_1'' of agent 1:

$t_1' =$

Item	Cost
a	23
b	41
c	16
Total	80

$t_1'' =$

Item	Cost
a	23
b	41
c	15
Total	79

Both payoff types are described with items in $\{a,b,c\}$. That is, both $t_1', t_1'' \in T_1^{\{a,b,c\}}$. When payoff information on item c is stripped away from payoff type t_1' we obtain payoff type t_1. That is, $r_{\{a,b\}}^{\{a,b,c\}}(t_1') = t_1$. Similarly, $r_{\{a,b\}}^{\{a,b,c\}}(t_1'') = t_1$. Thus, $t_1, t_1', t_1'' \in t_1^\uparrow$. □

For any agent $i \in I$, we define $\lambda : \mathscr{T}_i \longrightarrow L$ by $\lambda(t_i) = \ell$ if $t_i \in T_i^\ell$. Again, for brevity we do not index λ by agents. The function λ indicates the awareness level that is required to describe the payoff type. For instance, in Example 1, $\lambda(t_1) = \{a,b\}$ while $\lambda(t_1') = \{a,b,c\}$.

Our framework is general enough to capture payoff types described by all kinds of formal objects like sets of formulae in a formal language (e.g., [26]), abstract sets, vectors, matrices, formal concepts, presheaves etc. that may represent verbal, quantitative, or pictorial features of proposals, tenders, quotations, messages etc. in business practice. By abstracting from these particular features, we obtain a theory that works for more than one kind of formalism, ensure tractability, and focus on what is really essential for modeling awareness levels, namely the existence of an order of expressiveness of descriptions.

For each agent $i \in I$, payoff types and awareness levels are drawn consistently as follows: At awareness level $\bar{\ell}$, nature draws a payoff type $\bar{t}_i \in T_i^{\bar{\ell}}$ in the upmost payoff type space, interpreted as agent i's true payoff type if she were aware of everything, and an awareness level $\ell_i \in L$. Consequently, the agent's perceived payoff type is $r_{\ell_i}^{\bar{\ell}}(t_i)$. That is, agent i can "miss something" but he cannot perceive the "wrong" payoff type w.r.t. what he is aware. More generally, for any awareness level $\ell \in L$, nature draws

agent i's corresponding type $r_\ell^{\bar{\ell}}(\bar{t}_i)$ and agent i's awareness level $\ell' = \ell_i \wedge \ell$. This means in particular, that if $\ell = \ell_i$, then it is payoff type $r_{\ell_i}^{\bar{\ell}}(t_i)$ and awareness level ℓ_i as just discussed. If $\bar{\ell} \trianglerighteq \ell \trianglerighteq \ell_i$, nature draws agent i's corresponding type $r_\ell^{\bar{\ell}}(\bar{t}_i)$ and awareness level ℓ_i. If $\ell \in L$ with $\ell_i \trianglerighteq \ell$, nature draws agent i's corresponding type $r_\ell^{\bar{\ell}}(\bar{t}_i)$ and awareness level ℓ. Finally, if $\ell \in L$ with $\ell \not\trianglerighteq \ell_i$, nature draws agent i's corresponding type $r_\ell^{\bar{\ell}}(\bar{t}_i)$ and awareness level $\ell' = \ell_i \wedge \ell$. This specifies perceived payoff types and awareness levels consistently across the type spaces $\{T_i^\ell\}_{\ell \in L}$. Notice that if $t_i \in T_i^\ell$ for some $\ell \in L$, then the agent's perceived type is always in a type space with (weakly) less awareness than ℓ.

For any $\ell \in L$, let $\boldsymbol{T}^\ell := \times_{i \in I} T_i^\ell$. Moreover, let $\mathscr{T} := \times_{i \in I} \mathscr{T}_i$. Similarly, we let $\boldsymbol{T}_{-i}^\ell := \times_{j \in I \setminus \{i\}} T_j^\ell$ and $\mathscr{T}_{-i} := \times_{j \in \in I \setminus \{i\}} \mathscr{T}_j$.

When agents have the payoff type profile $\boldsymbol{t} = (t_i)_{i \in I} \in \mathscr{T}$, their *pooled* awareness level is $\check{\lambda}(\boldsymbol{t}) := \bigvee_{i \in I} \lambda(t_i) \in L$, i.e., the join of all agent's awareness levels at payoff type profile $\boldsymbol{t}$. Since L is a finite lattice, the join always exists in L. Note that $\check{\lambda}(\boldsymbol{t})$ may be a greater awareness level than any of the agent's awareness levels.

For each $\ell \in L$, there is a nonempty compact set of outcomes or allocations X^ℓ. This formulation allows for unawareness of outcomes. We require that $\ell \trianglerighteq k$ implies $X^k \subseteq X^\ell$. Denote by $X := \bigcup_{\ell \in L} X^\ell$. Clearly, $X = X^{\bar{\ell}}$.

Each agent $i \in I$ has an upper semi-continuous utility function $u_i : \bigcup_{\ell \in L} X^\ell \times T_i^\ell \longrightarrow \mathbb{R}$. From this formulation it is clear that we focus on private payoff types. To facilitate a commonly used notion of efficiency, we assume that each agent's utility function is quasilinear. I.e., for each $i \in I$, $u_i(x, t_i) := v_i(x_0, t_i) + x_i$ for $x = (x_0, x_1, ..., x_{|I|}) \in X^\ell := X_0^\ell \times \mathbb{R}^{|I|}$ and $t_i \in T_i^\ell$, $\ell \in L$. As usual, x_0 describes the physical properties of the outcome while $(x_i)_{i \in I}$ represents the vector of transfers made *to* agents.

We denote the outcome function by $f_0 : \mathscr{T} \longrightarrow X_0$. That is, $f_0(\boldsymbol{t})$ is the physical outcome prescribed by the outcome function f_0 to type profile $\boldsymbol{t}$. We require that for any $\boldsymbol{t} \in \mathscr{T}$, $f_0(\boldsymbol{t}) \in X_0^{\check{\lambda}(\boldsymbol{t})}$. That is, the social planner can use joint awareness to select the outcome/allocation. If the reported payoff type profile is $\boldsymbol{t}$, then the planner's awareness is $\check{\lambda}(\boldsymbol{t})$ and hence outcomes in $X_0^{\check{\lambda}(\boldsymbol{t})}$ are selected.[3] The formulation also makes clear that the social planner cannot necessarily describe the outcome function to agents in advance if the social planner is unaware of something herself. However, we assume that the social planner can commit to abstract properties of outcome functions such as efficiency, which in a quasi-linear setting takes the following form:

Definition 1. *The outcome function f_0 is* utilitarian ex-post efficient *if for all $\ell \in L$ and $\boldsymbol{t} = (t_i)_{i \in I} \in \boldsymbol{T}^\ell$,*

$$\sum_{i \in I} v_i(f_0(\boldsymbol{t}), t_i) \geq \sum_{i \in I} v_i(x_0, t_i) \text{ for all } x_0 \in X_0^\ell. \tag{1}$$

2.2 Dynamic Direct Elaboration Mechanisms

In the appendix, we show that VCG mechanisms are insufficient for efficient implementation under unawareness as they fail to pool awareness. In order to pool awareness among agents and allow agents to provide information on issues they are or became aware, we allow for communication not just *from* agents but also *to* agents. That is, we envision a mediator who receives messages about payoff types from agents, pools the awareness contained in those messages, and sends back messages to agents that

[3] Typically the planer is not considered as an agent in mechanism design and hence we do not explicitly consider the awareness that the planner may have. Yet, all of our results remain intact when the planner joins her awareness together with the awareness of all agents whenever communicating the pooled awareness level in the dynamic elaboration mechanisms introduced in the next section.

potentially raise the awareness of agents to the pooled awareness level. Subsequently, agents may want to elaborate on their prior messages at least to the details of the pooled awareness level. This procedure may be repeated till no agent wants to elaborate further. To this end, we introduce a new class of dynamic mechanisms.

Definition 2 (Dynamic Direct Elaboration Mechanism)**.** *The* dynamic direct elaboration mechanism *implementing outcome function f_0 is defined recursively by the following algorithm:*[4]

Stage $n = 1$: Each agent $i \in I$ must report a type t_i^1.

Stages $n > 1$: Each agent i must report a type $t_i^n \in (t_i^{n-1})^{\uparrow} \cap \left(T_i^{\check{\lambda}(\boldsymbol{t}^{n-1})} \right)^{\uparrow}$.

Stop: If $t_i^{n+1} = t_i^n$ for all $i \in I$, then $f_0(\boldsymbol{t}^{n+1})$ is implemented.

At the first stage, every agent reports a type. At later stages, agents can elaborate on their prior reported types. When agents report the payoff type profile $\boldsymbol{t}^{n-1}$ in stage $n-1$, the mediator awareness level is the pooled awareness level $\check{\lambda}(\boldsymbol{t}^{n-1})$. Consequently, he communicates back the pooled awareness level $\check{\lambda}(\boldsymbol{t}^{n-1})$ to all agents.[5] Importantly, all agents receive the same message from the mediator. That is, we can use public messages. At the next stage, each agent i must report a type at least at the pooled awareness level $T_i^{\check{\lambda}(\boldsymbol{t}^{n-1})}$. This report must be consistent with her prior reported type which is implied by the requirement $(t_i^{n-1})^{\uparrow} \cap \left(T_i^{\check{\lambda}(\boldsymbol{t}^{n-1})} \right)^{\uparrow}$. For instance, if an agent reported a type whose awareness is strictly below the pooled awareness of all agents' reports, then she must now report a type at least at the pooled awareness level $\check{\lambda}(t_i^{n-1})$ that is consistent with her prior reported type, i.e., a type in $(t_i^{n-1})^{\uparrow}$. However, it could be the case that she has previously pretended to have much less awareness so that the pooled awareness level did not yet incorporate all her awareness. Thus, we allow her to report a type with awareness strictly larger than the pooled awareness level. This motivates the requirement $t_i^n \in (t_i^{n-1})^{\uparrow} \cap \left(T_i^{\check{\lambda}(\boldsymbol{t}^{n-1})} \right)^{\uparrow}$ rather than $t_i^n \in (t_i^{n-1})^{\uparrow} \cap T_i^{\check{\lambda}(\boldsymbol{t}^{n-1})}$. If an agent did report at the pooled awareness level in the previous stage, then she is not able to change her report in the current stage. The mechanism stops once no agents revise their reports any further. Once it stops, it implements the physical outcome associated to the final reported type profile by f_0.

Example 1 (Continued) To continue our Example 1 above, suppose that agent 1 reported in stage 1 payoff type $t_1^1 = t_1$ and agent 2 reported $t_2^1 = t_2$. Pooling awareness leads to $\check{\lambda}(t_1, t_2) = \{a, b, c\}$. At stage 2, both agents must now elaborate their prior reported type and report a type in $T_i^{\{a,b,c\}}$. For instance, agent 1 could report $t_1^2 = t_1'$. Since no further awareness can be revealed after stage 2, the mechanism must conclude after stage 3. □

Note that awareness and information is transmitted both from agents to the mediator but also from the mediator to the agents. Thus, awareness of agents may change endogenously when interacting in the mechanism. Since agents report directly types and elaborate in later stages one their prior reported types,

[4]Transfers can be arbitrary at this point and will be specified later when we consider particular dynamic direct elaboration mechanisms.

[5]The current formulation presumes that the mediator, mechanism designer, or social planner has no relevant awareness herself. If she does have relevant awareness, she can incorporate it easily into the message communicated back to agents.

we call it a "direct elaboration" mechanism. Note that by the definition of join of the lattice of spaces, $\check{\lambda}(\boldsymbol{t}^{n-1}) \trianglerighteq \lambda(t_i^{n-1})$ for any $i \in I$.

The mechanism stops when no agent wants to further elaborate on her type. Clearly, since L is finite, the mechanism must stop at some finite stage n. Moreover, when it stops at n, then $\boldsymbol{t}^n \in \boldsymbol{T}^{\check{\lambda}(\boldsymbol{t}^n)}$. That is, all agents must have reported twice in a row types at the *same* awareness level.

Denote by σ_i agent i's reporting strategy, by σ_i^* agent i's truth-telling strategy, by $\boldsymbol{\sigma}$ a strategy profile, by h_i an information set of agent i's, $\Sigma_i(h_i)$ agent i's strategies that allow information set h_i, $\mathfrak{T}$ the set of sequences of reported payoff type profiles, by $\tau(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma})$ the profile of reported payoff type profiles the final when $\boldsymbol{t}$ is the profile of payoff types, $\boldsymbol{\ell}$ is the profile of awareness levels, and $\boldsymbol{\sigma}$ is the profile of strategies, by $\tau^*(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma})$ the final reported payoff type profile thereof. We also extend definition of λ to information sets. In the complete manuscript ([39]), we provide details on the dynamic game with unawareness induced by dynamic direct elaboration mechanisms, information sets, and strategies.

For any agent $i \in I$, let $f_i : \mathfrak{T} \longrightarrow \mathbb{R}$ denote the transfer paid *to* agent i in the dynamic direct elaboration mechanisms. These transfers will depend on the precise versions of the direct elaboration mechanism studied below. In contrast to the outcome function f_0 we allow transfers to depend on the entire sequence of reported type profiles.

We let the social choice function (i.e., the outcome function and transfers) be denoted by $f : \mathfrak{T} \longrightarrow \mathbb{R}$ and defined by $f = (f_0, f_1, ..., f_{|I|})$. That is, for any strategy profile $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$, $\ell \in L$, initial profiles of payoff types $\boldsymbol{t} \in \boldsymbol{T}^{\ell}$, initial profiles of awareness levels $\boldsymbol{\ell} \in L(\ell)^{|I|}$, $f(\tau(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma})) = (f_0(\tau^*(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma})), f_1(\tau(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma})), ..., f_{|I|}(\tau(\boldsymbol{t},\boldsymbol{\ell},\boldsymbol{\sigma}))\big)$.

Definition 3. *The dynamic direct elaboration mechanism truthfully implements the social choice function f in conditional dominant strategies if for all agents $i \in I$, information sets $h_i \in H_i(\sigma_i^*)$, opponents' strategy profiles $\boldsymbol{\sigma}_{-i} \in \boldsymbol{\Sigma}_{-i}$, initial profiles of payoff types $\boldsymbol{t} \in \boldsymbol{T}^{\lambda(h_i)}$ (as perceived by i in h_i), and initial profiles of awareness levels $\boldsymbol{\ell} \in L(\lambda(h_i))^{|I|}$ (as perceived by i in h_i) such that $h_i \in H_i(\boldsymbol{t},\boldsymbol{\ell},(\sigma_i^*,\boldsymbol{\sigma}_{-i}))$,*

$$u_i(f(\tau(\boldsymbol{t},\boldsymbol{\ell},(\sigma_i^*,\boldsymbol{\sigma}_{-i})),t_i(h_i)) \quad \geq \quad u_i(f(\tau(\boldsymbol{t},\boldsymbol{\ell},(\sigma_i,\boldsymbol{\sigma}_{-i})),t_i(h_i)) \tag{2}$$

for all $\sigma_i \in \Sigma_i(h_i)$.

Conditional dominance strengthens dominance by requiring each agent not only to select a (partial) strategy that is ex-ante dominant but also dominant conditional on each information set. This becomes important when agents cannot anticipate all information sets ex-ante and thus cannot select ex-ante a strategy for the entire game in extensive form. The use of conditional dominance as a solution concept for implementing outcomes in our mechanisms implies that our results do not depend on agents' beliefs about other agents, their awareness etc. That is, revelation of awareness and private information in the mechanisms is robust to possibly misspecified beliefs by the mediator and agents.

Definition 4. *A outcome function f_0 is truthfully implemented at the pooled awareness level if for any $\bar{\boldsymbol{t}} \in \boldsymbol{T}^{\bar{\ell}}$ and $(\ell_i)_{i\in I} \in L^{|I|}$, $f_0\left(r^{\bar{\ell}}_{(\bigvee_{i\in I}\ell_i)}(\bar{\boldsymbol{t}})\right)$ is implemented.*

Recall that $(\bar{\boldsymbol{t}},(\ell_i)_{i\in I}) \in \boldsymbol{T}^{\bar{\ell}} \times L^{|I|}$ represents the move of nature selecting both actual payoff types and awareness levels for all agents. Consequently, $\bigvee_{i\in I}\ell_i$ represents the pooled awareness level.

3 Efficient Implementation

To implement efficiently at the pooled awareness level, we specify for each agent $i \in I$ the transfers f_i that incentivize both the revelation of information and raising awareness. Revelation of information is achieved via VCG transfers. Raising awareness requires an extra term in the transfer functions.

Definition 5 (Dynamic Elaboration VCG Mechanism)**.** *We say that the dynamic direct elaboration mechanism implementing f is a dynamic elaboration VCG mechanism if f_0 is utilitarian ex-post efficient and transfers f_i to agent $i \in I$ are given by for any $(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \in \mathfrak{T}$,*

$$f_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n) := \sum_{j\neq i} v_j(f_0(\boldsymbol{t}^n), t_j^n) + y_i^{\breve{\lambda}(\boldsymbol{t}^n)}(\boldsymbol{t}^n_{-i}) + a_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n), \tag{3}$$

where, for each ℓ, $y_i^\ell : \boldsymbol{T}^\ell_{-i} \longrightarrow \mathbb{R}$ is an arbitrary function and $a_i(\cdot)$ is defined by:

$$a_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n) := \begin{cases} m_i(\breve{\lambda}(\boldsymbol{t}^n)) & \text{if } i = i^*(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \\ -\frac{1}{|I|-1} m_j(\breve{\lambda}(\boldsymbol{t}^n)) & \text{if } j = i^*(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \neq i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where

$$i^*(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \;:=\; \left\{ i \in I : \lambda(t_i^k) = \breve{\lambda}(\boldsymbol{t}^n) \text{ and } \nexists j \neq i, k' \leq k \text{ such that } \lambda(t_j^{k'}) = \breve{\lambda}(\boldsymbol{t}^n) \right\}$$

and $m_i(\ell)$ is defined recursively, as follows: $m_i(\underline{\ell}) := 0$ and for any $\ell \triangleright \underline{\ell}$,

$$m_i(\ell) \;:=\; \max_{\substack{\ell' \lhd \ell \\ \boldsymbol{t}' \in \boldsymbol{T}^{\ell'} \\ \boldsymbol{t} \in \boldsymbol{T}^{\ell}}} \left(m_i(\ell') + v_i(f_0(\boldsymbol{t}'), t_i) + \sum_{j\neq i} v_j(f_0(\boldsymbol{t}'), t_j') + y_i^{\ell'}(\boldsymbol{t}'_{-i}) - \sum_j v_j(f_0(\boldsymbol{t}), t_j) - y_i^{\ell}(\boldsymbol{t}_{-i}) \right). \tag{5}$$

The mechanism can viewed as a dynamic version of the Vickrey-Clarke-Groves (VCG) mechanisms ([20, 21]). Note that the transfers can be described without necessarily being aware of all payoff type profiles. The mechanism designer commits to implement a utilitarian ex-post efficient outcome given the agents' final reports of payoff types. Each agent is paid the total welfare of others given the final reports of payoff types plus a term that depends on only the opponents' payoff types (and the final pooled awareness level) and additionally a term incentivizing raising of awareness. This last term does not only depend on the final reported payoff type profile but on the sequence of reported payoff type profiles. It matters who reports the pooled awareness level first. Note that $i^*(\boldsymbol{t}^1,...,\boldsymbol{t}^n) = \emptyset$ if there is no agent who reports the pooled awareness level (which can happen if the individual awareness levels are incomparable) or if several agents simultaneously report the pooled awareness level first. If i is the unique agent who first reports the pooled awareness level, then a_i receives a payment related to the cost that raising awareness could impose on her utility from the mechanism, net of the $a_i(\cdot)$ term itself.

Theorem 1. *The dynamic elaboration VCG mechanism truthfully implements in conditionally dominant strategies a utilitarian ex-post efficient outcome under pooled awareness.*

The proof is contained the complete manuscript ([39]).

Proposition 1. *The utilitarian ex-post efficient outcome under pooled awareness is truth-fully implemented in conditional dominant strategies in the game induced by the dynamic elaboration VCG mechanism in at most three stages.*

4 No Deficit

Ideally, we like our mechanisms to satisfy desirable properties beyond efficiency. In the complete manuscript, we characterize budget balance of the dynamic elaboration VCG mechanism and show that

it does not impose constraints beyond static VCG mechanisms. In classical mechanism design, it is well known that the VCG mechanisms like the Groves mechanisms can not satisfy budget balance in general ([19]). We may want to look for weaker requirements. Not being budget balanced means that the mechanisms can run a deficit or surplus. In the context of unawareness, we are interested more in running no deficit than running no surplus for two reasons: First, if the mechanisms runs a deficit larger than the mechanism designer anticipated, agents may suspect that the mechanism designer is not committed to the mechanisms and become reluctant to truth-fully report their payoff types. Second, unforeseen contingencies are often used to justify budget overruns. But are deficits really inevitable in the presence of unawareness?

Recall that we used the convention that transfers f_i denote transfers *to* agent i. The following property is sometimes also called weak budget balance (e.g., [44]).

Definition 6 (No deficit)**.** *We say that the dynamic direct elaboration mechanism with transfer functions $(f_i)_{i\in I}$ satisfies no deficit if for all $(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \in \mathfrak{T}$,*

$$\sum_{i\in I} f_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n) \quad \leq \quad 0. \tag{6}$$

In classical mechanism design, it is well known that a Groves mechanism may run a deficit, but the Clarke (or Pivot) mechanism, in which each agent pays the negative externality that they impose on other agents through their effect on the social choice, does not. Therefore, the idea is to design a version of dynamic elaboration VCG mechanisms with Clarke transfers and show that it does not run a deficit.

To define the mechanism, we have to specify externalities. For any agent $i \in I$, define the $(-i)$-utilitarian ex-post efficient outcome function $f_0^{-i} : \mathscr{T} \longrightarrow X_0$ by $f_0^{-i}(\boldsymbol{t}) = \arg\max_{x_0 \in X_0^{\check{\lambda}(\boldsymbol{t})}} \sum_{j\neq i} v_j(x_0,t_j)$. That is, for each profile of payoff types $\boldsymbol{t} \in \mathscr{T}$, f_0^{-i} maximizes utilitarian ex-post welfare taking into account only the value functions of agent i's opponents. Note that different from restricted outcome functions in standard Clarke mechanism the argument of f_0^{-i} is the *full* profile of payoff types $\boldsymbol{t} = (t_j)_{j\in I}$ rather than just $\boldsymbol{t}_{-i}$. The reason is that in order to compute f_0^{-i} we need the awareness of *all* agents, not just agents $j \neq i$: Although agent i value is not considered when evaluating the social welfare of an outcome, for agents $j \neq i$, the efficiency of this outcome is still evaluated at the pooled awareness level.

Definition 7 (Dynamic Elaboration Clarke Mechanism)**.** *We say that the dynamic direct elaboration mechanism implementing f is a dynamic elaboration Clarke mechanism if it is a dynamic elaboration VCG mechanisms with, for all $\boldsymbol{t}^n \in \boldsymbol{T}^{\check{\lambda}(\boldsymbol{t}^n)}$ and $i \in I$,*

$$y_i^{\check{\lambda}(\boldsymbol{t}^n)}(\boldsymbol{t}^n_{-i}) := -\sum_{j\neq i} v_j(f_0^{-i}(\boldsymbol{t}^n), t_j^n).$$

Observe that for any $i \in I$, if $\sum_{j\neq i} v_j(f_0^{-i}(\boldsymbol{t}^n), t_j^n) \geq \sum_{j\neq i} v_j(f_0(\boldsymbol{t}^n), t_j^n)$ then $f_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n) - a_i(\boldsymbol{t}^1,...,\boldsymbol{t}^n)$ is non-positive. Since $f_0^{-i}(\boldsymbol{t}^n)$ maximizes the sum of values over $j \in I\setminus\{i\}$, we have $\sum_{j\neq i} v_j(f_0^{-i}(\boldsymbol{t}^n), t_j^n) \geq \sum_{j\neq i} v_j(f_0(\boldsymbol{t}^n), t_j^n)$. Together with budget neutrality of the a_i-terms, this implies now that the mechanism satisfies no deficit. Utilitarian ex-post efficiency is implied by Theorem 1.

Theorem 2. *The dynamic elaboration Clarke mechanism truthfully implements in conditionally dominant strategies a utilitarian ex-post efficient outcome under pooled awareness with no deficit.*

The proof is contained the complete manuscript ([39]).

We illustrate the dynamic elaboration Clarke mechanisms with two examples. In the first example, neither agent will announce the pooled awareness level as the pooled awareness level is the join that is

strictly greater than any agent's awareness.

Example 1 (Continuation) Consider again Example 1. In the conditional dominant solution to the dynamic elaboration Clarke mechanism, agents report in the first stage respectively,

$t_1 =$

Item	Cost
a	23
b	41
Total	64

$t_2 =$

Item	Cost
b	38
c	29
Total	67

Since $\lambda(t_1) = \{a,b\}$ and $\lambda(t_2) = \{b,c\}$, agents are made aware of the join $\check{\lambda}(t_1,t_2) = \{a,b,c\}$ and are invited to report an elaborated type in $T_i^{\{a,b,c\}}$, $i \in 1,2$, respectively. Extending the example slightly, let agents report truthfully their elaborations, respectively,

$t_1' =$

Item	Cost
a	23
b	41
c	16
Total	80

$t_2' =$

Item	Cost
a	19
b	38
c	29
Total	86

To complete the example, suppose that the possible physical outcomes (at any awareness level) are that the good is produced by agent 1, the good is produced by agent 2 or the good is not produced, i.e., $X_0 = \{1,2,\emptyset\}$. Finally, let there be a third agent, a buyer, who always values the good at 100 no matter whether it comes from agent 1 or agent 2, i.e., $v_3(1,t_3) = v_3(2,t_3) = 100$ and $v_3(\emptyset,t_3) = 0$ for any t_3.

Then, the efficient decision at the updated type profile is for the good to be produced by agent 1 (at a total cost of 80). Since no player is the first to announce the joint awareness level, the $a_i(\cdot)$-term is 0 for each agent. Agent 3 is pivotal in the sense that if agent 3's valuation is not considered, the good would not be produced at all. We have for all $\boldsymbol{t}$, $f(\boldsymbol{t}) = 1$, $f^{-1}(\boldsymbol{t}) = f^{-2}(\boldsymbol{t}) = 1$, and $f^{-3}(\boldsymbol{t}) = \emptyset$. The transfers to agent 1 are $100 - 0 - (100 - 0) + 0 = 0$, to agent 2 they are $20 - (100 - 80) - 0 = 0$, and to agent 3 are $-80 - 0 + 0 = -80$. Thus, the mechanism runs a surplus of 80.[6] □

Note that in this example awareness did not improve utilitarian ex-post welfare. Nevertheless, we are able to raise awareness to the pooled awareness level with out mechanisms.

In the complete manuscript ([39]), we provide additional examples illustrating the novel a_i-terms are non-trivial.

In the complete manuscript ([39]), we study participation constraints and show that the dynamic elaboration Clarke mechanism satisfies ex ante anticipated ex post participation constraints under standard assumptions on the value functions. Finally, also in the complete manuscript ([39]) we introduce dynamic elaboration reverse second price auctions for procurement contexts and show that they implement efficient outcomes under budget balance and ex post participation constraints.

5 Related Literature

Our paper contributes to the releated recent literature on contracting under unawareness (e.g., [33, 45, 47, 3, 12, 18, 9, 4, 34, 13]). Closer to mechanism design, [35] study the seller's decision to raise bidders' awareness of characteristics before a second-price auction with entry fees. Optimal entry fees capture an

[6] Note that the mechanism does not satisfy agent 1's ex-post participation constraints. This is true of the Clarke mechanism in this context even without any unawareness. It is known that the Clarke mechanisms may not necessarily satisfy ex-post participation constraints. We will analyze participation constraints in the next section.

additional unawareness rent due to unaware bidders misperceiving their probability of winning and the price to be paid upon winning. In contrast to our setting, the auctioneer is aware of everything ex-ante. [38] studies how a decision maker can incentivize an expert to reveal novel aspects about a decision problem via an iterated revelation mechanism in which at each round the expert decides on whether or not to raise awareness of a contingency that in turn the decision maker considers when proposing a new contract that the expert can accept or reject. This bears some similarity with our dynamic direct elaboration mechanisms. Most importantly, we focus on a multi-agent setting with transferable utilities. [38] shows that iterated revelation allows for efficient outcomes to emerge.

[31] study a procurement problem with a principal and two agents who may be aware of some design flaws. Our paper differs from theirs in many respects: First, in [31] agents can raise awareness of realized design flaws, agent's private costs are independent of the design flaw, and fixing the design flaw requires a known common cost that is interim verified by an industry expert. In our model, agents report payoff types, can raise awareness of potential factors affecting payoffs, and individually elaborate how the payoffs change in light of new awareness. Second, [31] construct an efficient direct mechanism under common awareness and then argue that there is an indirect mechanism that under asymmetric awareness that gives rise to the same outcomes and incentives. This leads them to conclude that there are also corresponding equilibria in these two mechanisms. Yet, an appropriate notion of equilibrium in mechanisms under asymmetric awareness should verify equilibrium behavior w.r.t what outcomes agents anticipate and how behavior is affect by changing anticipations of outcomes during the play. Our approach is more "direct": We conduct our entire analysis of efficient conditional dominant strategy implementation in a direct mechanisms under asymmetric awareness thereby modeling every potential deviation from equilibrium behavior from the agent's point view. Finally, [31] focus on the particular but very relevant application to procurement while we consider more generally the efficient mechanism design problem under asymmetric awareness.

Acknowledgment Burkhard gratefully acknowledges financial support via ARO Contract W911NF2210282.

References

[1] T. Agotnes & N. Alechina (2014): *A logic for reasoning about knowledge of unawareness*. *Journal of Logic, Language, and Information* 23, pp. 197–217. Available at `https://doi.org/10.1007/s10849-014-9201-4`.

[2] O. Alper & W.B. Boning (2003): *Using procurement auctions in the Department of Defense*. *Center for Naval Analysis* CRM D0007515.A3. Available at `https://www.cna.org/CNA_files/PDF/D0007515.A3.pdf`.

[3] S. Auster (2003): *Asymmetric awareness and moral hazard*. *Games and Economic Behavior* 82, pp. 503–521. Available at `https://doi.org/10.1016/j.geb.2013.08.011`.

[4] S. Auster & N. Pavoni (2024): *Optimal delegation and information transmission under limited awareness*. *Theoretical Economics* 19, pp. 245–284. Available at `https://doi.org/10.3982/TE5117`.

[5] P. Bajari, R. McMillan & S. Tadelis (2008): *Auctions versus negotiations in procurement: An empirical analysis*. *Journal of Law, Economics & Organization* 25, pp. 372–399. Available at `https://doi.org/10.1093/jleo/ewn002`.

[6] P. Bajari & S. Tadelis (2001): *Incentives versus transaction costs: A theory of procurement contracts*. *RAND Journal of Economics* 32, pp. 387–407, doi:10.2307/2696361. Available at `https://www.jstor.org/stable/2696361`.

[7] G. Belardinelli & B.C. Schipper (2024): *Implicit knowledge in unawareness structures*. *Synthese* 204. Available at `https://doi.org/10.1007/s11229-024-04706-9`.

[8] A. Burrieza & A Yuste-Ginel (2021): *An awareness epistemic framework for belief, argumentation and their dynamics*. In A. Perea, editor: *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge*. Available at https://doi.org/10.48550/arXiv.2106.11495.

[9] K.S. Chung & L. Fortnow (2016): *Loopholes*. *Economic Journal* 126, pp. 1774–1797. Available at https://doi.org/10.1111/ecoj.12203.

[10] N.R. Devanur & L. Fortnow (2009): *A computational theory of awareness and decision making*. In A. Heifetz, editor: *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 99–107. Available at https://doi.org/10.1145/1562814.1562830.

[11] H. van Ditmarsch, T. French, F.R. Velazquez-Quesada & Y.N. Wang (2018): *Implicit, explicit, and speculative knowledge*. *Artificial Intelligence* 256, pp. 35–67. Available at https://doi.org/10.1016/j.artint.2017.11.004.

[12] E. Filiz-Ozbay (2012): *Incorporating unawareness into contract theory*. *Games and Economic Behavior* 76, pp. 191–194. Available at https://doi.org/10.1016/j.geb.2012.05.009.

[13] A. Francetich & B.C. Schipper (2024): *Rationalizable screening and disclosure under unawareness*. Technical Report, University of Washington, Bothell, and University of California, Davis. Available at https://faculty.econ.ucdavis.edu/faculty/schipper/unawscreen.pdf.

[14] P. Fritz & H. Lederman (2015): *Standard state-space models of unawareness*. In R. Ramanujam, editor: *Proceedings of the 15. Conference on Theoretical Aspects of Rationality and Knowledge*. Available at https://doi.org/10.48550/arXiv.1606.07520.

[15] S. Galanis (2011): *Syntactic foundation for unawareness of theorems*. *Theory and Decision* 71, pp. 593–614. Available at https://doi.org/10.1007/s11238-010-9218-3.

[16] S. Galanis (2013): *Unawareness of theorems*. *Economic Theory* 52, pp. 41–73. Available at https://doi.org/10.1007/s00199-011-0683-x.

[17] V.P. Goldberg (1977): *Competitive bidding and the production of precontract information*. *Bell Journal of Economics* 8, pp. 250–261, doi:10.2307/3003497. Available at https://www.jstor.org/stable/3003497.

[18] S. Grant, J. Kline & J. Quiggin (2012): *Differential awareness, ambiguity, and incomplete contracts: A model of contractual disputes*. *Journal of Economic Behavior and Organization* 82, pp. 494–504. Available at https://doi.org/10.1016/j.jebo.2012.02.021.

[19] J. Green & J.-J. Laffont (1979): *Incentives in public decision-making*. North-Holland, Amsterdam.

[20] T. Groves (1973): *Incentives in teams*. *Econometrica* 41, pp. 617–631, doi:10.2307/1914085. Available at https://www.jstor.org/stable/1914085.

[21] T. Groves & M. Loeb (1975): *Incentives for public inputs*. *Journal of Public Economics* 4, pp. 211–226. Available at https://doi.org/10.1016/0047-2727(75)90001-8.

[22] J. Halpern & L.C Rêgo (2008): *Interactive unawareness revisited*. *Games and Economic Behavior* 62, pp. 232–262. Available at https://doi.org/10.1016/j.geb.2007.01.012.

[23] J. Halpern & L.C Rêgo (2013): *Reasoning about knowledge of unawareness revisited*. *Mathematical Social Sciences* 66, pp. 73–84. Available at https://doi.org/10.1016/j.mathsocsci.2012.08.003.

[24] J. Halpern & L.C. Rêgo (2014): *Extensive games with possibly unaware players*. *Mathematical Social Sciences* 70, pp. 42–58. Available at https://doi.org/10.1016/j.mathsocsci.2012.11.002.

[25] A. Heifetz, M. Meier & B.C. Schipper (2006): *Interactive unawareness*. *Journal of Economic Theory* 130, pp. 78–94. Available at https://doi.org/10.1016/j.jet.2005.02.007.

[26] A. Heifetz, M. Meier & B.C. Schipper (2008): *A canonical model of interactive unawareness*. *Games and Economic Behavior* 62, pp. 304–324. Available at https://doi.org/10.1016/j.geb.2007.07.003.

[27] A. Heifetz, M. Meier & B.C. Schipper (2013): *Dynamic unawareness and rationalizable behavior*. *Games and Economic Behavior* 81, pp. 100–121. Available at https://doi.org/10.1016/j.geb.2013.04.003.

[28] A. Heifetz, M. Meier & B.C. Schipper (2013): *Unawareness, beliefs, and speculative trade*. Games and Economic Behavior 77, pp. 100–121. Available at https://doi.org/10.1016/j.geb.2012.09.003.

[29] A. Heifetz, M. Meier & B.C. Schipper (2021): *Prudent rationalizability in generalized extensive-form games with unawareness*. B.E. Journal of Theoretical Economics 21, pp. 525–556. Available at https://doi.org/10.1515/bejte-2020-2013.

[30] S. Heinsalu (2012): *Equivalence of the information structure with unawareness to the logic of awareness*. Journal of Economic Theory 147, pp. 2453–2468. Available at https://doi.org/10.1016/j.jet.2012.05.010.

[31] F. Herweg & K. Schmidt (2020): *Procurement with unforeseen contingencies*. Management Science 66, pp. 2194–2212. Available at https://doi.org/10.1287/mnsc.2019.3290.

[32] K. Konolige (1986): *What awareness isn't: A sentential view of implicit and explicit belief*. In J. Halpern, editor: Proceedings of the 1st Conference on Theoretical Aspects of Reasoning and Knowledge, Morgan Kaufmann, pp. 241–250, doi:10.1016/B978-0-934613-04-0.50019-3. Available at https://dl.acm.org/doi/10.5555/1029786.1029807.

[33] J. Lee (2008): *Unforeseen contingencies and renegotiation with asymmetric information*. Economic Journal 118, pp. 678–694. Available at https://doi.org/10.1111/j.1468-0297.2008.02137.x.

[34] H. Lei & X.J. Zhao (2021): *Delegation and information disclosure with unforeseen contingencies*. B.E. Journal of Theoretical Economics 21, pp. 637–656. Available at https://doi.org/10.1515/bejte-2018-0184.

[35] Y.X. Li & B.C. Schipper (2024): *Raising bidders' awareness in second-price auctions*. Technical Report, University of California, Davis. Available at https://faculty.econ.ucdavis.edu/faculty/schipper/unawauction.pdf.

[36] Office of the Federal Register (2024): *Federal Acquisition Regulation: Prohibition on the use of reverse auctions for complex, specialized, or substantial design and construction services (FAR Case 2023–003)*. Federal Register 89, pp. 70157–70160.

[37] E. Ozbay (2008): *Unawareness and strategic announcements in games with uncertainty*. Technical Report, University of Maryland. Available at https://doi.org/10.1145/1324249.1324280.

[38] E. Piermont (2024): *Iterated revelation: How to incentive experts to complete incomplete contracts*. Technical Report, Royal Holloway. Available at https://doi.org/10.48550/arXiv.2304.05142.

[39] K. Pram & B.C. Schipper (2025): *Efficient mechanisms under unawareness*. Technical Report, University of Nevada, Reno, and University of California, Davis. Available at https://faculty.econ.ucdavis.edu/faculty/schipper/mechunaw.pdf.

[40] C. Proietti, F.R. Velazquez-Quesada & A. Yuste-Ginel (2023): *An abstract look at awareness models and their dynamics*. In R. Verbrugge, editor: Theoretical Aspects of Rationality and Knowledge 2023, pp. 455–469. Available at https://arxiv.org/pdf/2307.05049.

[41] B.C. Schipper: *The Unawareness Bibliography*. https://faculty.econ.ucdavis.edu/faculty/schipper/unaw.htm.

[42] B.C. Schipper (2021): *Discovery and equilibrium in games with unawareness*. Journal of Economic Theory 198, p. 105365. Available at https://doi.org/10.1016/j.jet.2021.105365.

[43] B.C. Schipper (2024): *Interactive awareness of unawareness*. Technical Report, University of California, Davis. Available at https://faculty.econ.ucdavis.edu/faculty/schipper/awunaw.pdf.

[44] Y. Shoham & K. Leyton-Brown (2012): *Multiagent systems. Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, Cambridge, M.A.

[45] S.C. Sommer & C.H. Loch (2009): *Incentive contracts in projects with unforeseeable uncertainty*. Production and Operations Management 18, pp. 185–196. Available at https://doi.org/10.1111/j.1937-5956.2009.01015.x.

[46] J. Sweet (1994): *Legal aspects of architecture, engineering and the construction process*. West Publishing Company, St. Paul.

[47] E.-L. von Thadden & X.J. Zhao (2012): *Incentives for unaware agents*. *Review of Economic Studies* 79, pp. 1151–1174. Available at `https://doi.org/10.1093/restud/rdr050`.

More Reasoning, Less Outcomes: A Monotonicity Result for Reasoning in Dynamic Games

Andrés Perea
Maastricht University
The Netherlands
Department of Quantitative Economics

a.perea@maastrichtuniversity.nl

A focus function in a dynamic game describes, for every player and each of his information sets, the collection of opponents' information sets he reasons about. Every focus function induces a rationalizability procedure in which a player believes, whenever possible, that his opponents choose rationally at those information sets he reasons about. Under certain conditions, we show that if the players start reasoning about more information sets, then the set of outcomes induced by the associated rationalizability procedure becomes smaller or stays the same. This result does not hold on the level of strategies, unless the players only reason about present and future information sets. The monotonicity result enables us to derive existing theorems, such as the relation in terms of outcomes between forward and backward induction reasoning, but also paves the way for new results.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 14–14, doi:10.4204/EPTCS.437.2

© Andrés Perea
This work is licensed under the
Creative Commons Attribution License.

Solving Four Open Problems about Core Stability in Altruistic Hedonic Games

Jörg Rothe
Institut für Informatik, MNF
Heinrich-Heine-Universität Düsseldorf
Düsseldorf, Germany
rothe@hhu.de

Ildikó Schlotter
Institute of Economics
HUN-REN Centre for Economic and Regional Studies
Budapest, Hungary
schlotter.ildiko@krtk.hun-ren.hu

Hedonic games—at the interface of cooperative game theory and computational social choice—are coalition formation games in which the players have preferences over the coalitions they can join. Kerkmann *et al.* [13] introduced *altruistic* hedonic games where the players' utilities depend not only on their own but also on their friends' valuations of coalitions. The complexity of the verification problem for core stability has remained open in four variants of altruistic hedonic games: namely, for the variants with average- and minimum-based "equal-treatment" and "altruistic-treatment" preferences. We solve these four open questions by proving the corresponding problems coNP-complete; our reductions rely on rather intricate gadgets in the related networks of friends.

1 Introduction

Hedonic games are at the interface of two areas: *cooperative game theory*—in particular, they are special coalition formation games (CFGs)—and *computational social choice*—as the players in a hedonic game have preferences over the coalitions (i.e., subsets of the players) they can join. The ultimate goal of a hedonic game (and of CFGs in general) is to form a coalition structure (i.e., a partition of the player set into coalitions) that is stable in some sense. To study which coalition structures are likely to form, various stability notions have been introduced among which *core stability* perhaps is the most natural and most central one. Suppose that there is a coalition structure Γ in a hedonic game but a group C of players are not satisfied with their coalitions in Γ: Instead, they would prefer leaving their current coalition in Γ so as to form a new coalition together. In such a case, we say that C *blocks* Γ, and a coalition structure is said to be *core-stable* if it is not blocked by any coalition. For more background on hedonic games, on core stability as well as other stability concepts for them, and on the computational complexity of the verification and existence problems for hedonic games with respect to various stability notions, we refer to the book chapters by Aziz and Savani [2] and Bullinger *et al.* [4] and to the survey by Woeginger [24].

Specifically, we will study *altruistic* hedonic games, which were introduced by Kerkmann *et al.* [13] (see also its predecessor by Nguyen *et al.* [17]). As noted by them (and previously already in the survey by Rothe [22]), their work was inspired in part by the—apparently unrelated—work of biologists: Hare and Woods [9] complement Darwin's celebrated thesis of *"survival of the fittest"* with a novel insight into how evolution works, put forward as their thesis of *"survival of the friendliest."* Indeed, they collect data and arguments (e.g., by comparing bonobos and chimpanzees, two species of great apes) showing that evolutionary success can also arise from a *friendlier* behavior. Transferring this idea to game theory, *altruistic* behavior in games can surpass aggressive selfishness that only cares about maximizing one's own utilities, regardless of the impact on others, especially so when a group of players need to coalesce. As surveyed by Rothe [22], a variety of altruistic models have been studied for noncooperative and, to a lesser extent, cooperative games, and Kerkmann *et al.* [13, 17] were the first to study altruism in hedonic

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 15–29, doi:10.4204/EPTCS.437.3

© J. Rothe & I. Schlotter
This work is licensed under the
Creative Commons Attribution License.

games. Later on, Schlueter and Goldsmith [23] extended their idea when they studied *"super-altruistic hedonic games,"* and Kerkmann *et al.* [11, 14] transferred their ideas and models to CFGs in general.

If game theory aims at modeling the behavior of players in the real world, thus providing a theoretical framework to realistically support their decision-making, it must take altruistic behavior into account. Indeed, the models of Kerkmann *et al.* [11, 13, 14, 17] are based on the *friend-oriented preference extension* due to Dimitrov *et al.* [7] where players divide the other players into friends and enemies, yielding a network of friends (a simple, undirected graph whose edges represent friendship relations among the players) that allows for a compact representation of hedonic games.[1]

Unlike in the model of Dimitrov *et al.* [7], however, in the models of Kerkmann *et al.* [11, 13, 14, 17] the players' utilities from a coalition are not solely determined by their own valuations; rather, also the (average or minimum) valuations of their friends are taken into account. Specifically, three degrees of altruism are defined that differ depending on the order in which players refer to their own or their friends' valuations: In the *selfish-first model* (SF), the players first look at their own and then at their friends' valuations—the latter only if they themselves are indifferent; in the *equal-treatment model* (EQ), they treat their own and their friends' valuations equally at the same time; and in the *altruistic-treatment* model (AL), the players first consider their friends' valuations, and only in the case of indifference they decide according to their own valuation which coalition they prefer.

The above papers provide long lists of related work. Instead of repeating them here, we just highlight some of the work that is most closely related to ours because it also studies core stability in hedonic games, such as the papers by Banerjee *et al.* [3], Dimitrov *et al.* [7], Alcade and Romero-Medina [1], Woeginger [25], Peters [20], Ohta *et al.* [18], and Chen *et al.* [6]. Bullinger and Kober [5] introduced the notion of *loyalty in cardinal hedonic games*, and their loyal variant of symmetric friend-oriented hedonic games is nothing other than the minimum-based altruistic hedonic games under EQ preferences.

For average- and minimum-based SF altruistic hedonic games, Kerkmann *et al.* [13] have shown that it is coNP-complete to verify whether a given coalition structure is core-stable, and Kerkmann *et al.* [11] showed coNP-completeness of the verification problem for average- and minimum-based SF altruistic CFGs, leaving these questions open for the other two degrees of altruism: average- and minimum-based EQ and AL preferences in both altruistic hedonic games and, more generally, altruistic CFGs. Note that EQ and AL preferences are particularly interesting, as they are "more altruistic" than SF preferences. Recently, Hoffjan *et al.* [10] solved these open questions for altruistic CFGs, again showing coNP-completeness of the verification problem for average- and minimum-based EQ and AL preferences.[2]

We solve all four questions that remained open along this line of research: We show that for average- and minimum-based EQ and AL altruistic hedonic games, the verification problem also is coNP-complete. Our proofs are based on constructing rather involved gadgets in the networks of friends of these games.

2 Preliminaries

The goal of a *coalition formation game* (CFG) is to partition a finite set N of players into *coalitions*, i.e., subsets of N, yielding a *coalition structure*, i.e., a partition of N; the *set of all possible coalition*

[1] Representing hedonic games *compactly* is important, as each of n players can join 2^{n-1} coalitions, and all players need to express their preferences on all coalitions they can join. For a large variety of ways of succinctly representing hedonic games, we refer to, e.g., the book chapters by Aziz and Savani [2] and Bullinger *et al.* [4] and also the work of Kerkmann *et al.* [12].

[2] In fact, Kerkmann *et al.* [11] and Hoffjan *et al.* [10] consider *sum-based* (not average-based) SF, EQ, and AL preferences. However, these are equivalent to (and a bit simpler than) average-based SF, EQ, and AL preferences for altruistic CFGs, as each player has the same number of friends, no matter which two coalition structures are being compared, so the denominator in the average can simply be omitted, leaving just the sum of their friends' valuations; cf. (1) in Section 2.

structures over N is denoted by $\mathscr{C}_N$. Given a coalition structure $\Gamma \in \mathscr{C}_N$, the *coalition containing player i* is denoted by $\Gamma(i)$. We are interested in studying hedonic games, which constitute a special type of CFGs. A *hedonic game* $(N, \succeq)$ is specified by the player set N and a preference profile $\succeq = (\succeq_1, \ldots, \succeq_n)$, where $n = |N|$ and $\succeq_i \subseteq \mathscr{N}^i \times \mathscr{N}^i$ is player i's preference relation (i.e., a complete, weak order) over $\mathscr{N}^i = \{C \subseteq N \mid i \in C\}$, the set of all coalitions containing i.

Specifically, for any two coalitions $C, D \in \mathscr{N}^i$, we write $C \succeq_i D$ to mean that *player i weakly prefers C to D*; we write $C \succ_i D$ to mean that *i prefers C to D* (i.e., $C \succeq_i D$ and not $D \succeq_i C$); and we write $C \sim_i D$ to mean that *i is indifferent between C and D* (i.e., $C \succeq_i D$ and $D \succeq_i C$). Since the players' preferences only depend on the coalitions containing them, $\succeq_i$ also induces a weak preference ranking of player $i \in N$ over the coalition structures in $\mathscr{C}_N$: For $\Gamma, \Delta \in \mathscr{C}_N$, $\Gamma \succeq_i \Delta$ if and only if $\Gamma(i) \succeq_i \Delta(i)$. For a hedonic game $(N, \succeq)$, a partition $\Gamma \in \mathscr{C}_N$, and a coalition $C \subseteq N$, let $\Gamma_{C \to \emptyset}$ denote the coalition structure that results from Γ when all players i in C leave $\Gamma(i)$ to form a new coalition, C, while all other players remain in their coalition in Γ.[3] Formally,

$$\Gamma_{C \to \emptyset} = \{C\} \cup \{C' \setminus C \mid C' \in \Gamma, C' \not\subseteq C\}.$$

We say that *a coalition C blocks* Γ if $C \succ_i \Gamma(i)$ for all players $i \in C$, and Γ is *core-stable* if no nonempty coalition blocks Γ.

Kerkmann *et al.* [13] (see also [17]) introduced *altruistic hedonic games* (AHGs), which have later been generalized by Kerkmann *et al.* [11] (see also [14]) to *altruistic coalition formation games* (ACFGs). Both AHGs and ACFGs are based on the *friend-oriented preference extension* due to Dimitrov *et al.* [7], which allows for a compact representation: Every player divides the other players into friends and enemies, which yields a *network of friends*—a simple, undirected graph $G = (N, \mathscr{E})$ whose vertices are the players and whose edges represent mutual friendship relations, whereas a missing edge indicates that these two players are enemies. For $i \in N$, let $F_i = \{j \in N \mid \{i, j\} \in \mathscr{E}\}$ be the set of *i's friends* and $E_i = N \setminus (F_i \cup \{i\})$ be the set of *i's enemies*, and define *i's friend-oriented valuation for a coalition* $C \subseteq \mathscr{N}^i$ by

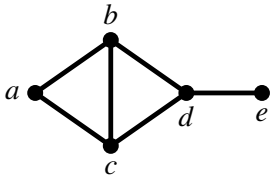
$$\mathrm{val}_i(C) = n \cdot |F_i \cap C| - |E_i \cap C|;$$

recall that $n = |N|$. Hence, for any two coalitions $C, D \in \mathscr{N}^i$, *player i friend-orientedly prefers C to D*, denoted as $C \succ_i^F D$, exactly if C contains more friends of i than D, or in case C and D contain the same number of players among i's friends, if C contains fewer enemies of i than D. Note that $C \succ_i^F D$ if and only if $\mathrm{val}_i(C) > \mathrm{val}_i(D)$.

In the altruistic model of Kerkmann *et al.* [13], the players' utilities from a coalition are not only determined by their own valuations, but also the (average or minimum) valuations of their friends in the same coalition are taken into account. Specifically, they define the following three degrees of altruism:

- In the *selfish-first model* (SF), the players' preferences mainly depend on their own valuations, consulting their friends' valuations only if they are indifferent between two coalitions;
- in the *equal-treatment model* (EQ), the players weigh their own and their friends' valuations equally; and
- in the *altruistic-treatment* model (AL), the players first consider their friends' valuations, consulting their own valuations only in the case of indifference between two coalitions.

[3] The notation $\Gamma_{X \to Y}$ usually denotes the coalition structure obtained from Γ when a set X of players joins some coalition Y; hence, our notation $\Gamma_{C \to \emptyset}$ reflects the intuition that establishing a new coalition C can be thought of as all members of C joining the empty coalition $\emptyset$.



	a	b,c	d	e
$\mathrm{val}_i(\Gamma)$	8	14	14	2
$\mathrm{util}_i^{\text{avg-SF}}(\Gamma)$	$8w+14$	$14w+12$	$14w+10$	$2w+14$
$\mathrm{util}_i^{\text{avg-EQ}}(\Gamma)$	12	12.5	11	8
$\mathrm{util}_i^{\text{avg-AL}}(\Gamma)$	$14w+8$	$12w+14$	$10w+14$	$14w+2$
$\mathrm{util}_i^{\text{min-SF}}(\Gamma)$	$8w+14$	$14w+8$	$14w+2$	$2w+14$
$\mathrm{util}_i^{\text{min-EQ}}(\Gamma)$	8	8	2	2
$\mathrm{util}_i^{\text{min-AL}}(\Gamma)$	$14w+8$	$8w+14$	$2w+14$	$14w+2$

Figure 1: Illustration for Example 1. The network of friends is depicted to the left. The table on the right contains players' valuations and utilities in the coalition structure $\Gamma = \{N\}$ under the six preference models considered.

Formally, setting the minimum of the empty set to zero by convention, define the *(friend-oriented) average* and *minimum value of player i's friends in a coalition $C \in \mathcal{N}^i$ (without and with i)* by

$$\mathrm{avg}_i^F(C) = \sum_{c \in C \cap F_i} \frac{\mathrm{val}_c(C)}{|C \cap F_i|} \quad \text{and} \quad \mathrm{avg}_i^{F+}(C) = \sum_{c \in (C \cap F_i) \cup \{i\}} \frac{\mathrm{val}_c(C)}{|(C \cap F_i) \cup \{i\}|}, \tag{1}$$

$$\mathrm{min}_i^F(C) = \min_{c \in C \cap F_i} \{\mathrm{val}_c(C)\} \quad \text{and} \quad \mathrm{min}_i^{F+}(C) = \min_{c \in (C \cap F_i) \cup \{i\}} \{\mathrm{val}_c(C)\}. \tag{2}$$

Now, define player i's average-based and minimum-based *utilities* from a coalition $C \in \mathcal{N}^i$ in an AHG according to the above three degrees of altruism:

$$\mathrm{util}_i^{\text{avg-SF}}(C) = w \cdot \mathrm{val}_i(C) + \mathrm{avg}_i^F(C) \quad \text{and} \quad \mathrm{util}_i^{\text{min-SF}}(C) = w \cdot \mathrm{val}_i(C) + \mathrm{min}_i^F(C),$$
$$\mathrm{util}_i^{\text{avg-EQ}}(C) = \mathrm{avg}_i^{F+}(C) \quad \text{and} \quad \mathrm{util}_i^{\text{min-EQ}}(C) = \mathrm{min}_i^{F+}(C), \text{ and}$$
$$\mathrm{util}_i^{\text{avg-AL}}(C) = \mathrm{val}_i(C) + w \cdot \mathrm{avg}_i^F(C) \quad \text{and} \quad \mathrm{util}_i^{\text{min-AL}}(C) = \mathrm{val}_i(C) + w \cdot \mathrm{min}_i^F(C),$$

where $w \geq n^4$ is a constant weight on some of the valuations that ensures that the SF utility is first determined by i's valuation and the AL utility is first determined by i's friends' valuations [13]. Again, since players due to their hedonism care only about the coalitions containing them, their utilities from coalitions immediately induce their *utilities from coalition structures*: $\mathrm{util}_i^{\text{avg-SF}}(\Gamma) = \mathrm{util}_i^{\text{avg-SF}}(\Gamma(i))$, etc.

Depending on which aggregation method and which degree of altruism is used, we denote our games as *average-based* or *min-based SF*, *EQ*, or *AL AHGs*. We study the computational complexity of the *verification problem* for them:

VERIFICATION-OF-CORE-STABILITY-IN-ALTRUISTIC-HEDONIC-GAMES	
Given:	An AHG I and a coalition structure Γ.
Question:	Is Γ core-stable for I?

Example 1. Consider a game I containing a set $N = \{a,b,c,d,e\}$ of players whose network of friends is depicted in Figure 1. Note that b and c have the same friends and enemies in N, and thus have the same valuations and utilities as long as they are contained in the same coalition.

The coalition $C = \{a,b,c,d\}$ yields the valuations and utilities for its players as shown in Table 1, where we fix some constant weight $w \geq n^4 = 5^4 = 625$ ensuring the right order for SF and AL utilities. Consider the coalition structure $\Gamma = \{N\}$ containing only the grand coalition N, which consists of all players. Observe that C blocks Γ in both the minimum-based EQ and AL models, but does not block Γ under the remaining four models.

	a	b,c	d	Does C block $\{N\}$?
$\text{val}_i(C)$	9	15	9	
$\text{util}_i^{\text{avg-SF}}(C)$	$9w+15$ ✔	$15w+11$ ✔	$9w+15$ ✗	no
$\text{util}_i^{\text{avg-EQ}}(C)$	13 ✔	12 ✗	13 ✔	no
$\text{util}_i^{\text{avg-AL}}(C)$	$15w+9$ ✔	$11w+15$ ✗	$15w+9$ ✔	no
$\text{util}_i^{\text{min-SF}}(C)$	$9w+15$ ✔	$15w+9$ ✔	$9w+15$ ✗	no
$\text{util}_i^{\text{min-EQ}}(C)$	9 ✔	9 ✔	9 ✔	yes
$\text{util}_i^{\text{min-AL}}(C)$	$15w+9$ ✔	$9w+15$ ✔	$15w+9$ ✔	yes

Table 1: The valuations and utilities of players in the coalition $C = \{a,b,c,d\}$. The sign ✔ means that the given player $i \in N$ prefers C to N under the given preference model, i.e., $C \succ_i N$, while the sign ✗ means that $N \succ_i C$.

This shows that Γ is *not* in the core of I if I is interpreted as a min-based EQ or AL AHG. By contrast, it is not hard to verify that Γ is in the core of I if I is interpreted as an average-based SF, EQ, or AL AHG or a min-based SF AHG.

Since our AHGs are compactly represented by their (undirected) friendship graphs $G = (N, \mathscr{E})$, we also need some graph-theoretic notation. For any subset $M \subseteq N$ of the vertices of G, let $G[M]$ denote the *subgraph of G induced by M*. For each vertex $i \in N$, let $\delta(i) \subseteq \mathscr{E}$ denote the set of edges incident to i in G. Note that $\delta(i) = \{\{i,j\} \mid j \in F_i\}$ gives player i's friendship relations in the game.

Referring to some standard textbooks [8, 19, 21], we assume the reader to be familiar with the basic notions of computational complexity theory, such as the complexity classes P (*deterministic polynomial time*), NP (*nondeterministic polynomial time*), and coNP $= \{\overline{L} \mid L \in \text{NP}\}$ (the *class of complements of* NP *problems*). Our reductions used to show coNP-hardness are based on the standard *polynomial-time many-to-one reducibility*, and problems that are coNP-hard and in coNP are said to be coNP-*complete*.

3 Core Stability in AHGs

Each of our reductions is from the NP-complete problem CLIQUE where we are given a graph $H = (V,E)$ with an integer k, and the question is whether H contains a clique of size k, i.e., a complete graph on k vertices, as a subgraph. Before delving into the details of our proofs, let us sketch a high-level description of the approach shared by all of them.

Given an instance (H,k) of CLIQUE with input graph $H=(V,E)$, each of our reductions defines an AHG I and a coalition structure Γ for I so that H has a clique of size k if and only if Γ is *not* core-stable for I. These reductions have a common approach: They construct *vertex*, *edge*, and *incidence gadgets* reflecting the structure of H. Namely, for each vertex $v \in V$, for each edge $e \in E$, and for each pair (v,e) where v is a vertex incident to some edge e in H, we define a certain *gadget*. Although vertex, edge, and incidence gadgets may differ from each other, each of these gadgets contains a distinguished player that represents the given vertex $v \in V$, edge $e \in E$, or pair $(v,e) \in \{(v,e \mid v \in V, e \in \delta(v)\}$; this player is called the *vertex*, *edge*, or *incidence player* corresponding to v, to e, or to (v,e), respectively.

The importance of distinguished players comes from the fact that, in each gadget, only the distinguished player has friends outside the gadget. As each of the gadgets forms a coalition in the coalition structure Γ for the constructed AHG, the properties of the gadgets will ensure that from among all players in the constructed gadgets, only distinguished players might be willing to form a coalition C that blocks Γ. For this to happen, the distinguished players need not only to have enough friends within C but, due to the altruistic preference models we use, their friends in C also need a certain number of friends within C. Moreover, some of the distinguished players will only be able to achieve a better utility value in C than in Γ if they have fewer enemies in C than they have within their gadget; hence, such players will impose an upper bound on the size of C.

Therefore, the computational intractability of finding a blocking coalition is caused by the need to ensure a lower bound on the number of friends for certain players in the blocking coalition C, while also respecting an upper bound on the size of C. Translating this into graph-theoretic terminology, the task is to find a subgraph in the network of friends that satisfies certain lower bounds on vertex degrees while respecting an upper bound on the number of vertices in the subgraph. The following observation offers a flexible way to deduce the presence of a clique from such requirements.

Observation 3.1. *If a simple graph $H=(V,E)$ satisfies $|V|+\alpha|E| \leq k+\alpha\binom{k}{2}$ for some $\alpha \geq 1$ and each of its vertices has degree at least $k-1$, then H is a clique of size k.*

Proof. Let v be any vertex in the graph. Since v has degree at least $k-1$, the graph contains at least k vertices. Since each of them has degree at least $k-1$, the number of edges is at least $\frac{k(k-1)}{2}$, so we get $|V|+\alpha|E| \geq k+\alpha\binom{k}{2}$. Thus all of these inequalities must hold with equality, so the number of edges must be *exactly* $\binom{k}{2}$, and the number of vertices must be *exactly* k. This means that the graph is indeed a clique of size k. □

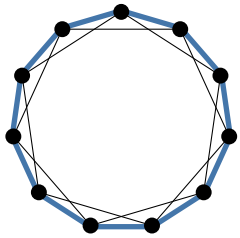
3.1 Min-Based EQ and AL AHGs

We show that verifying core stability in min-based AHGs is coNP-complete, using the same construction for both min-based EQ and min-based AL preferences.

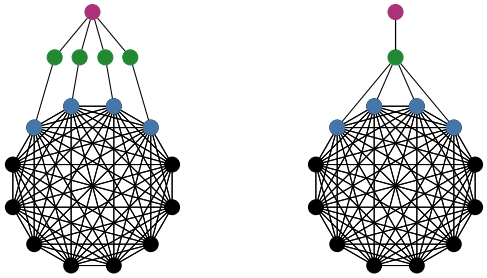
Theorem 3.2. *Verifying core stability in a min-based EQ or AL AHG is* coNP*-complete.*

Proof. Both problems clearly are in coNP because one can verify in polynomial time that a given coalition is blocking. To prove their coNP-hardness, we present a reduction from CLIQUE to their complements. Let our input for CLIQUE be the graph $H=(V,E)$ and integer k. If k is even, then we add an additional vertex connected to every vertex of H; the obtained graph has a clique of size $k+1$ if and only if H contains a clique of size k. Therefore, without loss of generality, we may assume that k is odd.

Construction. We construct from (H,k) a min-based EQ or AL AHG over player set N and an underlying friendship graph G. Let us start by defining a $(k-1,k')$*-circulant gadget* for $k'=k\binom{k}{2}+k+1$ as



(a) The friendship graph of a $(k-1,k')$-circulant gadget for $k=5$ with $k'=11$ players, with the base cycle shown with bold, blue lines.



(b) The friendship graphs of a normal (i.e., non-pinched) and a pinched (d,k')-dome gadget for $d=4$ with $k'=17$ players. The top, mid, and fringe players are depicted as purple, green, and blue circles, respectively.

Figure 2: Illustration of our gadgets in the proofs of Theorems 3.2, 3.6, and 3.8.

follows: It contains k' players arranged along a cycle Q of length k', and each player i on Q is friends with those $k-1$ players on Q who are at a distance of at most $\frac{k-1}{2}$ away from i along Q (not including i itself); see Figure 2a for an illustration. We refer to Q as the *base cycle* of the gadget.

We now introduce vertex, edge, and incidence gadgets; each of these gadgets will be a $(k-1,k')$-circulant gadget. For each vertex $v \in V$, we add a *vertex gadget* over player set P_v, with a special *vertex player* $v' \in P_v$ corresponding to the vertex v. Similarly, for each edge $e \in E$, we introduce an *edge gadget* over player set P_e, containing the *edge player* $e' \in P_e$ corresponding to the edge e. Next, for each vertex $v \in V$ and for each edge $e \in \delta(v)$, we introduce an *incidence gadget* over player set $P_{v,e}$, containing the *incidence player* $b_{v,e}$. Additionally, we define a set A_e of $k-3$ *dummy players* for each edge $e \in E$. We will use the notation $V' = \{v' \mid v \in V\}$, $E' = \{e' \mid e \in E\}$, $B = \{b_{v,e} \mid v \in V,\ e \in \delta(v)\}$, and $A = \bigcup_{e \in E} A_e$. The total set of players is

$$N = \bigcupdot_{v \in V} P_v \cup \bigcupdot_{e \in E} (P_e \cup A_e) \cup \bigcupdot_{v \in V,\, e \in \delta(v)} P_{v,e}.$$

Besides the friendships within gadgets, we let each incidence player $b_{v,e} \in B$ be friends with the vertex player v' and the edge player e', so $G[V' \cup E' \cup B]$ can be obtained from H by subdividing each of its edges once. Next, for each edge $e = xy \in E$, we let the dummy players in A_e be friends with players $e', b_{x,e}$, and $b_{y,e}$. Finally, for each edge $e \in E$, we let $G[A_e]$ be a clique of size $k-3$, so all players in A_e are friends with each other. This completes the definition of our AHG.

Let us define the coalition structure Γ as

$$\Gamma = \{P_v \mid v \in V\} \cup \{P_e \mid e \in E\} \cup \{P_{v,e} \mid v \in V, e \in \delta(v)\} \cup \{A_e \mid e \in E\},$$

so Γ contains the player set of each vertex, edge, and incidence gadget as well as the set of dummy players associated with any edge as a coalition. Note that, for $n = |N|$, each player i's valuation of Γ is

$$\mathrm{val}_i(\Gamma) = \begin{cases} (k-1)\cdot n - (k'-k) & \text{if } i \in N \setminus A, \\ (k-4)\cdot n & \text{if } i \in A. \end{cases}$$

Therefore, both in the min-based EQ and AL models, the utility of each player equals its valuation.

Proof of correctness. We claim that Γ is *not* core-stable if and only if H contains a clique of size k.

Only if: Let us first assume that Γ is *not* core-stable; we show that H contains a clique of size k. The following claim captures the key property of our gadgets.

Claim 3.3. *Let P be a set of players of a $(k-1,k')$-circulant gadget in the constructed instance whose unique player having friends outside P is $p^\star$ (a vertex, edge, or incidence player). If some coalition $C \subseteq N$ blocks Γ, then $C \cap P \subseteq \{p^\star\}$.*

Proof: Let us show first that each player $i \in C \cap (P \setminus \{p^\star\})$ must have at least $k-1$ friends in C. Suppose for the sake of contradiction that i has at most $k-2$ friends in C. On the one hand, if i has a friend $j \in C$, then $j \in C \cap P$ and, for $\varphi \in \{\text{min-EQ}, \text{min-AL}\}$, the utility of j in $\Gamma_{C \to \emptyset}$ is at most the valuation of i:

$$\text{util}_j^\varphi(\Gamma_{C\to\emptyset}) \leq \text{val}_i(\Gamma_{C\to\emptyset}) \leq (k-2)\cdot n < \text{util}_j^\varphi(\Gamma),$$

which contradicts $j \in C$. On the other hand, if i has no friends in C, then $\text{util}_i^\varphi(\Gamma_{C\to\emptyset}) = 0 < \text{util}_i^\varphi(\Gamma)$, which contradicts $i \in C$.

Now, to see the statement of the claim, assume for the sake of contradiction that C contains some player $i \in P \setminus \{p^\star\}$. Let Q denote the base cycle of the $(k-1,k')$-circulant gadget with players P. If $P \not\subseteq C$ were true, then consider the path in $Q[C]$ that contains i (a subpath of the base cycle); let player j be one endpoint of this path that does not coincide with $p^\star$ (such a player j exists because $i \neq p^\star$). Then j has fewer than $k-1$ friends in C because one of j's friends, though adjacent to j on Q, is not in C. This is not possible by the argument given in the previous paragraph, so $P \subseteq C$ follows.

To see that $P \subseteq C$ is not possible either, consider some player i in $P \setminus \{p^\star\}$ that is not a friend of $p^\star$; by $k' > k$, such a player exists. Assuming that $P \subseteq C$, we get that i and all friends of i have valuation exactly $(k-1)n - (|C| - k)$ in $\Gamma_{C\to\emptyset}$, which is at most $(k-1)n - (k'-k) = \text{util}_i^\varphi(\Gamma)$ by $|C| \geq k'$. This contradicts $i \in P \subseteq C$. ◁

We will also need the following simple claim.

Claim 3.4. *Let $C \subseteq N$ be a coalition blocking Γ. Each player $i \in C$ has at least $k-1$ friends in C.*

Proof: First, if some player i has a non-dummy friend j in C (i.e., a friend $j \in C \setminus A$), then i must have at least $k-1$ friends in C, as otherwise, for $\varphi \in \{\text{min-EQ}, \text{min-AL}\}$, we have $\text{util}_j^\varphi(\Gamma_{C\to\emptyset}) \leq \text{val}_i(\Gamma_{C\to\emptyset}) \leq (k-2)n < \text{util}_j^\varphi(\Gamma)$ which contradicts $j \in C$. In particular, since dummy players in C must have non-dummy friends in C (as $C \subseteq A$ is not possible), and each dummy player has $k-1$ friends in total, we get that whenever $C \cap A_e \neq \emptyset$ for some $e = xy \in E$, then C must contain all players in $\{e', b_{x,e}, b_{y,e}\} \cup A_e$.

To prove the claim, it remains to show that every non-dummy player i in C has a non-dummy friend in C. First, every player in C must have some friend in C, as otherwise this player's utility is zero in $\Gamma_{C\to\emptyset}$. For the sake of contradiction, assume that some non-dummy player $i \in C$ has only dummy friends in C. Then, by construction, i must be among the players in $\{e', b_{x,e}, b_{y,e}\}$ for some edge $e = xy \in E$ for which $C \cap A_e \neq \emptyset$. Since e' is friends with both $b_{x,e}$ and $b_{y,e}$, it follows that i has a friend in $\{e', b_{x,e}, b_{y,e}\}$. However, by the argument given in the previous paragraph, $C \cap A_e \neq \emptyset$ implies $\{e', b_{x,e}, b_{y,e}\} \subseteq C$, showing that i has a non-dummy friend in C. This contradiction proves the claim. ◁

Let $C \subseteq N$ be a coalition that blocks Γ. Using Claim 3.3, we obtain that $C \subseteq V' \cup E' \cup B \cup A$. Let us create the subgraph H_C of H that contains a given vertex or edge if and only if the corresponding vertex or edge player is contained in C. We now show that H_C is a well-defined subgraph of H. First, recall that each incidence player $b_{v,e}$ has exactly $k-1$ friends in $V' \cup E' \cup B \cup A$, namely the players v', e', and the $k-3$ players in A_e; therefore, if $b_{v,e}$ is contained in C, then due to Claim 3.4 all friends of $b_{v,e}$ must be in C as well. Keeping this in mind, assume that $e' \in C$ for some edge $e = xy \in E$. Since e' also has exactly $k-1$ friends in $V' \cup E' \cup B \cup A$, namely those in A_e together with the two incidence player $b_{x,e}$

and $b_{y,e}$, all of them must be in C by Claim 3.4. However, $\{b_{x,e}, b_{y,e}\} \subseteq C$ in turn implies $\{x', y'\} \subseteq C$, so H_C is indeed well-defined.

Using again Claim 3.4, we know that each vertex in H_C has degree at least $k-1$. To see this, note that $x' \in C \cap V'$ implies that there are at least $k-1$ incidence-player friends of x' in C, together with their edge-player friends, which means that there are at least $k-1$ edges incident to x in H_C. To prove that H_C is a clique of size k, we will use Observation 3.1.

By our previous arguments, $e' \in C \cap E'$ implies that all $k-1$ friends of e' in $B \cup A$ are also in C. Since no player in $B \cup A$ has two edge players as a friend, we get

$$|C| = |C \cap V'| + |C \cap E'| + |C \cap (B \cup A)| \geq |C \cap V'| + k|C \cap E'|.$$

Consider now an edge player e' in C. Since every friend of e' in C, as well as e' itself, has exactly $k-1$ friends in C, the utility of e' is $\mathrm{util}^{\varphi}_{e'}(\Gamma_{C \to \emptyset}) = (k-1)n - (|C| - k)$ for $\varphi \in \{\text{min-EQ}, \text{min-AL}\}$, which exceeds $\mathrm{util}^{\varphi}_{e'}(\Gamma)$ only if $|C| < |k'|$. Thus $|C| \leq k\binom{k}{2} + k$, which implies $|C \cap V'| + k|C \cap E'| \leq k + k\binom{k}{2}$. Applying Observation 3.1 with $\alpha = k$, it follows that H_C is a clique of size k.

If: Let us now show that if H contains a clique on a set K of k vertices, then there exists a coalition $C \subseteq N$ that blocks Γ. Let C contain the vertex and edge players corresponding to the vertices and edges of the clique K, together with all friends of these edge players in $A \cup B$, so $|C| = k + k\binom{k}{2}$. The utility of each player in $\Gamma_{C \to \emptyset}$ is then $(k-1)n - (|C| - k)$, which is more than its utility in Γ, because $|C| = k + k\binom{k}{2} < k'$. Therefore, C blocks Γ, and so Γ is *not* core-stable. □

3.2 Average-Based EQ and AL AHGs

To show the coNP-completeness of verifying core stability in an average-based EQ or AL AHG, we introduce a new type of gadgets. For some integers $d \geq 1$ and $k' > 2d+1$, we define a (d,k')-*dome gadget* as follows: It contains k' players, among them a *top player* $p^\star$ with friends $p_1, \ldots, p_d$ that we call *mid players*, and with the remaining $k'-d-1$ players forming a clique in the friendship graph that we call the *base clique*. Additionally, for each mid player p_i, we select a player p'_i in the base clique in a way that the players $p'_1, \ldots, p'_d$, called *fringe players*, are all distinct, and we let p_i be friends with p'_i.

We further introduce a modification of this gadget: Let a *pinched* (d,k')-*dome* be obtained from a (d,k')-dome gadget by the identification of all d mid players into a single mid player.

See Figure 2b for an illustration of such gadgets. The following claim captures their key property.

Proposition 3.5. *Suppose that an instance of an average-based EQ or AL AHG over n players contains a (possibly pinched) (d,k')-dome gadget on player set P for some integers d and k' satisfying $n \geq k'(d+1)$ and $k' > 2d+3$. Assume further that no player in P except for its top player $p^\star$ has a friend outside P. If $P \in \Gamma$ is a coalition in some coalition structure Γ, and a coalition C blocks Γ, then C contains no player from the base clique of the gadget.*

Proof. Let $k'' = k' - d - 1$ denote the size of the base clique K in our gadget over P. First observe that if some non-fringe player i within the base clique K is in C, then all k'' players in the base clique are in C, as otherwise each friend of i (and i itself) loses a friend in C when compared to P, and thus the average number of friends among i's friends decreases by at least 1. The average number of enemies may only decrease by at most $d+1$, so $n > d+1$ implies $\mathrm{util}^{\varphi}_i(\Gamma_{C \to \emptyset}) < \mathrm{util}^{\varphi}_i(\Gamma)$, a contradiction to $i \in C$. In fact, $i \in C$ further implies that each player in the base clique K must have the same number of friends in C as in Γ, as otherwise the average number of friends among i's friends (possibly counting also i itself) decreases by at least $\frac{1}{k''}$, while the average number of enemies of these players can only decrease by at

most $d+1$, so $\frac{n}{k''} > d+1$ implies $\text{util}_i^{\varphi}(\Gamma_{C\to\emptyset}) < \text{util}_i^{\varphi}(\Gamma)$. Thus we have $P \setminus \{p^\star\} \subseteq C$. The presence of some fringe player j in C further implies $p^\star \in C$ by the same arguments, since $p^\star \notin C$ would mean a decrease of at least $\frac{1}{k''+1}$ in the utility of j. Therefore, we get $P \subseteq C$; however, then the utility of every player in the base clique is at most its utility in Γ, which contradicts our assumption that C blocks Γ.

We thus have proven that C cannot contain non-fringe players in the base clique of P. However, from this it follows that no fringe player i of P can be contained in C, as its utility in $\Gamma_{C\to\emptyset}$ would be less than its utility in Γ: Observe that all fringe-player friends of i (if any) lose at least one friend when switching from Γ to $\Gamma_{C\to\emptyset}$ (because there are $k'-2d-1>0$ non-fringe players in the base clique K), while the mid-player friend of i has at most $d+1$ friends, which is strictly fewer than the $k''-1 \geq d+2$ friends of the base clique players in Γ. Thus, no player in K can be contained in C. □

We are now ready to show coNP-completeness of verifying core stability in an AHG with either avg-EQ or avg-AL preferences, starting with the former.

Theorem 3.6. *Verifying core stability in an average-based EQ AHG is* coNP*-complete.*

Proof. Membership of the verification problem in coNP again is obvious. To show its coNP-hardness, we again reduce from CLIQUE. Let (H,k) be our input, where $H=(V,E)$ is a graph and k an integer. We may assume, without loss of generality, that $k \geq 4$ and that $|V|+|E| \geq k^2$. Let $k' = k+3\binom{k}{2}+1$. The reduction will, to some degree, be similar to the one given in the proof of Theorem 3.2, but we need to carefully modify the construction and proof of correctness to make it work for avg-EQ preferences.

Construction. We construct vertex, edge, and incidence gadgets as follows. For each vertex $v \in V$, we add a $(k-1,k')$-dome gadget over player set P_v whose top player is the *vertex player* v' corresponding to v. For each edge $e \in E$, we add a $(2,k')$-dome gadget over player set P_e whose top player is the *edge player* e' corresponding to e. Further, for each $v \in V$ and $e \in \delta(v)$, we introduce a $(2,k')$-dome gadget over player set $P_{v,e}$ whose top player is the *incidence player* $b_{v,e}$. Let n denote the total number of players in the constructed gadgets, and we will also use the notation $V' = \{v' \mid v \in V\}$, $E' = \{e' \mid e \in E\}$, and $B = \{b_{v,e} \mid v \in V, e \in \delta(v)\}$. Besides the friendships within gadgets, we let each incidence player $b_{v,e} \in B$ be friends with the vertex player v' and the edge player e', so the friendship graph induced by $V' \cup E' \cup B$ can be obtained from H by subdividing each of its edges once. This completes the definition of our average-based EQ AHG.

Let Γ be the coalition structure that contains the player set of each vertex, edge, and incidence gadget as a coalition; formally,

$$\Gamma = \{P_v \mid v \in V\} \cup \{P_e \mid e \in E\} \cup \{P_{v,e} \mid v \in V, e \in \delta(v)\}.$$

Due to Proposition 3.5, we will not be interested in non-fringe base players, so we only need to compute the valuation of the remaining players in Γ, which are as follows:

$$\text{val}_i(\Gamma) = \begin{cases} (k-1)\cdot n - (k'-k) & \text{if } i \text{ is a vertex player,} \\ 2\cdot n - (k'-3) & \text{if } i \text{ is an edge, an incidence, or a mid player,} \\ (k'-k)\cdot n - (k-1) & \text{if } i \text{ is a fringe player in a vertex gadget,} \\ (k'-3)\cdot n - 2 & \text{if } i \text{ is a fringe player in an edge or incidence gadget.} \end{cases}$$

It is now straightforward to compute the utilities of the players in Γ in the average-based EQ model; we provide these for top and mid players below:

$$\text{util}_i^{\text{avg-EQ}}(\Gamma) = \begin{cases} \frac{3k-3}{k}\cdot n - \left(k'-1-\frac{3k-3}{k}\right) & \text{if } i \in V', \\ 2\cdot n - (k'-3) & \text{if } i \in E' \cup B, \\ \frac{k'+1}{3}\cdot n - (k'-1-\frac{k'+1}{3}) & \text{if } i \text{ is a mid player in } \Gamma(i). \end{cases}$$

Proof of correctness. We claim that Γ is *not* core-stable if and only if H contains a clique of size k.

Only if: Let us first assume that Γ is *not* core-stable; we show that H contains a clique of size k. Let C be a coalition blocking Γ. Observe that C can contain no mid players from an edge or incidence gadget: By Proposition 3.5, a mid player i can have only one friend in C, namely the top player in $\Gamma(i)$, while the top player in such a gadget has four friends in total; hence, the utility of i in C can be at most $\frac{5}{2}n < \text{util}_i^{\text{avg-EQ}}(\Gamma)$, a contradiction to $i \in C$. Therefore, by Proposition 3.5, C can only contain players in $V' \cup E' \cup B$, and possibly some mid players from vertex gadgets.

Next, we show an analogue of Claim 3.4.

Claim 3.7. *If $e' \in C$ for some edge $e = xy \in E$, then $\{b_{x,e}, b_{y,e}, x', y'\} \subseteq C$ and $|C| < k'$.*

Proof: The friends of e' in C can only be the two incidence players $b_{x,e}$ and $b_{y,e}$, and they both can have at most two friends in C (besides e', one of the vertex players x' and y'). Thus the utility of e' in $\Gamma_{C\to\emptyset}$ can exceed $\text{util}_{e'}^{\text{avg-EQ}}(\Gamma)$ only if $\{b_{x,e}, b_{y,e}\} \subseteq C$ and, moreover, both $b_{x,e}$ and $b_{y,e}$ have two friends in C, which leads to $\{x', y'\} \subseteq C$. Therefore, the utility of e' in $\Gamma_{C\to\emptyset}$ is exactly $\text{util}_{e'}^{\text{avg-EQ}}(\Gamma_{C\to\emptyset}) = 2n - (|C| - 3)$, which exceeds $\text{util}_{e'}^{\text{avg-EQ}}(\Gamma)$ if and only if $|C| < k'$. ◁

Let us create a subgraph H_C of H that contains a given vertex or edge if and only if the corresponding vertex or edge player is contained in C. By Claim 3.7, H_C is a well-defined subgraph of H.

We proceed by showing that every vertex v in H_C has at least $k-1$ incident edges. Intuitively, the reason for this is that adding friends with only one friend in the coalition cannot raise the utility of a player. For a formal proof, assume for the sake of contradiction that v has at most $k-2$ incident edges. Then v' has at most $k-2$ friends with two friends being in C, and v' may have an additional number ℓ of friends, each of whom has only v' as a friend in C. Then we get

$$\begin{aligned}\text{util}_{v'}^{\text{avg-EQ}}(\Gamma_{C\to\emptyset}) &\leq \frac{2(k-2)+(k-2+\ell)+\ell}{k+\ell-1}\cdot n = \left(3 - \frac{\ell+3}{k+\ell-1}\right)\cdot n \\ &< \left(\frac{3k-3}{k} - \frac{1}{k^2}\right)\cdot n < \frac{3k-3}{k}\cdot n - \left(k' - 1 - \frac{3k-3}{k}\right) = \text{util}_{v'}^{\text{avg-EQ}}(\Gamma),\end{aligned}$$

where the first strict inequality holds because $k \geq 4$ and $\ell \geq 0$, as can be checked through simple calculation, and the second strict inequality follows from our assumptions on the size of the graph H that guarantees $n \geq k^2 \cdot k'$. Therefore, we obtain a contradiction to $v' \in C$, proving that each vertex in H_C has degree at least $k-1$.

To show that C is a clique, we use Observation 3.1. Recall that $|C| < k'$ due to Claim 3.7. Since each incidence player has only one edge-player friend, but each edge player in C has two incidence-player friends in C, we get

$$k + 3\binom{k}{2} = k' - 1 \geq |C| = |C \cap V| + |C \cap E| + |C \cap B| \geq |C \cap V| + 3|C \cap E|. \tag{3}$$

Applying Observation 3.1 with $\alpha = 3$, it follows that H_C is a clique of size k.

If: Let us now show that if H contains a clique on a set K of k vertices, then there exists a coalition C that blocks Γ. Similarly as in the proof of Theorem 3.2, let C contain the vertex and edge players corresponding to the vertices and edges of the clique K, together with all friends of these edge players in B, so $|C| = k + 3\binom{k}{2}$. In the new coalition structure $\Gamma_{C\to\emptyset}$ (after the members of C have deviated from

their coalitions in Γ), the utility of each player i in C is as follows:

$$\text{util}_i^{\text{avg-EQ}}(\Gamma_{C\to\emptyset}) = \begin{cases} \frac{3k-3}{k}\cdot n - \left(|C| - 1 - \frac{3k-3}{k}\right) & \text{if } i \in V', \\ 2\cdot n - (|C|-3) & \text{if } i \in E', \\ \frac{k+3}{3}\cdot n - \left(|C| - 1 - \frac{k+3}{3}\right) & \text{if } i \in B. \end{cases}$$

Since $|C| = k + 3\binom{k}{2} < k'$ and $k \geq 4$, it follows that $\text{util}_i^{\text{avg-EQ}}(\Gamma_{C\to\emptyset}) < \text{util}_i^{\text{avg-EQ}}(\Gamma)$ for each player i in C. Therefore, C blocks Γ, and so Γ is *not* core-stable. □

Finally, we turn to showing that verifying core stability in an average-based AL AHG is coNP-complete. The proof of Theorem 3.8 is similar to the proof Theorem 3.6 and relies on pinched dome-gadgets.

Theorem 3.8. *Verifying core stability in an average-based AL AHG is* coNP*-complete.*

Proof. Again, membership of the verification problem in coNP is obvious and, to show its coNP-hardness, we present a reduction from CLIQUE. The proof will be quite similar to that of Theorem 3.6: We use the same ideas and an analogous construction, however, the details need to be carefully adjusted. Let (H,k) be our input instance of CLIQUE, where $H = (V,E)$ is a graph and k an integer. Set $k' = k + 3\binom{k}{2} + 1$. Without loss of generality, we assume that $|V| + |E| > k \geq 3$ and that $\frac{k+1}{3}$ is an integer.

Construction. We construct vertex, edge, and incidence gadgets as follows. For each vertex $v \in V$, we introduce a pinched $(2,k')$-dome gadget over player set P_v whose top player is the *vertex player* v' corresponding to v. For each edge $e \in E$, we introduce a pinched $(2,k')$-dome gadget over player set P_e whose top player is the *edge player* e' corresponding to e. Additionally, for each vertex $v \in V$ and each edge $e \in \delta(v)$, we introduce a pinched $\left(\frac{k+1}{3},k'\right)$-dome gadget over player set $P_{v,e}$ whose top player is the *incidence player* $b_{v,e}$. We let n denote the total number of players in the constructed gadgets, and we will use the notation $V' = \{v' \mid v \in V\}$, $E' = \{e' \mid e \in E\}$, and $B = \{b_{v,e} \mid v \in V, e \in \delta(v)\}$. Besides the friendships within gadgets, we let each incidence player $b_{x,e} \in B$ for some edge $e = xy \in E$ be friends with the vertex player x', the edge player e', and—in addition—the incidence player $b_{y,e}$. This completes the definition of our AHG.

Let Γ be the coalition structure that contains the player set of each vertex, edge, and incidence gadget as a coalition; formally,

$$\Gamma = \{P_v \mid v \in V\} \cup \{P_e \mid e \in E\} \cup \{P_{v,e} \mid v \in V, e \in \delta(v)\}.$$

Due to Proposition 3.5, we will only be interested in the utilities of the top and mid players within each gadget in Γ. Recall that the mid player in a pinched (d,k')-gadget has $d+1$ friends in the gadget. Simple calculation and $k \geq 3$ gives us the following:

$$\text{util}_i^{\text{avg-AL}}(\Gamma) = \begin{cases} 3\cdot n - (k'-4) & \text{if } i \in V' \cup E', \\ \frac{k+4}{3}\cdot n - \left(k' - 1 - \frac{k+4}{3}\right) & \text{if } i \in B'; \end{cases}$$

$$\text{util}_i^{\text{avg-AL}}(\Gamma) \geq \quad \frac{2k'-3}{3}\cdot n - \left(k' - 1 - \frac{2k'-3}{3}\right) \quad \text{if } i \text{ is a mid player in } \Gamma(i).$$

Proof of correctness. We claim that Γ is *not* core-stable if and only if H contains a clique of size k.

Only if: Assume first that some coalition C blocks Γ. Observe that a blocking coalition C can contain no mid players from an edge or incidence gadget: By Proposition 3.5, a mid player i can have

only one friend in C, namely the top player in $\Gamma(i)$, while the top player in such a gadget has at most four friends in total; hence, the utility of i in C can be at most $4n < \frac{2k'-4}{3} \cdot n < \text{util}_i^{\text{avg-AL}}(\Gamma)$, a contradiction to $i \in C$. Therefore, by Proposition 3.5, C can only contain players in $V' \cup E' \cup B$, and possibly some mid players from vertex gadgets.

We define the graph H_C as in the proof of Theorem 3.6, i.e., H_C contains a given vertex or edge of H if and only if the corresponding vertex or edge player is in C.

Next, we show that Claim 3.7 remains true for the modified construction.

Claim 3.9. *If $e' \in C$ for some edge $e = xy \in E$, then $\{b_{x,e}, b_{y,e}, x', y'\} \subseteq C$ and $|C| < k'$.*

Proof: The friends of e' in C can only be the two incidence players $b_{x,e}$ and $b_{y,e}$, and both of these players can have at most three friends in C. If one of $b_{x,e}$ and $b_{y,e}$ is not in C, then the utility of e' in $\Gamma_{C \to \emptyset}$ could be at most $2n < \text{util}_{e'}^{\text{avg-AL}}(\Gamma)$. Hence, we get $\{b_{x,e}, b_{y,e}\} \subseteq C$; moreover, both $b_{x,e}$ and $b_{y,e}$ must have three friends in C, which leads to $\{x', y'\} \subseteq C$. Therefore, the utility of e' in $\Gamma_{C \to \emptyset}$ is exactly $\text{util}_{e'}^{\text{avg-AL}}(\Gamma_{C \to \emptyset}) = 3n - (|C| - 4)$, which exceeds $\text{util}_{e'}^{\text{avg-AL}}(\Gamma)$ if and only if $|C| < k'$. ◁

By Claim 3.9, we know that H_C is a well-defined subgraph of H. We next prove that every vertex in H_C has degree at least $k-1$.

Assume $x' \in C$. First notice that each friend of x' has at most three friends in C. Thus all friends of x' in C need to have *exactly* three friends in C, as otherwise the utility of x' in $\Gamma_{C \to \emptyset}$ would be lower than $\frac{2+3(\ell-1)}{\ell} \cdot n = 3n - \frac{n}{\ell} < 3n - k' < \text{util}_{x'}^{\text{avg-AL}}(\Gamma)$ where ℓ denotes the number of friends that x' has in C; notice that the inequality $n > k' \cdot \ell$ we use here follows from the fact that x must have degree at least $\ell - 1$ in H. In particular, no mid player in the vertex gadget containing x' can be in C. Furthermore, whenever $b_{x,e} \in C$ for some edge e incident to x in H, then $e' \in C$ follows. This means that the degree of x in H_C is exactly the number of friends that x' has in C.

Consider now the utility of some incidence player $b_{x,e}$ in C: If x' has at most $k-2$ friends in C, then

$$\text{util}_{b_{x,e}}^{\text{avg-AL}}(\Gamma_{C \to \emptyset}) \leq \frac{2+3+k-2}{3} \cdot n < \frac{k+4}{3} \cdot n - k' < \text{util}_{b_{x,e}}^{\text{avg-AL}}(\Gamma)$$

where the first strict inequality follows from $n > 3k'$ which in turn holds due to our assumptions on the size of the graph H. This proves that x' needs to have at least $k-1$ friends in C, that is, every vertex in H_C has degree at least $k-1$. Finally, observe that the inequality (3) holds; using also Claim 3.9, we can apply Observation 3.1 with $\alpha = 3$. It follows that H_C is a clique of size k.

If: Given a clique of size k in H, proving that the corresponding vertex, edge, and incidence players form a coalition blocking Γ is a straightforward adaptation of the arguments presented in the proof of Theorem 3.6. □

4 Conclusions and Open Questions

Having solved the last four open problems related to the computational complexity of verifying core stability in altruistic hedonic games, the picture for this property is now complete: For all three degrees of altruism, in both the average-based and the minimum-based case, and for both altruistic hedonic games and the more general altruistic CFGs, it is coNP-complete to verify whether a given coalition structure is core-stable. Of course, many related problems remain open and can be tackled in future research.

For example, we only have an upper bound of containment in coNP for verifying *strict* core stability, and it remains to show a matching lower bound of coNP-hardness. A coalition structure Γ is *strictly*

core-stable if it is not weakly blocked by any coalition, i.e., for each coalition $C \subseteq N$, we either have $\Gamma(i) \succ_i \Gamma_{C\to\emptyset}(i)$ for some player $i \in C$, or we have $\Gamma(i) \sim_i \Gamma_{C\to\emptyset}(i)$ for all players $i \in C$. Furthermore, the existence problems for core stability and strict core stability have not been classified in terms of their complexity yet, in any of the models of AHGs or ACFGs we have considered here.

In addition, the existence problems for other properties of AHGs or ACFGs remain open as well, such as the existence of a (strictly) popular coalition structure, even though the corresponding verification problems have recently been settled for AHGs [15]. A coalition structure Γ is *popular* if it is preferred to any other coalition structure Δ by at least as many players as there are players preferring Δ to Γ.

Finally, in addition to classical complexity, it would be very interesting to study these problems in terms of their fixed-parameter tractability and parameterized complexity (see, e.g., the work of Chen *et al.* [6]) or in terms of their approximability (see, e.g., the work of Munagala *et al.* [16] who study core stability in the context of participatory budgeting).

Acknowledgments

This work was supported in part by Deutsche Forschungsgemeinschaft under DFG research grant RO-1202/21-2 (project 438204498) and by the Hungarian Academy of Sciences under its Momentum Programme (LP2021-2) and its János Bolyai Research Scholarship.

References

[1] J. Alcade & A. Romero-Medina (2006): *Coalition Formation and Stability*. *Social Choice and Welfare* 27(2), pp. 365–375, doi:10.1007/s00355-006-0123-1.

[2] H. Aziz & R. Savani (2016): *Hedonic Games*. In F. Brandt, V. Conitzer, U. Endriss, J. Lang & A. Procaccia, editors: *Handbook of Computational Social Choice*, chapter 15, Cambridge University Press, pp. 356–376, doi:10.1017/CBO9781107446984.016.

[3] S. Banerjee, H. Konishi & T. Sönmez (2001): *Core in a Simple Coalition Formation Game*. *Social Choice and Welfare* 18(1), pp. 135–153, doi:10.1007/s003550000067.

[4] M. Bullinger, E. Elkind & J. Rothe (2024): *Cooperative Game Theory*. In J. Rothe, editor: *Economics and Computation. An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, 2nd edition, chapter 3, Classroom Companion: Economics, Springer, pp. 139–229, doi:10.1007/978-3-031-60099-9_3.

[5] M. Bullinger & S. Kober (2021): *Loyalty in Cardinal Hedonic Games*. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, pp. 66–72, doi:10.24963/ijcai.2021/10.

[6] J. Chen, G. Csáji, S. Roy & S. Simola (2023): *Hedonic Games With Friends, Enemies, and Neutrals: Resolving Open Questions and Fine-Grained Complexity*. In: *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, IFAAMAS, pp. 251–259. Available at https://dl.acm.org/doi/abs/10.5555/3545946.3598644.

[7] D. Dimitrov, P. Borm, R. Hendrickx & S. Sung (2006): *Simple Priorities and Core Stability in Hedonic Games*. *Social Choice and Welfare* 26(2), pp. 421–433, doi:10.1007/s00355-006-0104-4.

[8] M. Garey & D. Johnson (1979): *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, doi:10.5555/574848.

[9] B. Hare & V. Woods (2020): *Survival of the Friendliest: Understanding Our Origins and Rediscovering Our Common Humanity*. Penguin Random House. Available at https://www.penguinrandomhouse.com/books/550437/survival-of-the-friendliest-by-brian-hare-and-vanessa-woods/.

[10] M. Hoffjan, A. Kerkmann & J. Rothe (2024): *Core Stability in Altruistic Coalition Formation Games*. In: *Proceedings of the 16th Latin American Theoretical Informatics Symposium*, Springer-Verlag *Lecture Notes in Computer Science #14579*, pp. 320–333, doi:10.1007/978-3-031-55601-2_21.

[11] A. Kerkmann, S. Cramer & J. Rothe (2024): *Altruism in Coalition Formation Games*. *Annals of Mathematics and Artificial Intelligence* 92(3), pp. 601–629, doi:10.1007/s10472-023-09881-y.

[12] A. Kerkmann, J. Lang, A. Rey, J. Rothe, H. Schadrack & L. Schend (2020): *Hedonic Games with Ordinal Preferences and Thresholds*. *Journal of Artificial Intelligence Research* 67, pp. 705–756, doi:10.1613/jair.1.11531.

[13] A. Kerkmann, N. Nguyen, A. Rey, L. Rey, J. Rothe, L. Schend & A. Wiechers (2022): *Altruistic Hedonic Games*. *Journal of Artificial Intelligence Research* 75, pp. 129–169, doi:10.1613/jair.1.13706.

[14] A. Kerkmann & J. Rothe (2020): *Altruism in Coalition Formation Games*. In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, pp. 347–353, doi:10.24963/ijcai.2020/49.

[15] A. Kerkmann & J. Rothe (2024): *The Complexity of Verifying Popularity and Strict Popularity in Altruistic Hedonic Games*. *Journal of Autonomous Agents and Multi-Agent Systems* 38(2):47, doi:10.1007/s10458-024-09679-0.

[16] K. Munagala, Y. Shen & K. Wang (2022): *Auditing for Core Stability in Participatory Budgeting*. In: *Proceedings of the 18th International Workshop on Internet & Network Economics*, Springer-Verlag *Lecture Notes in Computer Science #13778*, pp. 292–310, doi:10.1007/978-3-031-22832-2_17.

[17] N. Nguyen, A. Rey, L. Rey, J. Rothe & L. Schend (2016): *Altruistic Hedonic Games*. In: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, IFAAMAS, pp. 251–259. Available at https://dl.acm.org/doi/10.5555/2936924.2936963.

[18] K. Ohta, N. Barrot, A. Ismaili, Y. Sakurai & M. Yokoo (2017): *Core Stability in Hedonic Games among Friends and Enemies: Impact of Neutrals*. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, pp. 359–365, doi:10.24963/ijcai.2017/51.

[19] C. Papadimitriou (1995): *Computational Complexity*, second edition. Addison-Wesley.

[20] D. Peters (2017): *Precise Complexity of the Core in Dichotomous and Additive Hedonic Games*. In: *Proceedings of the 5th International Conference on Algorithmic Decision Theory*, Springer-Verlag *Lecture Notes in Artificial Intelligence #10576*, pp. 214–227, doi:10.1007/978-3-319-67504-6_15.

[21] J. Rothe (2005): *Complexity Theory and Cryptology. An Introduction to Cryptocomplexity*. EATCS Texts in Theoretical Computer Science, Springer-Verlag, doi:10.1007/3-540-28520-2.

[22] J. Rothe (2021): *Thou Shalt Love Thy Neighbor as Thyself When Thou Playest: Altruism in Game Theory*. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, AAAI Press, pp. 15070–15077, doi:10.1609/aaai.v35i17.17770.

[23] J. Schlueter & J. Goldsmith (2020): *Super Altruistic Hedonic Games*. In: *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference*, AAAI Press, pp. 160–165. Available at https://aaai.org/papers/160-flairs-2020-18422/.

[24] G. Woeginger (2013): *Core Stability in Hedonic Coalition Formation*. In: *Proceedings of the 39th International Conference on Current Trends in Theory and Practice of Computer Science*, Springer-Verlag *Lecture Notes in Computer Science #7741*, pp. 33–50, doi:10.1007/978-3-642-35843-2_4.

[25] G. Woeginger (2013): *A Hardness Result for Core Stability in Additive Hedonic Games*. *Mathematical Social Sciences* 65(2), pp. 101–104, doi:10.1016/j.mathsocsci.2012.10.001.

A Modular Equilibrium Theory for Admissibility in Games

Abstract

Admissibility constitutes the most common tool for refining Nash equilibrium predictions in normal-form games. Curiously, equilibrium notions that exclude weakly dominated strategies (e.g. perfect equilibrium) strictly refine the combined predictions of corresponding non-equilibrium solution concepts (e.g. admissibility) and Nash equilibrium. We use an epistemic approach to cautious reasoning in games to understand the nature of this "extra refinement". It emerges that the gap between equilibrium refinements and their non-equilibrium counterparts stems from a strong requirement that players be fully lexicographically correct about each others' cautious theories. This is precisely what drives perfect equilibrium to strictly refine the intersection of admissibility and Nash equilibrium. More seriously, full lexicographic correctness also causes an iteratively admissible refinement of perfect equilibrium to be impossible for almost all normal-form games. This is true even though any game possesses a Nash equilibrium in iteratively admissible strategies. As an alternative to full lexicographic correctness, a notion of primary lexicographic correctness is developed. This epistemic condition smoothly combines with cautious non-equilibrium solution concepts. Primary lexicographic correctness together with common assumption of rationality provide an epistemic foundation for a novel solution concept called iteratively admissible equilibrium that we propose.

Christian W. Bach
c.w.bach@reading.ac.uk

Stephan Jagau
sjagau@uci.edu

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 30–30, doi:10.4204/EPTCS.437.4

Optimal Information Design in Sender-Receiver Cheap Talk Interactions

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

Ivan Geffner
Utrecht University
Utrecht, Netherlands
i.e.geffnerfuenmayor@uu.nl

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

This paper considers the dynamics of cheap talk interactions between an oblivious receiver and a sender with different amounts of information. Even though it may seem that having additional information about the state of the game is always beneficial to the sender, we show that there are cases in which garbling the information of a fully informed sender can improve not only receiver's utility in equilibrium, but also that of the sender herself. We also provide efficient algorithms that output the optimal amount of information in sender-receiver scenarios with binary actions and extend some of these results to settings with multiple senders and one receiver.

1 Introduction

Consider a cheap talk interaction between a seller and a buyer in which the seller has in its possession an art piece that can be either an original (**OG**), a fake that is almost indistinguishable from an original (**IF**), or a fake easily distinguishable by an art expert (**DF**). The buyer wants to buy the piece only if it is an original. The seller wants to sell it only if it is of types **OG** or **IF**, since otherwise it is highly likely that she will eventually lose reputation for selling fakes (even if the buyer cannot tell right away). If the item is not sold, both agents get 0 utility. Instead, if the buyer buys the item, the exact utilities are shown in the following table.

Type	Seller	Buyer
OG	1	1
IF	1	-5
DF	-5	-5

Suppose that, initially, the item is equally likely of being **OG**, **IF** or **DF**. Moreover, assume that the seller is an art expert, and thus can distinguish between types **OG** and **DF**, and also that she knows if the item is a fake or not from the source (which means that she can distinguish between all three types). Suppose that the buyer is an art collector, but not necessarily an art expert (i.e., the buyer has no information about the item's type besides the prior). It is easy to check that, in this scenario, the only Nash equilibrium is the one in which the buyer never buys the item, giving both agents 0 utility. To see this, note that, if the seller could convince the buyer to buy the item in some cases, she would definitely do so whenever the item is of type **IF**, giving the buyer a strictly negative expected utility.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 31–45, doi:10.4204/EPTCS.437.5

© I. Arieli, I. Geffner & M. Tennenholtz
This work is licensed under the
Creative Commons Attribution License.

Consider an identical scenario but in which the seller is **not** an art expert. This means that the seller still knows if the piece is a fake or not from the source but, whenever the item is fake, she can't tell how good of a fake it is (i.e., she can't tell if it is of type **IF** or **DF**.) Here, there is a Nash equilibrium that gives $\frac{1}{3}$ expected utility to each agent in which the seller signals if the item is an original or not, and the buyer buys it only if it is original (note that the sender does not want to sell the item whenever it is fake). Thus, in this example, the receiver is better off by not being knowledgeable about art. This raises the natural question of, given a general sender-receiver setting, determining the *optimal* amount of knowledge that the sender must have in order for her (resp., the receiver) to maximize her expected utility in equilibrium. More precisely, we start with a setting in which the sender has full information about the state of the game, and our aim is constructing an information filter that garbles the information that is disclosed to the sender in a way that the best equilibrium of the resulting game gives her (resp., the receiver) the maximum possible utility. In this paper we focus on constructing such filter in a setting with binary actions. In particular, we provide $O(k \log k)$ algorithms that output the optimal disclosure of information to the sender, where k is the number of states. If we restrict our attention to the best possible equilibrium for the receiver, we also provide an efficient algorithm when there are multiple senders that are equally knowledgeable. A perhaps surprising property of our setting is that, in the example above, the fact that both players can simultaneously increase their utility is no coincidence. In fact, we prove in our analysis that, if the sender can increase her utility by garbling her information, the utility of the receiver increases as well (see Proposition 3).

1.1 Related Literature

Restricting the information available to the sender has been increasingly gaining traction in the community. Most notably, Bergemann, Brooks and Morris [3] considered a buyer-seller setting and studied possible ways to limit the seller's information. In their work, they characterized the set of all pairs of possible utilities achievable in equilibrium. In particular, they provide an algorithm that achieves the optimal way to limit the seller's information to maximize the expected buyers' utility. Ichihashi [9] considered a Bayesian persuasion setting and studied how the outcome of the interaction is affected when the sender's information is restricted. One of their results is that, if the receiver restricts sender information in a pre-play stage, the best utility that the receiver can get in this setting coincides with the one that the receiver would get in the "flipped game", where the receiver persuades the sender. We study a similar problem, except that in our case agents have no commitment power. Other papers have studied settings with limited communication between the sender and the receiver both in the context of common-interest coordination games [4, 6] and cheap talk models [10, 8]. Restricting the information available also became quite relevant in the context of moderating large language model models in order for them to avoid producing undesirable responses in some prompts [14, 13, 7]. Our underlying model is also similar to that of Bayesian persuasion [11], especially when the sender has no commitment power [12, 5, 1, 2].

The rest of the paper is organized as follows. In Section 2 we introduce the main concepts and definitions used throughout the paper. In Section 3 we state our main results regarding the computation of the best filters in sender-receiver games with one and two senders. These results are later proved in Sections 4 and 6. We end with a conclusion in Section 7.

2 Basic Definitions

2.1 Information Transmission Games

An *information transmission game* involves a *sender* s and a *receiver* r, and is defined by a tuple $\Gamma = (A, \Omega, p, M, u)$, where $A = \{a_1, \ldots, a_\ell\}$ is the set of actions, $\Omega = \{\omega_1, \ldots, \omega_n\}$ is the set of possible states, p is a commonly known prior distribution over Ω that assigns a strictly positive probability to each possible state, M is a finite set that contains the messages that the sender can send to the receiver (M is usually assumed to be finite), and $u : \{s, r\} \times \Omega \times A \longrightarrow \mathbb{R}$ is a utility function such that $u(i, \omega, a)$ gives the utility of player i (where i is either the sender or the receiver) when action a is played at state ω. Each information transmission game instance is divided into three phases.

- **Phase 1:** A state $\omega \in \Omega$ is sampled according to the distribution p and is disclosed to the sender.
- **Phase 2:** The sender sends a message $m \in M$ to the receiver.
- **Phase 3:** The receiver plays an action $a \in A$ and each player i receives $u(i, \omega, a)$ utility.

Given an information transmission game $\Gamma = (A, \Omega, p, M, u)$, a strategy profile $\vec{\sigma}$ for Γ consists of a pair of strategies (σ_s, σ_r) for the sender and the receiver, where σ_s is a map from Ω to distributions over M (which is denoted by $\Delta(M)$), and σ_r is a map from M to $\Delta(A)$. We say that $\vec{\sigma}$ is a Nash equilibrium if no player can increase its utility by defecting from $\vec{\sigma}$. More precisely, if we denote by $u_i(\vec{\sigma})$ the expected utility that i gets when players play $\vec{\sigma}$, then $\vec{\sigma}$ is a Nash equilibrium if $u_i(\vec{\sigma}_{-i}, \tau_i) \leq u_i(\vec{\sigma})$ for all players $i \in \{s, r\}$ and all strategies τ_i for i.

2.2 Information Aggregation Games

At a high level, *information aggregation games* are information transmission games with multiple senders. For a rigorous definition, an information aggregation game consists of a receiver r, and a tuple $\Gamma = (S, A, \Omega, p, M, u)$ where $S = \{s_1, s_2, \ldots, s_k\}$ is the set of senders and $A = \{a_1, \ldots, a_\ell\}$, $\Omega = \{\omega_1, \ldots, \omega_k\}$, $p \in \Delta(\Omega)$, M and u are defined as in information transmission games, with the only exception that u is a function from $S \cup \{r\} \times \Omega \times A$ to $\mathbb{R}$ instead of a function from $\{s, r\} \times \Omega \times A$ to $\mathbb{R}$. An information aggregation game runs in the same way as an information transmission game, except that in phase 1 the state is disclosed to all senders, and in phase 2 each sender s_i sends a message m_i to the receiver.

2.3 Filtered Information Transmission Games

We are interested in settings where the sender does not have a full picture about the state of the game. In order to model this, we will assume that, in Phase 1, the state ω is not directly disclosed to the sender, but that instead it goes through an *information filter* $X : \Omega \to \Delta(\{0,1\}^*)$ that maps each state ω to a distribution over possible signals, and these signals are binary strings of arbitrary length (we use the notation $\{0,1\}^* := \bigcup_{n \geq 0} \{0,1\}^n$). The sender receives a signal sampled from $X(\omega)$ instead of ω itself. If $\Gamma = (A, \Omega, p, M, u)$ is an information transmission game and $X : \Omega \to \Delta(\{0,1\}^*)$ is a filter, we denote by $\Gamma(X)$ the resulting filtered information game where the state goes through X before being disclosed to the

sender. Filtered information aggregation games are defined analogously. For future reference, given an information transmission game Γ, we define a *sender-optimal filter* X_s^Γ as a filter in which the utility that the sender gets in the best equilibrium is maximal. We define a *receiver-optimal filter* X_r^Γ analogously.

3 Main Results

We next address the question of finding the optimal amount of information the sender must have in information transmission games and information aggregation games with binary actions.

Theorem 1. *Let $\Gamma = (A, \Omega, p, M, u)$ be an information transmission game with $A = \{0,1\}$. Then, there exists an algorithm π that outputs a receiver-optimal filter (resp., a sender-optimal filter). The algorithm runs in $O(k \log k)$ time, where $k = |\Omega|$.*

If there are two senders instead, the following result shows that we can reduce the problem of computing a receiver-optimal filter to a linear programming instance where the number of variables and constraints are linear over the number of states.

Theorem 2. *Let $\Gamma = (S, A, \Omega, p, M, u)$ be an information aggregation game with $|\Omega| = k$, $S = \{s_1, s_2\}$, and $A = \{0,1\}$. Then, finding a receiver-optimal filter reduces to a linear programming instance with k variables and $2k+2$ constraints.*

Note that, if we focus on the receiver, we only provide results for the cases of one and two senders. If $\Gamma = (S, A, \Omega, p, M, u)$ is an information aggregation game with $|S| \geq 3$ there is a trivial strategy profile $\vec{\sigma}_{MAX}$ that gives the receiver the maximal possible utility of the game with no filters. The senders simply forward the state of the game to the receiver, and the receiver plays the action that gives her the most utility on the state sent by a majority of the senders. It is easy to check that this is indeed a Nash equilibrium: the senders do not get any additional utility by defecting since the other senders will still send the true state (which means that the receiver will be able to compute the true state as well), and the receiver does not get any additional utility either since she is always playing the optimal action for each possible state. This gives the following result.

Theorem 3. *Let $\Gamma = (A, \Omega, p, M, u)$ be an information aggregation game with $|S| \geq 3$. Then, $\vec{\sigma}_{MAX}$ is a Nash equilibrium that gives the receiver the maximal possible utility of the game.*

In the following sections, we give algorithms that output the best Nash equilibria for the settings described in Theorems 1 and 2, respectively.

4 Proof of Theorem 1

In this section we provide an algorithm that outputs the receiver-optimal filter for any information transmission game $\Gamma^F = (A, \Omega, p, M, u)$ with $S = \{s\}$ and $A = \{0,1\}$. The sender-optimal filter can be computed analogously (see Section 5 for more details). We start with the following proposition:

Proposition 1. *Let $\Gamma = (A, \Omega, p, M, u)$ be an information transmission game and let $X : \Omega \to \Delta(\{0,1\}^*)$ be an information filter. Then, there exists a Pareto-optimal Nash equilibrium $\vec{\sigma}$ in $\Gamma(X)$ such that, in $\vec{\sigma}$, either:*

- *The receiver always plays 0.*
- *The receiver always plays 1.*
- *The receiver always plays the best action for the sender.*

Before proving Proposition 1 we need additional notation. First, given a filtered information transmission game $\Gamma(X)$, let $u_i(x,a)$ be the expected utility of player i on signal x and action a. This expected utility can be computed with the following equation:

$$u_i(x,a) = \sum_{\omega\in\Omega} \Pr[\omega \mid x] \cdot u_i(\omega,a),$$

where $\Pr[\omega \mid x]$ is the probability that the realized state is ω conditional on the fact that the sender received signal x. Similarly, $\Pr[\omega \mid x]$ has the following expression:

$$\Pr[\omega \mid x] = \frac{\Pr[\omega \longleftarrow p,\ x \longleftarrow X(\omega)]}{\sum_{\omega\in\Omega} \Pr[\omega \longleftarrow p,\ x \longleftarrow X(\omega)]}.$$

Let X_0 be the set of signals x in $\{X(\omega)\}_{\omega\in\Omega}$ such that $u_s(x,0) > u_s(x,1)$ (i.e., the set of signals in which the sender prefers 0), let X_1 be the set of signals such that $u_s(x,0) < u_s(x,1)$, and $X_=$ be the set signals such that $u_s(x,0) = u_s(x,1)$. Given a strategy profile $\vec{\sigma}$ for $\Gamma(X)$, denote by $\sigma_s(x,m)$ the probability that the sender sends message m given signal x and denote by $\sigma_r(a,m)$ the probability that the receiver plays action a given message m. Moreover, let $M_0^{\vec{\sigma}}$ denote the set of messages that have a strictly positive probability to be sent by the sender on at least one signal x in which the sender prefers 0 (i.e., $M_0^{\vec{\sigma}}$ is the set of messages m such that there exists $x \in X_0$ such that $u_s(x,m) > 0$). We define $M_1^{\vec{\sigma}}$ and $M_=^{\vec{\sigma}}$ analogously.

With this notation, the following lemma describes all strategy profiles in $\Gamma(X)$ that are incentive-compatible for the sender.

Lemma 1. *A strategy profile $\vec{\sigma}$ for $\Gamma(X)$ is incentive-compatible for the sender if and only if the following is satisfied:*

(a) If $m \in M_0^{\vec{\sigma}}$, then $\sigma_r(0,m) \geq \sigma_r(0,m')$ for all messages m'.

(b) If $m \in M_1^{\vec{\sigma}}$, then $\sigma_r(1,m) \geq \sigma_r(1,m')$ for all messages m'.

Lemma 1 states that, for a strategy profile to be incentive-compatible for the sender, the receiver should play 0 with maximal probability on all messages that could be sent on signals in which the sender prefers 0, and the receiver should play 0 with minimal probability on all messages that could be sent on signals in which the sender prefers 1. In particular, we have the following Corollary.

Corollary 1. *A strategy profile $\vec{\sigma}$ for $\Gamma(X)$ is incentive-compatible for the sender if and only if there exist two real numbers $\ell_{\min}, \ell_{\max} \in [0,1]$ with $\ell_{\min} \leq \ell_{\max}$ such that:*

(a) $m \in M_0^{\vec{\sigma}} \implies \sigma_r(0,m) = \ell_{\max}$.

(b) $m \in M_1^{\vec{\sigma}} \implies \sigma_r(0,m) = \ell_{\min}$.

(c) $m \in M_=^{\vec{\sigma}} \implies \ell_{\min} \leq \sigma_r(0,m) \leq \ell_{\max}$.

Proof of Lemma 1. Clearly, if (a) and (b) are satisfied, then $\vec{\sigma}$ is incentive-compatible for the sender. Conversely, suppose that $\vec{\sigma}$ is incentive-compatible for the sender but it doesn't satisfy (a). This means that there exists a signal $x \in X_0$ and a message m that satisfies $u_s(x,m) > 0$ and such that $u_r(0,m) < u_r(0,m')$ for some other message m'. Therefore, if the sender sends m' instead of m whenever it receives signal x, it could increase its expected utility. This contradicts the fact that $\vec{\sigma}$ is incentive-compatible for the sender. The proof of the case in which $\vec{\sigma}$ doesn't satisfy (b) is analogous. □

Corollary 1 characterizes the necessary and sufficient conditions for a strategy profile $\vec{\sigma}$ in $\Gamma(X)$ to be incentive-compatible for the sender. We show next that Proposition 1 follows from adding the receiver incentive-compatibility constraints into the mix. Denote by $u_r^{\vec{\sigma}}(m,0)$ the receiver's expected utility when playing action 0 conditioned on the fact that it received message m and that the sender plays σ_s. Then, we have the following cases:

Case 1: There exists $m \in M_0^{\vec{\sigma}}$ such that $u_r^{\vec{\sigma}}(m,0) < u_r^{\vec{\sigma}}(m,1)$: Since $\vec{\sigma}$ is incentive-compatible for the receiver, it must be the case that $\ell_{\min} = 1$, and therefore that $\sigma_r(0,m) = 1$ for all messages m.

Case 2: There exists $m \in M_1^{\vec{\sigma}}$ such that $u_r^{\vec{\sigma}}(m,0) > u_r^{\vec{\sigma}}(m,1)$: This time it must be the case that $\ell_{\max} = 0$, and therefore that $\sigma_r(0,m) = 0$ for all messages m.

Case 3: $u_r^{\vec{\sigma}}(m,0) \geq u_r^{\vec{\sigma}}(m,1)$ for all $m \in M_0^{\vec{\sigma}}$ and $u_r^{\vec{\sigma}}(m,0) \leq u_r^{\vec{\sigma}}(m,1)$ for all $m \in M_1^{\vec{\sigma}}$: In this case, consider a strategy profile $\vec{\sigma}'$ that is identical to $\vec{\sigma}$ except that, whenever the receiver receives a message $m \in M_0^{\vec{\sigma}}$, it plays action 0 with probability 1 (as opposed to probability $\ell_{\max}$), and when the receiver receives a message $m \in M_1^{\vec{\sigma}}$, it plays action 1 with probability 1 (as opposed to $1 - \ell_{\min}$). It is easy to check that, by construction, $\vec{\sigma}'$ is a Nash equilibrium that Pareto-dominates $\vec{\sigma}$. Even so, $\vec{\sigma}'$ can be further improved: consider a strategy profile $\vec{\sigma}^s$ such that the sender sends message 0 whenever she strictly prefers 0 to 1 or whenever she is indifferent and the receiver strictly prefers 0 to 1, and she sends message 1 otherwise. Additionally, the receiver plays the action suggested by the receiver. It is easy to check that $\vec{\sigma}^s$ is a Nash equilibrium that Pareto-dominates $\vec{\sigma}'$ since the receiver plays the best action for the sender in both cases, but in $\vec{\sigma}^s$ she also plays the best action for herself whenever the sender is indifferent while she may not necessarily do so in $\vec{\sigma}'$.

This analysis provides a refinement of Proposition 1. In fact, consider the following two strategy profiles:

- Strategy $\vec{\sigma}^r$: Regardless of the signal received, the sender signals an empty $\perp$ message. The receiver plays the action that gives her the most utility with no information.
- Strategy $\vec{\sigma}^s$: After receiving the signal, the sender signals her preferred action to the receiver. The receiver plays the action sent by the sender.

We have the following characterization of Pareto-optimal Nash equilibria:

Proposition 2 (Refinement of Proposition 1)**.** *If $\vec{\sigma}^s$ is incentive-compatible for the receiver, then $\vec{\sigma}^s$ is a Pareto-optimal Nash equilibrium of $\Gamma(X)$. Otherwise, $\vec{\sigma}^r$ is a Pareto-optimal Nash equilibrium of $\Gamma(X)$.*

Proposition 2 shows that Theorem 1 reduces to find the filter X such that the receiver (resp., the sender) maximizes her utility by playing $\vec{\sigma}^s$ (note that $\vec{\sigma}^r$ gives the same utility to both players independently of the filter applied). In the next sections we show how to efficiently compute such filters. However, before going into it, our analysis also shows the following.

Proposition 3. *If there exists a filter that strictly increases the sender's utility, it also (weakly) increases the receiver's utility.*

Proof. If $\vec{\sigma}^s$ is a Nash equilibrium of Γ, it gives the maximum possible utility to the receiver in every single state. Therefore, if there exists a filter X such that the sender gets more utility in $\Gamma(X)$ than in X, it must be that $\vec{\sigma}^r$ is a Pareto-optimal Nash equilibrium of Γ while $\vec{\sigma}^s$ is a Pareto-optimal Nash equilibrium of $\Gamma(X)$. This lemma follows from the fact that, if $\vec{\sigma}^s$ is a Nash equilibrium of $\Gamma(X)$, it gives more utility to the receiver than $\vec{\sigma}^r$. □

4.1 Computation of X_r^Γ

In this section we show how to compute a receiver-optimal filter. As discussed in Section 4, our aim is to find a filter X_r^Γ such that $\vec{\sigma}^s$ is incentive-compatible for the receiver in $\Gamma(X_r^\Gamma)$ and such that the expected utility for the receiver with $\vec{\sigma}^s$ is as large as possible. We begin by showing an algorithm that runs in $O(k\log k)$ time, where k is the number of states (i.e., the size of Ω), and then we show the proof of correctness for the algorithm provided.

4.1.1 An $O(k\log k)$ Algorithm

For our algorithm, we restrict our search to *binary* filters (i.e., filters that only send signals in $\{0,1\}$) and later, in Lemma 2, we show that this is done without loss of generality. With this restriction, we can describe possible filter candidates X by a function X_* that maps each state ω to the probability that $X(\omega)$ outputs 0. Note that, without loss of generality, we can also restrict our search to filters in which the sender prefers action 0 on signal 0 and prefers action 1 on signal 1 (otherwise we can re-label the signals). If a filter satisfies this condition, we say that it is incentive-compatible for the sender. Analogously, if the receiver prefers action 0 on signal 0 and prefers action 1 on signal 1, we say that X is incentive-compatible for the receiver. This shows that, without loss of generality, we can search for the binary filter that (a) is incentive-compatible for both the sender and the receiver, and (b) that gives the most utility for the receiver.

Before we start, let Ω^0 (resp., Ω^1) denote the set of states in which both the sender and the receiver prefer 0 (resp., both prefer 1) and let Ω_0^1 (resp., Ω_1^0) denote the set of states in which the sender strictly prefers 0 and the receiver strictly prefers 1 (resp., the sender strictly prefers 1 and the receiver strictly prefers 0). The following algorithm outputs the optimal filter X^{OPT} for the receiver.

1. **Step 1:** Set $(X_r^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega^0$.
2. **Step 2:** Set $(X_r^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega^1$.
3. **Step 3:** Sort all states $\omega \in \Omega_0^1 \cup \Omega_1^0$ according to the value $\frac{u_r(\omega,0)-u_r(\omega,1)}{u_s(\omega,1)-u_s(\omega,0)}$. Let $\omega^1, \omega^2, \ldots, \omega^{k'}$ be the resulting list.
4. **Step 4:** Set $(X_r^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega_1^0$ and $(X_r^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega_0^1$. If the resulting filter is incentive-compatible for the sender, output X_r^Γ.

5. **Step 5:** For $i = 1,2,\ldots,k'$ do the following. If $(X_r^\Gamma)_*(\omega^i)$ is set to 1 (resp., to 0), set it to 0 (resp., to 1). If the resulting filter is incentive-compatible for the sender, find the maximum (resp., the minimum) q such that setting $(X_r^\Gamma)_*(\omega^i)$ to q is incentive-compatible for the sender (we show in Section 4.1.2 that it can be computed in amortized linear time). If the resulting filter is incentive-compatible for the receiver, output X_r^Γ, otherwise output the constant filter $(X_r^\Gamma)_* \equiv 0$.

It is important to note that the algorithm always terminates, since if we get to k' in step 5, the resulting filter will always output the signal that the sender prefers. In Section 4.1.2 we show how to compute q and check the incentive-compatibility of the sender and of the receiver in amortized linear time, and in Section 4.2 we show that the algorithm indeed produces the correct output.

4.1.2 Checking Incentive-Compatibilities and Computing q in Amortized Linear Time

In this section we show how to check the players' incentive-compatibilities and how to compute q in amortized linear time. For this purpose, let d_i^s (resp., d_i^r) denote the difference in utility for the sender (resp., for the receiver) between playing 0 and playing 1. More precisely, $d_i^t := u_t(\omega^i,0) - u_t(\omega^i,1)$ for $t \in \{s,r\}$. Then, given a binary filter X, $\vec{\sigma}^s$ is incentive-compatible for the sender and the receiver in Γ_X^F if and only if

$$\begin{array}{l} \sum_i p(\omega^i) \cdot d_i^t \cdot X_*(\omega^i) \geq 0 \\ \sum_i p(\omega^i) \cdot d_i^t \cdot (1 - X_*(\omega^i)) \leq 0 \end{array} \tag{1}$$

for $t \in \{s,r\}$. The first and second equations state that the sender (resp., the receiver) prefers 0 to signal 0 and prefers 1 on signal 1, respectively.

It is straightforward to check that the incentive-compatibility equations for the sender monotonically increase and decrease, respectively, whenever we iterate in the fifth step of the algorithm. Therefore, we can simply pre-compute the following sums for $t \in \{s,r\}$, $b \in \{0,1\}$ and $i \in \{1,2,\ldots,k'\}$.

$$\begin{array}{l} Y_b^t := \sum_{\omega \in \Omega^b} p(\omega) \cdot (u_t(\omega,0) - u_t(\omega,1)) \\ W_{i,b}^t := \sum_{j<i, \omega^j \in \Omega_b^{1-b}} p(\omega^j) \cdot d_j^t \\ X_{i,b}^t := \sum_{j>i, \omega^j \in \Omega_b^{1-b}} p(\omega^j) \cdot d_j^t. \end{array}$$

Note that these sums can be computed in amortized linear time over the total number of states. Once these sums are computed, if we are in the ith iteration of Step 5, to find q we can check in constant time if the solution of any of the following linear equations is in $[0,1]$:

$$\begin{array}{l} Y_0^s + W_{i,0}^s + p(\omega^i) \cdot q \cdot d_i^s + X_{i,0}^s = 0 \\ Y_1^s + W_{i,1}^s + p(\omega^i) \cdot (1-q) \cdot d_i^s + X_{i,1}^s = 0. \end{array}$$

If there exists such a solution q, to check that it is incentive-compatible for both players, we should simply check if the equations in 1 are satisfied, which can be rewritten as $Y_0^t + W_{i,0}^t + p(\omega^i) \cdot q \cdot d_i^t + X_{i,0}^t \geq 0$ and $Y_1^t + W_{i,1}^t + p(\omega^i) \cdot (1-q) \cdot d_i^t + X_{i,1}^t \leq 0$, respectively. Note that all of these computations can be performed in constant time if the necessary sums are pre-computed. Also, the whole process runs in $O(k \log k)$ time since it takes $O(k \log k)$ operations to sort the elements of $\Omega_0^1 \cup \Omega_1^0$, $O(k)$ operations to

compute the sums, and $O(1)$ operations in each iteration of Step 5. In the next section, we show that the algorithm's output is correct.

4.2 Proof of Correctness of π

In this section, we show that the algorithm π presented in Section 4.1 is correct. We begin by showing that we can indeed restrict our search to binary filters.

Lemma 2. *Let X be a filter such that $\vec{\sigma}^s$ is incentive-compatible for the receiver in $\Gamma(X)$. Then, there exists a binary filter X' such that*

(a) $\vec{\sigma}^s$ is incentive-compatible for the receiver in $\Gamma(X')$.

(b) With strategy profile $\vec{\sigma}^s$, the expected utility of the receiver in $\Gamma(X)$ and $\Gamma(X')$ is identical.

Proof. Recall that $\vec{\sigma}^s$ is a strategy profile in which, for each possible signal x, the sender sends its preferred action and then the receiver plays whatever is sent by the sender. This means that, if $\vec{\sigma}$ is incentive-compatible for the receiver in $\Gamma(X)$, then we can merge all signals in which the sender prefers 0, and also merge all signals in which the sender prefers 1. More precisely, given filter X, let X_0 and X_1 be the sets of signals in which the sender prefers 0 and 1 respectively. Consider a filter X' that sends signal 0 whenever X would send a signal in X_0, and sends signal 1 whenever X would send a signal in X_1. In $\Gamma(X')$, by construction, both the sender and the receiver prefer action 0 on signal 0 and action 1 on signal 1. This means that $\vec{\sigma}^s$ is a Nash equilibrium of $\Gamma(X')$. Moreover, again by construction, the expected utility of the receiver (and the sender) when playing $\vec{\sigma}^s$ in $\Gamma(X')$ is identical to the one they'd get in $\Gamma(X)$. □

Recall that binary filters X can be described by a function $X_* : \Omega \longrightarrow [0,1]$ that maps each state $\omega \in \Omega$ to the probability that $X(\omega)$ assigns to 0. Because of Lemma 1, our aim is to find which values in $[0,1]$ we should assign to each element in Ω. The following lemmas characterizes these values.

Lemma 3. *There exists a filter X_r^Γ that maximizes the utility for the receiver such that*

(a) $(X_r^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega_0^0$.

(b) $(X_r^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega_1^1$.

(c) At most one state $\omega \in \Omega_0^1 \cup \Omega_1^0$ satisfies that $(X_r^\Gamma)_*(\omega) \notin \{0,1\}$.

Proof. Given a binary filter X, if we increase $X_*(\omega_i)$ by ε, the expected utility of the receiver increases by $\varepsilon \cdot p(\omega_i) \cdot d_i^r$. This means that, if $d_i^s \geq 0$ and $d_i^r \geq 0$, setting $X_*(\omega_i)$ to 1 increases the receiver's utility and preserves the incentive-compatibility constraints (see Equation 1). Analogously, the same happens by setting $X_*(\omega_i)$ to 0 when $d_i^s \leq 0$ and $d_i^r \leq 0$. This proves (a) and (b).

To prove (c) suppose that there exist two states ω_i and ω_j such that $X_*^{OPT}(\omega_i), X_*^{OPT}(\omega_j) \notin \{0,1\}$. Then, because of (a) and (b) we can assume without loss of generality that $\omega_i, \omega_j \in \Omega_0^1 \cup \Omega_1^0$. Therefore, if we increase $X_*(\omega_i)$ by $\varepsilon \cdot d_j^s$ and decrease $X_*(\omega_j)$ by $\varepsilon \cdot d_j^s$, we'd have that the sender's utility and incentive-compatibility constraints remain unchanged, but the receiver's utility increases by $\varepsilon \cdot (d_j^s d_i^r - d_i^s d_j^r)$ (note

that this value can be negative). This means that if we choose an ε that is small enough and is of the same sign as $d_j^s d_i^r - d_i^s d_j^r$, not only the expected utility of the receiver increases, but also the values $\sum_i p(\omega_i) \cdot d_i^r \cdot X_*(\omega_i)$ and $\sum_i p(\omega_i) \cdot d_i^r \cdot (1 - X_*(\omega_i))$ increase and decrease respectively. This contradicts the fact that the filter X_r^Γ is optimal for the receiver. □

Lemma 3 shows that we can assign $(X_r^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega_0^0$, $(X_r^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega_0^0$, and either probability 1 or probability 0 to all but at most one of the remaining states. Ideally, we would like to assign probability 1 to all states in Ω_0^1 and probability 0 to all states in Ω_1^0, since this guarantees that the receiver gets the maximum possible utility. However, this may not always be incentive-compatible for the sender, which means that we may have to assign to some of the states the probability that the sender prefers, as opposed to the probability that the receiver prefers. The following lemma characterizes these states.

Lemma 4. *Given an information transmission game $\Gamma^F = (A, \Omega, p, u)$, let $\omega^1, \omega^2, \ldots, \omega^\ell$ be the states in $\Omega_0^1 \cup \Omega_1^0$ sorted by $\frac{u_r(\omega,0)-u_r(\omega,1)}{u_s(\omega,1)-u_s(\omega,0)}$. Then, there exists a binary filter X_r^Γ that is optimal for the receiver and a value $\ell' \leq \ell$ such that:*

(a) $(X_r^\Gamma)_*(\omega) = 1$ *for all* $\omega \in \Omega_0^0$.
(b) $(X_r^\Gamma)_*(\omega) = 0$ *for all* $\omega \in \Omega_1^1$.
(c) For $i < \ell'$, $(X_r^\Gamma)_*(\omega^i) = 0$ *if* $\omega \in \Omega_1^0$ *and* $(X_r^\Gamma)_*(\omega^i) = 1$ *if* $\omega \in \Omega_0^1$.
(d) For $i > \ell'$, $(X_r^\Gamma)_*(\omega^i) = 0$ *if* $\omega \in \Omega_0^1$ *and* $(X_r^\Gamma)_*(\omega^i) = 0$ *if* $\omega \in \Omega_1^0$.

Lemma 4 states that, if we sort the states in which the sender and the receiver have different preferences by the ratio between how much the receiver prefers 0 with respect to 1 and how much the sender prefers 1 with respect to 0 (note that these ratios are always positive), we get that we can split these states into two contiguous blocks in which, in the first block, we assign them the probability that is most convenient for the sender, and for the second block we assign them the probability that is most convenient for the receiver. The proof follows the lines of that of Lemma 3 part (c).

Proof of Lemma 4. Given two states $\omega, \omega' \in \Omega_0^1 \cup \Omega_1^0$, we say that $\omega < \omega'$ if $\frac{u_r(\omega,0)-u_r(\omega,1)}{u_s(\omega,1)-u_s(\omega,0)} < \frac{u_r(\omega',0)-u_r(\omega',1)}{u_s(\omega',1)-u_s(\omega',0)}$. Using the same argument as in the Proof of Lemma 3, we can assume w.l.o.g. that (a) and (b) are satisfied. Thus, it only remains to show (c) and (d), which follow from the fact that there exists an there exists an optimal binary filter X_r^Γ for the receiver such that:

(s1) If $\omega, \omega' \in \Omega_1^0$, $\omega < \omega'$ and $(X_r^\Gamma)_*(\omega) > 0$, then $(X_r^\Gamma)_*(\omega') = 1$.
(s2) If $\omega, \omega' \in \Omega_0^1$, $\omega < \omega'$ and $(X_r^\Gamma)_*(\omega) < 1$, then $(X_r^\Gamma)_*(\omega') = 0$.
(s3) If $\omega \in \Omega_1^0$, $\omega' \in \Omega_0^1$, $\omega < \omega'$ and $(X_r^\Gamma)_*(\omega) > 0$, then $(X_r^\Gamma)_*(\omega') = 0$.
(s4) If $\omega \in \Omega_0^1$, $\omega' \in \Omega_1^0$, $\omega < \omega'$ and $(X_r^\Gamma)_*(\omega) < 1$, then $(X_r^\Gamma)_*(\omega') = 1$.

We will prove (s1) and (s3), the proofs of (s2) and (s4) are analogous.

Proof of (s1): Suppose that there exist $\omega, \omega' \in \Omega_1^0$ such that $\omega < \omega'$ and $(X_r^\Gamma)_*(\omega) > 0$, but $(X_r^\Gamma)_*(\omega') < 1$. If this happens, we can set

$$(X_r^\Gamma)_*(\omega) \longleftarrow (X_r^\Gamma)_*(\omega) + \varepsilon(u_s(\omega', 0) - u_s(\omega', 1))$$
$$(X_r^\Gamma)_*(\omega') \longleftarrow (X_r^\Gamma)_*(\omega) + \varepsilon(u_s(\omega, 1) - u_s(\omega, 0)).$$

Since $u_s(\omega',0)-u_s(\omega',1))<0$ and $u_s(\omega,0)-u_s(\omega,1)<0$, there exists some $\varepsilon>0$ that is small enough so that $X_*^{OPT}(\omega)$ and $(X_r^\Gamma)_*(\omega)$ stay between 0 and 1. Moreover, as in the proof of Lemma 3 part (c), if we perform this change the sender's utility remains unchanged, but the receiver's utility increases by

$$\Delta := \varepsilon((u_s(\omega',0)-u_s(\omega',1))(u_r(\omega,0)-u_r(\omega,1))+(u_s(\omega,1)-u_s(\omega,0))(u_r(\omega',0)-u_r(\omega',1))).$$

By assumption, we have that

$$\frac{u_r(\omega,0)-u_r(\omega,1)}{u_s(\omega,1)-u_s(\omega,0)} < \frac{u_r(\omega',0)-u_r(\omega',1)}{u_s(\omega',1)-u_s(\omega',0)}$$

which, together with the fact that $u_s(\omega,1)-u_s(\omega,0)>0$ and $u_s(\omega',1)-u_s(\omega',0)>0$, implies that $\Delta\geq 0$ whenever $\varepsilon>0$.

Proof of (s3): The proof is almost identical to that of (s1). Suppose that there exist $\omega\in\Omega_1^0$, $\omega'\in\Omega_0^1$ such that $\omega<\omega'$, $(X_r^\Gamma)_*(\omega)>0$ and $(X_r^\Gamma)_*(\omega')>0$. In this case, we can again set

$$\begin{array}{l}(X_r^\Gamma)_*(\omega) \longleftarrow (X_r^\Gamma)_*(\omega)+\varepsilon(u_s(\omega',1)-u_s(\omega',0))\\ (X_r^\Gamma)_*(\omega') \longleftarrow (X_r^\Gamma)_*(\omega)+\varepsilon(u_s(\omega,0)-u_s(\omega,1)).\end{array}$$

The only difference with the proof of (s1) is that, in this case, $u_s(\omega,0)-u_s(\omega,1)<0$ and $u_s(\omega',0)-u_s(\omega',1)>0$. Again, there exists a sufficiently small $\varepsilon>0$ such that $(X_r^\Gamma)_*(\omega)$ and $(X_r^\Gamma)_*(\omega')$ remain between 0 and 1. Moreover, a straightforward computation shows that the sender's utility remains unchanged, but that the receiver's utility changes by

$$\Delta := \varepsilon((u_s(\omega',1)-u_s(\omega',0))(u_r(\omega,0)-u_r(\omega,1))+(u_s(\omega,0)-u_s(\omega,1))(u_r(\omega',0)-u_r(\omega',1))).$$

Using that

$$\frac{u_r(\omega,0)-u_r(\omega,1)}{u_s(\omega,1)-u_s(\omega,0)} < \frac{u_r(\omega',1)-u_r(\omega',0)}{u_s(\omega',0)-u_s(\omega',1)},$$

$u_s(\omega,1)-u_s(\omega,0)>0$, and $u_s(\omega',0)-u_s(\omega',1)>0$, we get that $\Delta\geq 0$ whenever $\varepsilon>0$. □

Lemma 4 shows that the algorithm provided in Section 4.1 outputs the correct solution as long as it computes the right value of q in Step 5. The next lemma shows that this value is precisely the one that the algorithm finds. This completes the proof of correctness.

Lemma 5. *Let X_R^Γ be the filter that assigns $(X_R^\Gamma)_*(\omega)=1$ for $\omega\in\Omega_0^0\cup\Omega_1^0$ and $(X_R^\Gamma)_*(\omega)=0$ for $\omega\in\Omega_1^1\cup\Omega_0^1$. If X_R^Γ is incentive-compatible for the sender, (X_R^Γ) is the optimal filter for the receiver that is incentive-compatible for both players. Otherwise, all filters X that are incentive-compatible for both players and are optimal for the receiver satisfy at least one of the following equations:*

$$\begin{array}{l}\sum_i p(\omega_i)\cdot d_i^s\cdot X_*(\omega_i)=0\\ \sum_i p(\omega_i)\cdot d_i^s\cdot (1-X_*(\omega_i))=0.\end{array}$$

Proof. Filter X_R^Γ gives the maximum utility to the sender of all possible filters. Therefore, if it is incentive-compatible for the sender, it is the optimal for the receiver. Suppose instead that X_R^Γ is not

incentive-compatible for the sender but that an optimal filter X_r^Γ satisfies

$$\begin{array}{l} \sum_i p(\omega_i) \cdot d_i^s \cdot X_*(\omega_i) > 0 \\ \sum_i p(\omega_i) \cdot d_i^s \cdot (1 - X_*(\omega_i)) < 0. \end{array} \tag{2}$$

Since X_R^Γ is not IC for the sender, there exists a state $\omega \in \Omega_{0,1}$ such that $(X_r^\Gamma)_* > 0$ or a state $\omega \in \Omega_{1,0}$ such that $(X_r^\Gamma)_* < 1$. In the first case, we can decrease $(X_r^\Gamma)_*$ by a small value $\varepsilon > 0$ such that Equation 2 is still satisfied. By doing this, we increase the receiver's expected utility while obtaining a new filter that is still incentive-compatible for both players. The latter case is analogous except that we increase $(X_r^\Gamma)_*$ instead of decreasing it. □

5 Constructing a Sender-Optimal Filter

Constructing a sender-optimal filter X_s^Γ in an information transmission game is analogous the one given in Section 4.1 but "reversing" the roles of the sender and the receiver. More precisely, the construction is as follows.

1. **Step 1:** Set $(X_s^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega^0$.
2. **Step 2:** Set $(X_s^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega^1$.
3. **Step 3:** Sort all states $\omega \in \Omega_0^1 \cup \Omega_1^0$ according to the value $\frac{u_s(\omega,0)-u_s(\omega,1)}{u_r(\omega,1)-u_r(\omega,0)}$. Let $\omega^1, \omega^2, \ldots, \omega^{k'}$ be the resulting list.
4. **Step 4:** Set $(X_s^\Gamma)_*(\omega) = 1$ for all $\omega \in \Omega_0^1$ and $(X_s^\Gamma)_*(\omega) = 0$ for all $\omega \in \Omega_1^0$. If the resulting filter is incentive-compatible for the receiver, output X_s^Γ.
5. **Step 5:** For $i = 1, 2, \ldots, k'$ do the following. If $(X_s^\Gamma)_*(\omega^i)$ is set to 1 (resp., to 0), set it to 0 (resp., to 1). If the resulting filter is incentive-compatible for the receiver, find the maximum (resp., the minimum) q such that setting $(X_s^\Gamma)_*(\omega^i)$ to q is incentive-compatible for the receiver. If the resulting filter is incentive-compatible for the sender, output X_s^Γ, otherwise output the constant filter $(X_s^\Gamma)_* \equiv 0$.

The proof of correctness is analogous to the one shown in Section 4.2.

6 Proof of Theorem 2

The proof of Theorem 2 follows from the following result of Arieli et al. [1] describing the characterization of the best Nash equilibrium in information aggregation games with two senders, one receiver, binary actions and a mediator.

Theorem 4 ([1]). *Let $\Gamma = \{S, A, \Omega, p, u\}$ be an information aggregation game with $S = \{s_1, s_2\}$ and $A = \{0,1\}$ in which the players can communicate with a third-party mediator. Given $i, j, k \in \{0,1\}$, let $\Omega_{i,j}^k$ be the set of states in which s_1 prefers action i, s_2 prefers action j, and the receiver prefers action k. Let also $\Omega_{i,j} := \Omega_{i,j}^0 \cup \Omega_{i,j}^1$. Consider the following six maps $M_1, M_2, M_3, M_4, M_5, M_6$ from Ω to $[0,1]$:*

(a) $M_1(\omega) = 1$ *for all* $\omega \in \Omega^0_{0,0}$ *and* $M_1(\omega) = 0$ *otherwise.*

(b) $M_2(\omega) = 0$ *for all* $\omega \in \Omega^1_{1,1}$ *and* $M_2(\omega) = 1$ *otherwise.*

(c) $M_3(\omega) = 1$ *for all* $\omega \in \Omega_{0,0} \cup \Omega_{0,1}$ *and* $M_3(\omega) = 0$ *otherwise.*

(d) $M_4(\omega) = 1$ *for all* $\omega \in \Omega_{0,0} \cup \Omega_{1,0}$ *and* $M_4(\omega) = 0$ *otherwise.*

(e) $M_5(\omega) = 1$ *for all* $\omega \in \Omega$.

(f) $M_6(\omega) = 0$ *for all* $\omega \in \Omega$.

Then, there exists an $i \in \{1,2,3,4,5,6\}$ *and a Nash equilibrium of* Γ *that maximizes the receiver's utility in which the function that maps each state to the probability that the receiver ends up playing* 0 *is equal to* M_i.

Intuitively, Theorem 4 states that if there were no filters and the two senders had access to a third-party mediator, there exists a Nash equilibrium that is optimal for the receiver in which the either (a) the receiver only plays 0 if all three players prefer 0, (b) the sender plays 1 if and only if all three players prefer 1, (c) the receiver always plays what the first sender prefers, (d) the receiver always plays what the second sender prefers, (e) the receiver plays always 0, or (f) the receiver plays always 1. In the setting described in Section 2, players have no access to a mediator. However, we can easily check if the outcomes described in (a), (b), (c), (d), (e) or (f) are incentive-compatible for the receiver (i.e., if the receiver gets more utility with the outcome than when playing with no information), there exists a strategy in Γ (without the mediator) that implements these outcomes.

For instance, suppose that we want to implement the outcome described in (a). Consider the strategy profile $\vec{\sigma}^{(0,0)}$ in which each sender sends 0 to the receiver if and only if the realized state is in $\Omega^0_{0,0}$, and the receiver plays 0 only if both signals are 0. It is easy to check that this is incentive-compatible for the senders: if the realized state is indeed in $\Omega^0_{0,0}$, it is a best response for both to send 0 since it guarantees that the receiver will play 0. Moreover, if the realized state is not in $\Omega^0_{0,0}$, none of the senders gets any additional utility by defecting from the main strategy since the other sender will always send a non-zero signal (which implies that the receiver will play 1). To implement (c), consider the strategy profile $\vec{\sigma}^{s_1}$ in which both senders send a binary signal $m \in \{0,1\}$ that is equal to 0 if and only if the realized state is in $\Omega_{0,0} \cup \Omega_{0,1}$, and the receiver plays 0 if and only if the signal from the first sender is 0. Again, it is straightforward to check that this is incentive-compatible for the senders: it is a best-response for sender 1 to send its preference, while it doesn't matter what sender 2 sends since it will be ignored. To implement (e) (resp., (f)) consider the strategy profile $\vec{\sigma}^0$ (resp., $\vec{\sigma}^1$) in which the senders send signal 0 regardless of the realized state and the receiver always plays 0 (resp., 1). It is easy to check that $\vec{\sigma}^0$ (resp., $\vec{\sigma}^1$) are incentive-compatible for the senders. Implementing the outcomes in (b) and (d) is analogous to implementing the outcomes in (a) and (c). Since all outcomes that are implementable without a mediator are also implementable with a mediator, this implies the following proposition:

Proposition 4. *Let* $\Gamma = \{S, A, \Omega, p, u\}$ *be an information aggregation game with* $S = \{s_1, s_2\}$ *and* $A = \{0,1\}$. *Then, either* $\vec{\sigma}^0$, $\vec{\sigma}^1$, $\vec{\sigma}^{(0,0)}$, $\vec{\sigma}^{(1,1)}$, $\vec{\sigma}^{s_1}$ *or* $\vec{\sigma}^{s_2}$ *is a Nash equilibrium that is optimal for the receiver.*

Proposition 4 implies that we can break the problem of finding the receiver-optimal filter in an information aggregation game into four sub-problems:

(a) Finding a filter $(X^\Gamma_r)_{(0,0)}$ such that $\vec{\sigma}^{(0,0)}$ gives the maximal utility for the receiver in $\Gamma((X^\Gamma_r)_{(0,0)})$.

(b) Finding a filter $(X_r^\Gamma)_{(1,1)}$ such that $\vec{\sigma}^{(1,1)}$ gives the maximal utility for the receiver in $\Gamma((X_r^\Gamma)_{(1,1)})$.

(c) Finding a filter $(X_r^\Gamma)_{s_1}$ such that $\vec{\sigma}^{s_1}$ gives the maximal utility for the receiver in $\Gamma((X_r^\Gamma)_{s_1})$.

(d) Finding a filter $(X_r^\Gamma)_{s_2}$ such that $\vec{\sigma}^{s_2}$ gives the maximal utility for the receiver in $\Gamma((X_r^\Gamma)_{s_2})$.

Note that these filters might not always exist (for instance, $(X_r^\Gamma)_{(0,0)}$ does not exist when the receiver always prefers 0 and the senders always prefer 1). Moreover, we are not including the optimal filters for $\vec{\sigma}^0$ and $\vec{\sigma}^1$ since all filters give the same utility with these strategies. Finding $(X_r^\Gamma)_{s_1}$ and $(X_r^\Gamma)_{s_2}$ (whenever they exist) reduce to the case of one sender (by ignoring s_2 and s_1, respectively), and thus can be solved with the $O(k \log k)$ algorithm presented in Section 4.1. We next show how to compute $(X_r^\Gamma)_{(0,0)}$ using a linear program with k variables and $2k+2$ constraints. Computing $(X_r^\Gamma)_{(1,1)}$ is analogous.

Suppose that X is a filter that maximizes the receiver utility in $\Gamma(X)$ when playing strategy $\vec{\sigma}^{(0,0)}$. Denote by M_0 the set of signals of X in which both senders and the receiver prefer 0 to 1. Consider a filter X' that samples a signal m according to X and does the following: if $m \in M_0$, it sends signal 0. Otherwise, it sends signal 1. It is straightforward to check that if $\vec{\sigma}^{(0,0)}$ is a Nash equilibrium in Γ_X^F, it is also a Nash equilibrium in $\Gamma(X')$. Moreover, players get the same utilities with X and X', which means that X' also maximizes the receiver's utility. This implies that we can restrict our search to binary filters in which all players prefer action 0 on signal 0 and the receiver prefers action 1 on signal 1.

Let $\omega_1, \omega_2, \ldots, \omega_k$ be the elements of Ω and define $A_i := u(s_1, \omega_i, 0) - u(s_1, \omega_i, 1)$, $B_i := u(s_2, \omega_i, 0) - u(s_2, \omega_i, 1)$ and $C_i := u(r, \omega_i, 0) - u(r, \omega_i, 1)$. A binary filter X over Ω can be described by a sequence of k real numbers $x_1, \ldots, x_k$ between 0 and 1 such that x_i denotes the probability that the filter sends signal 0 on state ω_i. The condition that both senders prefer action 0 on signal 0 translates into $\sum_{i=1}^k A_i x_i \geq 0$ and $\sum_{i=1}^k B_i x_i \geq 0$.

Moreover, the utility of the receiver with filter X and strategy $\vec{\sigma}^{(0,0)}$ is given by $\sum_{i=1}^k (x_i \cdot u(r, \omega_i, 0) + (1 - x_i) \cdot u(r, \omega_i, 1))$. This sum can be rearranged into $\sum_{i=1}^k u(r, \omega_i, 1) + \sum_{i=1}^k C_i x_i$. Therefore, finding a filter $(X_r^\Gamma)_{(0,0)}$ such that $\vec{\sigma}^{(0,0)}$ gives the maximal utility for the receiver in $\Gamma((X_r^\Gamma)_{(0,0)})$ reduces to solving the following linear programming instance:

$$\begin{aligned} &\max \textstyle\sum_{i=1}^k C_i x_i \\ &\textstyle\sum_{i=1}^k A_i x_i \geq 0 \\ &\textstyle\sum_{i=1}^k B_i x_i \geq 0 \\ 0 \leq x_i \leq 1 \quad &\forall i \in [k]. \end{aligned}$$

Note that, even though we are maximizing the receiver's utility, it may be the case that $\vec{\sigma}^{(0,0)}$ is not incentive-compatible for the receiver in $\Gamma((X_r^\Gamma)_{(0,0)})$. For instance, the receiver might prefer playing 0 whenever the senders send 1. If such a thing happens (which can be tested in linear time), it means that there is no filter satisfying the desired conditions.

7 Conclusion

Garbling the information of the sender can sometimes increase the players' utilities in information transmission and information aggregation games, and the optimal garbling for both sender and receiver can be computed with efficient algorithms. However, this paper leaves two important open questions. First if the algorithm presented in Section 4.1 can be generalized to three or more actions and, second, if we can construct an efficient algorithm that outputs the best filter for one of the senders in an information aggregation game with multiple senders.

References

[1] Itai Arieli, Ivan Geffner & Moshe Tennenholtz (2023): *Mediated cheap talk design*. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, pp. 5456–5463, doi:10.1609/aaai.v37i5.25678.

[2] Itai Arieli, Ivan Geffner & Moshe Tennenholtz (2023): *Resilient Information Aggregation*. In: *Proceedings of the 2023 Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, *EPTCS* 379, pp. 31–45, doi:10.4204/EPTCS.379.6.

[3] Dirk Bergemann, Benjamin Brooks & Stephen Morris (2015): *The limits of price discrimination*. *American Economic Review* 105(3), pp. 921–957, doi:10.2139/ssrn.2501403.

[4] Andreas Blume & Oliver Board (2013): *Language barriers*. *Econometrica* 81(2), pp. 781–812, doi:10.3982/ecta9183.

[5] Roberto Corrao & Yifan Dai (2023): *Mediated Communication with Transparent Motives*. In: *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 489–489, doi:10.1145/3580507.3597808.

[6] Kris De Jaegher (2003): *A game-theoretic rationale for vagueness*. *Linguistics and Philosophy* 26(5), pp. 637–659, doi:10.1023/A:1025853728992.

[7] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot & Vardan Papyan (2023): *Llm censorship: A machine learning challenge or a computer security problem?* *arXiv preprint arXiv:2307.10719*, doi:10.48550/arXiv.2307.10719.

[8] Jeanne Hagenbach & Frédéric Koessler (2020): *Cheap talk with coarse understanding*. *Games and Economic Behavior* 124, pp. 105–121, doi:10.1016/j.geb.2020.07.015.

[9] Shota Ichihashi (2019): *Limiting Sender's information in Bayesian persuasion*. *Games and Economic Behavior* 117, pp. 276–288, doi:10.2139/ssrn.3233072.

[10] Gerhard Jäger, Lars P Metzger & Frank Riedel (2011): *Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals*. *Games and economic behavior* 73(2), pp. 517–537, doi:10.1016/j.geb.2011.03.008.

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

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

[13] Zichuan Liu, Zefan Wang, Linjie Xu, Jinyu Wang, Lei Song, Tianchun Wang, Chunlin Chen, Wei Cheng & Jiang Bian (2024): *Protecting your llms with information bottleneck*. *Advances in Neural Information Processing Systems* 37, pp. 29723–29753, doi:10.48550/arXiv.2404.13968.

[14] Vaidehi Patil, Peter Hase & Mohit Bansal (2023): *Can sensitive information be deleted from llms? objectives for defending against extraction attacks*. *arXiv preprint arXiv:2309.17410*, doi:10.48550/arXiv.2309.17410.

Choquet Expected Utility and Never Best Choice

Abstract

For a given set of capacities, describing uncertainty over a finite set of states, and a finite set of acts, the question is considered when an act is never a best choice, specifically, when for every capacity in the given set there is another act with higher Choquet expected utility. This question is answered for several sets of capacities, distinguished by their supports, where the focus is on four different definitions of a support. One consequence of the analysis is that an act is never a best choice against the set of all capacities if and only if it is strictly dominated by a convex combination of the other acts after making these comonotonic with the act under consideration. This result can be seen as the counterpart of the analogous result for additive capacities, such as mixed strategies in games.

Christopher Kops
j.kops@maastrichtuniversity.nl

Hans Peters
h.peters@maastrichtuniversity.nl

Dries Vermeulen
d.vermeulen@maastrichtuniversity.nl

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 46–46, doi:10.4204/EPTCS.437.6

Robust Price Discrimination

Abstract

Itai Arieli
iarieli@technion.ac.il

Yakov Babichenko
yakovbab@technion.ac.il

Omer Madmon
omermadmon@gmail.com

Moshe Tennenholtz
moshet@technion.ac.il

We consider a model of third-degree price discrimination where the seller's product valuation is unknown to the market designer, who aims to maximize buyer surplus by revealing buyer valuation information. Our main result shows that the regret is bounded by a $(1/e)$-fraction of the optimal buyer surplus when the seller has zero valuation for the product. This bound is attained by randomly drawing a seller valuation and applying the segmentation of Bergemann et al. (2015) with respect to the drawn valuation. We show that this bound is tight in the case of binary buyer valuation. The full version of this paper is available at https://arxiv.org/abs/2401.16942

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 47–47, doi:10.4204/EPTCS.437.7

Hyperintensional Intention

Daniil Khaitovich
ILLC, Univesity of Amsterdam
Amsterdam, The Netherlands
d.khaitovich@uva.nl

Aybüke Özgün
ILLC, Univesity of Amsterdam
Amsterdam, The Netherlands
a.ozgun@uva.nl

Intentions are crucial for our practical reasoning. The rational intention obeys some simple logical principles, such as agglomeration and consistency, among others, motivating the search for a proper logic of intention. However, such a logic should be weak enough not to force the closure under entailment; otherwise, we cannot distinguish between intended consequences of agents' choices and their unintended side-effects. In this paper we argue that we should avoid not only the closure under entailment, but the weaker closure under equivalence as well. To achieve this, we develop a hyperintensional logic of intention, where what an agent intends is constrained by the agent's decision problem. The proposed system combines some elements of inquisitive and topic-sensitive theories of intensional modals. Along the way, we also show that the existing closest relatives of our framework overgenerate validities by validating some instances of closure under equivalence. Finally, we provide a sound and strongly complete axiomatization for this logic.

1 Introduction

Intentions do not obey many of the *closure principles* logicians bake in in their formalisms. While we are rationally committed to intend some of the consequences of what we intend, we are not rationally committed to intend, e.g., all logical, known, or believed consequences of our intentions. To cite a classic example [12, p. 218], one intends to go to the dentists to get one's tooth filled, believing or knowing that filling tooth will lead to being in pain (maybe because the agent doesn't know about anesthetics or they have a high tolerance to anesthetics or they cannot take much of them due to other health issues), while not intending to be in pain. Agents not only do not intend the unforeseen side-effects of their actions, but they also do not intend all of their foreseen side-effects [8, 12].

This has already been well documented in the literature on logics of intentions [13, 24, 34, 33], and much ink has been spilled to meet the challenge of finding a logical formalism for intentions that neither over- nor undergenerates the *intended* consequences of one's intentions such that the overgeneration problem is solved in a principled, well-motivated way [10, 3, 35, 28]. This is exactly what this paper aims to do. It provides a semantics for intentions that at least alleviates, if not completely avoids, the overgeneration problem - also known as *the problem of side-effects* [17] - for the right reasons and that explains which closure principles should be preserved.

The key idea we fall back on to achieve our goal is the so-called *issue-relevance* [4, 25] or *question-sensitivity* [29, 43, 21, 2] of propositional mental states, which has been explored extensively in a body of rapidly growing literature. Question-sensitivity of, e.g., epistemic/doxastic attitudes can explain the failure of some closure principles by modeling these attitudes as dependent on inquiry relevant to an agent's epistemic/doxastic agenda, alleviating the infamous problem of logical omniscience [39, 42, 20] or accounting for the effects of inquiry on a rational, idealized agent's knowledge and belief [2, 25]. In a similar vein, observing the parallel between the problems of logical omniscience and side-effects, recent work [3] developed a question-sensitive theory of rational intentions, modeling intentions

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 48–64, doi:10.4204/EPTCS.437.8

© Khaitovich & Özgün
This work is licensed under the
Creative Commons Attribution License.

as dependent on the agent's practical question *what to do?*, namely, their decision problem. We take our cue from this theory but argue that it still overgenerates validities: it is still subject to the problem of side-effects. To briefly explain (and to be further elaborated below), the purely possible-worlds, partition-based mechanism employed in [3] avoids closure under logical, necessary or known/believed entailments but still validates closure under (undesired instances of) logical and necessary *equivalents* for intentions. We provide compelling examples that challenge closure under logical, necessary, and known/believed equivalents and, in turn, argue that a logic of intentions should also avoid these closure principles of equivalents. In particular, it should be *hyperintensional*.

Our proposal employs tools from topic- or subject matter-sensitive semantics for knowledge and belief [31, 5, 20] and use them to model *decision problems* in a finer, hyperintensional way. We fully develop this formalism based on a bi-modal language with a global and an intention modality, and provide a sound and strongly complete axiomatization for the proposed hyperintensional logic of intention.

The paper is organized as follows. In Section 2, we present the problem of side-effects and challenges it poses for any logic of intention: we argue that closure under (logical, necessary and known/believed) equivalence is too strong for the logic of intention and list some principles that the logic of intention should validate. In Section 3, we review the closest relatives of our framework, namely the proposals in [3], and show that they do not satisfy the requirements we state in Section 2. In Section 4, we present a hyperintensional logic of intention that is weak enough not to force the closure under equivalence, but strong enough to validate all "good" principles. The framework is closely related to topic-sensitive theories of doxastic propositional attitudes [31, 4]. Moreover, we provide a sound and strongly complete axiom system for the hyperintensional logic of intention. As the proofs are not central to the conceptual contributions of this work, the longer proofs are omitted from the main body of the paper and presented in the appendices.

2 Logics of Intension and The Problem of Side-Effects

Formal epistemologists and philosophical logicians often turn to Kripke semantics to model propositional mental attitudes such as knowledge, belief and, of particular interest for us in this paper, *intention*. As well known, however, this mainstream approach validates certain closure principles for these attitudes that are fit to model only highly idealized knowers and believers. As also emphasized, e.g., in [24], these closure principles are a more serious threat for intention, as they cannot be defended for intentions even as idealizations. In this paper, we will focus on the following closure principles, formulated for intentions via operator I:

1. **Closure under logical entailment:** $\models \varphi \to \psi$ then $\models I\varphi \to I\psi$
2. **Necessary entailment:** $\models \boxplus(\varphi \to \psi) \to (I\varphi \to I\psi)$
3. **Closure under logical equivalence:** $\models \varphi \leftrightarrow \psi$ then $I\varphi \leftrightarrow I\psi$
4. **Necessary equivalence:** $\models \boxplus(\varphi \leftrightarrow \psi) \to (I\varphi \leftrightarrow I\psi)$

A few notes on notation is warranted. In the formulation of principles 1-4, $\models$ represents the logical consequence relation of the relevant logic, $I\varphi$ is read as "the agent intends that φ". $\boxplus\varphi$ is read as "it is a priori that φ" or "the agent knows that φ", depending on the context and examples we target. The reason why we do not introduce distinct operators for knowledge and a priori necessity is merely practical and the lack of such distinct operators does not bear on any substantive conceptual points we want to make in this work. In our formalism, we interpret $\boxplus$ as the global modality, closer to the standard interpretation of a priori necessity. Introducing an additional knowledge operator K interpreted on a subset of the

logical space, representing epistemically accessible worlds of the agent (as in, e.g., the Kripke semantics of epistemic logic) does not make any difference for the kinds of logical principles we focus on in this work.[1] This much formalism should be sufficient for our exposition in Sections 2 and 3. All these notions and notation will be properly introduced in Section 4.

We say that a logic *suffers from the problem of side-effects* if it validates at least one of 1-4 (even when φ and ψ are propositional variables). Just as an epistemic agent can be non-omniscient for diverse reasons [5, 20], the problem of side-effects can be explained by different factors. Since the early stages of intention logic research, principles 1 and 2 have been considered undesirably strong [24, 10]. There are at least two factors explaining the failure of these principles:

1. *Epistemic flaws*: One can intend that φ only if one has the epistemic resources to reason about φ. Agents cannot derive and do not always know/believe all the consequences of what they intend, thus, they do not always intend all necessary or logical consequences of what they intend.

For example, one may intend to maintain a vegan diet without intending to avoid eating anything that contains cysteine, simply because one has no idea what cysteine is or does not know that it is made out of animal products.[2]

2. *Control constraint*: One can intend that φ only if one believes that one has control over φ. So that, if one intends that φ, φ entails ψ, but one does not believe that one has control over ψ, one will not intend ψ. [3, p.357].

Epistemic flaws are not specific to intention and are widely studied in formal theories of knowledge and belief (see, e.g., [16, 42, 38, 20]). On the contrary, control constraint is a unique feature of intention and similar motivational attitudes. Cases of the failure of closure under entailment due to control constraint have been discussed in the existing literature [35, 3]. Authors often point to *foreseeing without intending*: one may intend to do some action, foresee some necessary consequences of that action (hence know that the intended action entails such consequences), yet still not intend these consequences [1]. We focus on the less discussed cases, where even the weaker principle of closure under equivalence fails.[3]

Below we present a few examples motivating the failure of 3-4 exactly due to control constraint. In all such cases, the bearer of intention intends that φ, knows that φ is (necessarily or materially) equivalent to ψ, but does not intend that ψ. Since the equivalences are known by the agent in question, such failures cannot be explained by epistemic flaws.

Example 1. *Imagine a patient who suffers from a severe dental disease. The doctor tells the patient that without any treatment, he will certainly lose his teeth, but the treatment decreases the chance only by*

[1] This claim assumes that the set of so-called conative alternatives w.r.t. which the operator I is interpreted is a subset of the epistemically accessible worlds. Moreover, one can also read $\boxplus$ as belief. In this extended abstract, we often refer to knowledge to keep the presentation more concise.

[2] This example is inspired by Stalnaker's famous William III example for belief [38], which is also used in [3, p. 348] for intention to make a similar point.

[3] In [10], authors approach the issue from the perspective of resource-bounded agents who can only conceive the so-called *cognitively finite objects*. Instead of possible worlds, which are maximally consistent sets of formulae that might lack finite cognitive representations, they use *cognitive abstractions*, i.e. consistent (but not necessarily maximally consistent) set of formulae that are built from finitely many propositional variables. This solution allows one to avoid the closure under logical equivalence, but is not sufficient to explain cases of foreseeing without intending: the agent may have enough cognitive resources to grasp and foresee some consequence of their intention, but might still not intend it. A similar worry can also be raised against [28], where all undesired closure principles are avoided via the machinery of awareness functions. This is conceptually insufficient, since, as foreseeing without intending demonstrates, agents might be aware of some unintended consequences of their intended actions. A similar solution based solely on syntactic restriction is given in [14]: closure under logical equivalence (and entailment) is restricted by arbitrary *practical rules of reasoning*, by means of which we can avoid not only undesired instances of such closure principles, but also virtually any closure, such as agglomeration.

50%. By agreeing to the treatment, the patient intends to save his teeth with 50% chance, but he doesn't intend to lose his teeth with 50% chance.[4]

Example 2. *In some monarchy, it is obligatory to pay taxes. A citizen of the monarchy does not want to face any legal consequences of tax avoidance, hence, she intends to pay taxes. The monarch necessarily spends tax money on official ceremonies. In this context, the propositions "one's taxes are paid" and "one's paid taxes are spent on official ceremonies" are necessarily equivalent (in all the worlds that comply with the customs of the monarchy): if one pays taxes, then the monarch spends them on ceremonies, while the spending one's paid taxes on ceremonies entails that one's taxes are paid. Being familiar with the customs of the monarch, this equivalence is known by the citizen. Nevertheless, the citizen may not intend that her paid taxes are spent on official ceremonies, despite intending to pay taxes.*

We cannot explain these failures of the closure under equivalence by epistemic flaws. Instead, they can be explained via control constraint.[5] Intuitively, the patient has control over the decrease of the chance of teeth loss by 50%, since there is an option to take the treatment, but has no control over the fact that there still will be 50% chance of losing the teeth. The taxpayer controls whether her taxes are paid or not, but it is not in her control to prevent the monarch from spending her tax money not to her liking.[6]

Therefore, the control constraint poses a challenge not only to the principles of closure under entailment (1 & 2), but also to the principles of closure under equivalences (3 & 4). A logic of rational intention should avoid them. Yet, there are a number of simple and intuitive logical principles that rational intention should validate; that is, a logic of intention is still possible. For example, the restricted versions of the above mentioned closure principles, constrained to the cases where the agent has control over both propositions in question and there are no epistemic obstacles, should be validated. Moreover, it is broadly endorsed that rational intentions are consistent (if one intends that φ, one doesn't intend that $\neg\varphi$) [12, 24, 10, 35] and they agglomerate (if one intends that φ and one intends that ψ, then one intends that $\varphi \wedge \psi$) [8, 3].[7]

[4] Similar examples can be found in the literature on framing effects [22].

[5] While early theories of agency define control strictly in terms of causation [11] – the agent has control over φ iff they can force φ by their actions – the more recent accounts refine these definitions, connecting control with explanatory relations. In [23], it is argued that an agent has control over φ only if the agent can act in a way that will lead to φ and the explanation why φ will obtain is centered around the actions of the agent. This leads to the failure of the closure under equivalents: it is possible that φ holds iff so does ψ, but explanations why those propositions are true differ. Explanation is often seen as an example of hyperintensional phenomena. All mathematical truths are necessary true and hence logically and a priori equivalent, nevertheless, they have different explanations. Such phenomena occur with contingent propositions as well: "*it is true that grass is green*" is explained by "*the grass is green*", but not vise versa, while two propositions are obviously necessarily equivalent. Since the metaphysics and logic of explanation are out of the scope of the present paper, we do not elaborate further on this argument. See, e.g., [30, p.157] and [36] for further discussion.

[6] As pointed out by one of the reviewers, the last case resembles familiar examples of intentions that involve the actions of others [9, Chapter 8]. Since our current framework is single-agent, what matters for the purposes of this paper is whether the agent has control over the relevant proposition – regardless of whether the lack of control stems from another agent's actions or from something else. Exploring how a lack of control arises from the actions of others in a multi-agent extension of this framework is an interesting direction for future work, which we plan to pursue in a follow-up paper.

[7] Admittedly, the agglomeration principle for intention is more contentious. [40], for example, argues that the unqualified agglomeration principle pressures agents to have "*one enormous compound intention*" [p. 517] that guide their actions, which is arguably implausible and might hinder their ability to satisfy their relevant desires. Still, there are ways to defend the agglomeration of intentions against such objections: see, e.g., [44] for an elaborate discussion. In this paper we adopt the agglomeration principle and leave a more detailed analysis of it for future work.

3 Formalizing Decision Problems as Partitions

One of the recent solutions to the problem of side-effects borrows another tool from formal epistemology – question-sensitivity. Some have argued that our knowledge and beliefs exist in the context of *questions* [38, 43, 2, 25]. Likewise, one can argue that our intentions exist in the context of *decision problems*. Just like any belief we hold is an answer to some question, any intention we have is a solution to a problem "*what to do?*" [3]. Question-sensivity explains why we may not intend some consequences of our intentions, even when we know what follows from them: some propositions might be *undefined* in the context of our decision problems; they might not constitute (partial) solutions to them.

How to formally model a decision problem? So far, authors take a decision problem to be a question and formalize it in terms of a partition of a logical space [3, 21, 2, 41, 29]:

Definition 1 (Decision problem as partition)**.** *Given a non-empty set of possible worlds W, a decision problem Π is a partition of W.*[8]

Definition 2 (Complete and partial solutions)**.** *Every cell $X \in \Pi$ is a complete solution to the decision problem Π. A union of any set of cells $A = \bigcup_{i\in I} X_i$ for some $\{X_i\}_{i\in I} \subseteq \Pi$ is a partial solution to Π.*

It is easy to see that every complete solution $X \in \Pi$ is a partial solution to Π: $X = \bigcup\{X\}$ with $\{X\} \subseteq \Pi$.

Intuitively, we associate a decision problem with a set of its mutually exhaustive complete solutions that the agent considers implementing. *Partial* solutions may be viewed as indeterministic choices between a set of compelete solutions: agents may intend only to partially solve their decision problem, given that sometimes the difference between some complete solutions is of no importance to them. For example, one may be dealing with the quest of getting to the railway station and consider three possible ways to do that: by bus, by tram or by foot. Both the bus and the tram depart from the same place and take roughly equal amount of time to get to the destination, so one may intend to simply go to the bus/tram station and take a bus or a tram without settling on either of the two complete solutions.

A decision problem restricts what propositions the agent can intend. Namely, if one intends that φ, then φ should be *defined on the agent's decision problem*. Beddor & Goldstein [3] has two proposals to formalize the latter notion of definedness on a decision problem. To recap briefly, the first one takes it that φ is defined on the agent's decision problem Π iff φ is a partial solution to Π (in the sense described in Definition 2). The corresponding logic of intention invalidates closure under entailment and validates agglomeration, but still forces closure under equivalents in full generality. The second proposal is a closer rival to our account presented in this paper, so we introduce its formal components in more detail and argue that it still overgenerates validities. Below, we assume that the object of intention is a sentence of a language of classical propositional logic, i.e., a formula that is built from atomic propositions using classical Boolean connectives ($\neg, \wedge, \vee$. etc). The semantics for such expressions is the standard possible worlds semantics, where $[\![\varphi]\!]$ denotes the set of possible worlds that make φ true.

The second proposal (see [3, Section 6]) takes it that a proposition φ is defined on a decision problem iff the *subject matter* of φ is included in the decision problem, where the notion of subject matter is also defined via partitions, in line with Lewisian theory of subject matters [26, 27].[9] Here, we briefly restate the relevant definitions of [3].

Definition 3 (Subject matters)**.** *Given a logical space W and any formula φ, $sm(\varphi) \subseteq 2^{2^W}$ is the subject matter of φ, defined recursively as follows (where p is atomic):*

[8] A partition Π of W is a subset $\Pi \subseteq 2^W$ such that $\emptyset \notin \Pi$, $\bigcup\Pi = W$, and for any $X, Y \in \Pi$ such that $X \neq Y$, $X \cap Y = \emptyset$.

[9] A detailed presentation of theories of subject matter is beyond the scope of this work. We refer the reader to [27, 19, 32, 26].

$sm(\top) = \{W\}$ $\quad sm(\neg\varphi) = sm(\varphi)$
$sm(p) = \{[\![p]\!], W \setminus [\![p]\!]\} \setminus \{\emptyset\}$ $\quad sm(\varphi \wedge \psi) = sm(\varphi \vee \psi) = \{X \cap Y \mid X \in sm(\varphi), Y \in sm(\psi)\} \setminus \{\emptyset\}$

Note that for any formula φ, $sm(\varphi)$ is a partition on W[10]*; Moreover, $\exists P \subseteq sm(\varphi) : \bigcup P = [\![\varphi]\!]$ and $\exists N \subseteq sm(\varphi) : \bigcup N = [\![\neg\varphi]\!]$.*

Definition 4 (Parthood on subject matter)**.** *Given a logical space W and two propositions φ, ψ, a subject matter of φ is a part of subject matter of ψ, $sm(\varphi) \sqsubseteq sm(\psi)$, iff for any $X \in sm(\varphi)$ there exists a subset $S \subseteq sm(\psi)$, such that $\bigcup S = X$.*

We are now set up to define the formal framework. Given some logical space W, we can represent intention via (1) the set of conative alternatives, i.e. the non-empty set of possible worlds $Con \subseteq W$, such that the agent's intention is satisfied in Con-worlds; (2) the decision problem Π , which is a partition of the logical space W.

Definition 5 (Question-sensitive intention)**.** *Given a logical space W, a set of conative alternatives Con and a decision problem Π, an agent intends that φ iff (1) $Con \subseteq [\![\varphi]\!]$ and (2) $sm(\varphi) \sqsubseteq \Pi$, i.e. for any $X \in sm(\varphi)$ there exists a partial solution $S \subseteq \Pi$, s.t. $X = \bigcup S$.*

The intuition behind Definition 5 is the following: the agent intends that φ iff (1) they intend to solve their decision problem in a way that will force φ and (2) φ is defined on the agent's decision problem. The agent has control over φ, since φing is considered a partial solution to their decision problem. By the very same reason, the agent is aware that they bring about that φ by the solution they are committed to implement. Hence, both control constraint and epistemic flaws are accounted for in this semantics.

Definition 5 validates consistency and agglomeration for intention. The closure under a priori entailment (2) and equivalence (4) fail, since the scope of the intention is restricted by the decision problem. Nevertheless, the usage of subject matters does not rule out all cases when the closure under a priori or known equivalences may fail. The closure cannot be falsified for atomic propositions: it not hard to see that under Definition 5, for any atomic propositions p, q, if $[\![p]\!] = [\![q]\!]$, then $sm(p) = sm(q)$ and hence one intends that p iff one intends that q. Examples 1 and 2 falsify this principle.

Yet another, more conceptual, argument against formalizing decision problems as partitions comes from the way the notion of parthood on questions is defined. Intuitively, one decision problem is a part of another iff every complete solution to the former is a partial solution to the latter (a complete solution to a coarser problem is a partial solution to a finer problem). More formally, Π_1 is a part of Π_2, denoted by $\Pi_1 \sqsubseteq \Pi_2$, iff for all $A \in \Pi_1$ there is $\{X_i\}_{i \in I} \subseteq \Pi_2$ such that $A = \bigcup_{i \in I} X_i$. We refer to $\sqsubseteq$ relation as *extensional parthood*, i.e. $\Pi_1 \sqsubseteq \Pi_2$ reads as "Π_1 is an extensional part of Π_2". Another useful acronym: $\Pi_1 \simeq \Pi_2 := \Pi_1 \sqsubseteq \Pi_2 \,\&\, \Pi_2 \sqsubseteq \Pi_1$, referred to as *extensional equivalence* (Π_1 is extensionally equivalent to Π_2). The claim according to which there is nothing more to parthood than an extensional parthood, is going to be referred as *extensional parthood claim*. Modeling decision problems as partitions supports the extensional parthood claim, affecting both proposals of question-sensitivity in [3]. On the contrary, we reject the extensional parthood claim, since there are decision problems Π_1, Π_2, where $\Pi_1 \sqsubseteq \Pi_2$, but the agent can intentionally solve Π_1 without any intention to solve Π_2. Consider the next example.

Example 3 (Taking a train service without minding the time and place of departure)**.** *A ticket office clerk works at a train station with two platforms. At every platform, a train departs once an hour, at j:00. Every train has a unique code of the form $(i; j)$: the train $(i; j)$ departs from the ith platform at j:00. Let*

[10] Our definition of *sm* function differs from the one given in [3]. The only differences are that we define $sm(\top)$ as a primitive notion and make sure that for any $\varphi \in \mathscr{L}$, $\emptyset \notin sm(\varphi)$. It affects the framework neither conceptually nor technically; and is done for the sake of consistency with our notation and definitions.

r^i_j be the proposition "someone has a ticket to the train with service number $(i;j)$", t_j – "someone has a ticket to a train that departs at j:00", p_i – "someone has a ticket to a train that departs from the ith platform". Note that every atomic proposition r^i_j is equivalent to the corresponding conjunction $p_i \wedge t_j$: every train service has its unique departure platform and time.

Someone comes to the train station and asks the clerk to sell them a ticket for a train with number $(2;13)$: the customer only knows the train service number, since they were asked to buy that ticket with no further details provided. The clerk knows the system well and is aware that the train with number $(2;13)$ departs from platform 2 at 13:00. There is a pile of 48 tickets with numbers: $(1;1),(1;2),\ldots,(2;24)$. The clerk needs to choose which ticket they should take from the pile and intends that the customer gets the ticket for the train with service number $(2;13)$. The clerk intends neither that the customer gets a ticket for a train that departs from the second platform nor that the customer gets a ticket for a train that departs at 13:00, since that is simply not the focus of the clerk.

Intuitively, Example 3 is just another instance of foreseeing without intending. The clerk intends to give the customer the ticket the customer has asked for. The clerk also foresees that as a consequence of that, the customer will receive a ticket for a train that departs at 13:00 from platform number 2. Nevertheless, the clerk does not intend to give the customer a ticket for a train with a specific departure time or place, since it is just irrelevant to the decision problem the clerk is facing. Unfortunately, formalization via subject matters of solutions cannot capture this nuance. The clerk needs to choose the ticket from the pile, i.e., they consider all the tickets, and their decision problem is $\Pi = \{[\![r^i_j]\!] \mid i \in \{1,2\}, 0 \leq j \leq 24\} \cup \{W \setminus \bigcup^{i\in\{1,2\}}_{0\leq j\leq 24} [\![r^i_j]\!]\}$. In other words, $\Pi = sm(\bigvee^{i\in\{1,2\}}_{0\leq j\leq 24} r^i_j)$. Under Definition 4, $sm(\bigvee^{i\in\{1,2\}}_{0\leq j\leq 24} r^i_j) = sm(\bigvee^{i\in\{1,2\}}_{0\leq j\leq 24} (p_i \wedge t_j))$. So that, r^i_j is defined on Π iff $p_i \wedge t_j$ is defined on Π. Under Definition 5, the clerk cannot intend that r^i_j without intending that p_i and that t_j.

We may come up with a number of similar examples: one may look for a specific building without minding the street and the house number of that building or wonder about one's age without thinking about neither the day nor the month nor the year one was born, etc.: in all such cases, an agent solves a problem without minding the extensional parts of the problem.

Summarizing, modeling decision problems as partitions does not seem to be sufficiently fine-grained, since it enforces extensional parthood and, as a consequence, the closure under equivalence 4 is satisfied in contexts where it should fail.

4 A Hyperintensional Logic of Intention

As we have just seen, modeling decision problems as partitions overgenerates validities: it validates undesired, atomic instaces of 4. Luckily, there is a concurrent way to represent decision problems: as atomic objects. Namely, we can simply state that there is a set P of problems that do not have internal structure. Then, we can define a solution function s, which maps every formula φ to a set of problems: if $a \in s(\varphi)$, then bringing about that φ is a partial solution to a. Definedness on a decision problem and control constraint are expressible via s. If the agent is solving a decision problem $a \in P$ and $a \in s(\varphi)$, then bringing about that φ is a partial solution to a the agent considers. Hence, φ is defined on a, which means that the agent has control over φ and is aware that some solution to the decision problem forces φ. To have an adequate notion of parthood on decision problems, we define the corresponding partial order $\leq$ on P. Then, following our intuition about parthood, function s should be $\leq$-upward closed: if $a \leq b$, then for any φ, $a \in s(\varphi)$ entails $b \in s(\varphi)$. Informally, if a is a part of b, then any partial solution to a is a partial

solution to b. As we are to see, this approach steps on the hyperintensional terrain, where the closure under logical and necessary equivalences fail. Moreover, it will allow us to differentiate even between the necessarily equivalent atomic propositions and adequately model cases such as Examples 1-3. We also elaborate on how our proposed way of modeling decision problems relates to the partition-based perspective presented in Section 3.

4.1 Formalizing decision problems as atomic objects

First, we define the formal language we are to work with. Let $\mathscr{L}_{CPL}$ be the language of classical propositional logic with a countable set of atomic propositions Prop, defined recursively as follows:

$$\alpha := \top \mid p \mid \neg\alpha \mid (\alpha \vee \alpha) \mid (\alpha \wedge \alpha)$$

where $p \in \mathsf{Prop}$. Note that we have $\top$, $\neg$, $\vee$ and $\wedge$ as primitives and do not define one in terms of another. Nevertheless, we use the interdefinability of $\neg$, $\vee$ and $\wedge$ in terms of each other as our logic cannot discern, e.g. $\varphi \wedge \psi$ from $\neg(\neg\varphi \vee \neg\psi)$. However, we have specifically defined $\top$ as a primitive: it is pure tautology without non-logical context. The reader should not confuse $\top$ with another tautology expressed using the elements of Prop, such as $p \vee \neg p$ and $p \to p$. They will be logically equivalent but they won't be substitutable with each other under the scope of our intention modality. For $\bot$, we set $\bot := \neg\top$. Based on $\mathscr{L}_{CPL}$, the well-formed formulas of the language $\mathscr{L}$ of our logic of intention are given by the following grammar:

$$\alpha \mid \neg\varphi \mid (\varphi \vee \varphi) \mid (\varphi \wedge \varphi) \mid \boxplus\varphi \mid I\alpha$$

where $\alpha \in \mathscr{L}_{CPL}$. $I\alpha$ reads as "*agent intends that* α". Notice that the maximum modal depth of a formula w.r.t. I is one, that is, I takes only the elements of $\mathscr{L}_{CPL}$ in its scope. It is done so to exclude expressions about higher order intentions, such as intention to intend. The reason for that restriction is to be discussed later. $\boxplus\varphi$ read as "*it is necessarily true that* φ". We employ the usual abbreviations for propositional connectives $\to$ and $\leftrightarrow$ as $\varphi \to \psi := \neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi := (\varphi \to \psi) \wedge (\psi \to \varphi)$. We interpret this language on the so-called *problem-sensitive Kripke models*, obtained by endowing Kripke models with a join-semilattice that represents decision problems and their parthood relation. The former is familiar in modal logic and requires no elaboration [6]. The component for the representation of decision problems, called *problems model*, is a variation of topic-sensitive models [5] and warrants further explanation.

Definition 6 (Problems model)**.** *A problems model $\mathscr{P}$ is a tuple $(P, \oplus, s)$, where:*

1. *$P \neq \emptyset$ is a non-empty set of decision problems;*
2. *$\oplus : P \times P \to P$ is a binary idempotent, commutative, associative operation: problem fusion. We assume the unrestricted fusion, that is, $\oplus$ is always defined on P: $\forall A \subseteq P \exists a \in P : \bigoplus A = a$. The parthood relation $\leq$ is defined in terms of $\oplus$: $\forall a, b \in P : a \leq b$ iff $a \oplus b = b$;*
3. *$s : \mathsf{Prop} \to 2^P$ is a function assigning a set of decision problems to each element in* Prop*. Namely, if $a \in s(p)$, then bringing it about that p is a suitable (partial) solution to a. s extends to the whole propositional language $\mathscr{L}_{CPL}$ as follows:*

$$\begin{array}{ll} s(\top) = P & s(\varphi \vee \psi) = s(\varphi) \cap s(\psi) \\ s(\neg\varphi) = s(\varphi) & s(\varphi \wedge \psi) = \{a \oplus b \mid a \in s(\varphi), b \in s(\psi)\} \end{array}$$

Given any $p \in Prop$, $s(p)$ is upward closed: $a \leq b \Rightarrow (a \in s(p) \Rightarrow b \in s(p))$.

The extension of s to the whole propositional language is motivated as follows. Given any problem $a \in P$, forcing φ is a partial solution to the problem iff to decide if φ or not to φ is a part of a, hence, $s(\varphi) = s(\neg\varphi)$. $\varphi \vee \psi$ partially solves a iff both disjuncts – to bring about that φ, to bring about that ψ – solve a: this corresponds to our intuition about partial solutions as indeterministic choice between a set of solutions. Finally, $\varphi \wedge \psi$ is a partial solution to a iff a can be split into two problems, i.e. $a = a_1 \oplus a_2$, such that φing partially solves a_1 and ψing partially solves a_2. The following lemma shows that s is upward closed over the whole language $\mathscr{L}_{CPL}$.

Lemma 1. *Given a problems model $\mathscr{P} = (P, \oplus, s)$, $a, b \in P$ and $\varphi \in \mathscr{L}_{CPL}$, the following holds:*

$$a \leq b \Rightarrow (a \in s(\varphi) \Rightarrow b \in s(\varphi)).$$

Proof. The proof follows by induction on the structure of φ. The case for atomic propositions holds by the definition of s. Case $\varphi := \top$ is trivial as $s(\top) = P$ and case $\varphi := \neg\psi$ follows from the induction hypothesis (IH) and the fact that $s(\neg\psi) = s(\psi)$.

Case $\varphi := \psi \vee \chi$: Suppose that $a \leq b$ and $a \in s(\psi \vee \chi)$. By the defn. of s, this means that $a \in s(\psi)$ and $a \in s(\chi)$. Then, by the IH, we have $b \in s(\psi)$ and $b \in s(\chi)$, thus, $b \in s(\psi) \cap s(\chi) = s(\psi \vee \chi)$.

Case $\varphi := \psi \wedge \chi$: Suppose that $a \leq b$ and $a \in s(\psi \wedge \chi)$. By the defn. of s, this means that $a = d \oplus c$ for some $d \in s(\psi)$ and $c \in s(\chi)$. Observe that $a = d \oplus c$ implies that $d \leq a$ and $c \leq a$, and, in turn, $d \leq b$ and $c \leq b$. Then, by IH, we have $b \in s(\psi)$ and $b \in s(\chi)$. As $b = b \oplus b$, by defn. of s, we have $b \in s(\psi \wedge \chi)$. □

4.2 Semantics, Axiomatization, and Completeness

Definition 7 (Problem-sensitive frames and models)**.** *A problem-sensitive frame is a tuple $F = (W, R, \mathscr{P}, f)$, where:*

- *$W \neq \emptyset$ is a non-empty set of possible worlds;*
- *$R \subseteq W \times W$ is a serial binary relation on W, such that $R(w)$ is the set of conative alternatives to w, i.e. possible worlds where agent's intention in w is satisfied;*
- *$\mathscr{P} = (P, \oplus, s)$ is a problems model;*
- *$f : W \to P$ is a function, which takes a world and returns a decision problem that the agent is solving in the given world;*

As usual, any frame F may be extended to a model $\mathscr{M} = (F, V)$, where $V : \mathsf{Prop} \to 2^W$ is a standardly defined valuation function.

We show the connection between the decision problems as atomic objects and as partitions. Namely, every atomic decision problem $a \in P$, as it is given in Definition 6, can be transformed into a partition Π_a. Intuitively, Π_a is an *extensional characterization* of a: if we accepted the extensional parthood claim, then Π_a could be identified with a and treated as in Definition 5. Moreover, we can show that if $a \leq b$, then $\Pi_a \sqsubseteq \Pi_b$. In other words, $\leq$ relation is not arbitrary, but meaningful w.r.t. solutions: if $a \leq b$, then every partial solution to a is a partial solution to b, hence, $a \sqsubseteq b$. Nevertheless, the converse does not hold: we may have decision problems $a, b \in P$, such that $\Pi_a \sqsubseteq \Pi_b$, but $a \not\leq b$. This corresponds to our denial of the extensional parthood claim: extensional parthood does not entail parthood in general. See Appendix A for the constructions and proofs of the mentioned properties. We interpret $\mathscr{L}$ on problem-sensitive models as in the following definition.

Definition 8 (Semantic for $\mathscr{L}$)**.** *Given a problem-sensitive model $\mathscr{M} = (W, R, \mathscr{P}, f, V)$ and a possible world $w \in W$, the satisfiability relation $\models$ is defined as follows, where $[\![\alpha]\!]_{\mathscr{M}} = \{w \in W \mid \mathscr{M}, w \models \varphi\}$:*

$$\begin{array}{lll} \mathscr{M}, w \models p & \textit{iff} & w \in V(p) \\ \mathscr{M}, w \models \neg\varphi & \textit{iff} & \mathscr{M}, w \not\models \varphi \\ \mathscr{M}, w \models \varphi \vee \psi & \textit{iff} & \mathscr{M}, w \models \varphi \textit{ or } \mathscr{M}, w \models \psi \\ \mathscr{M}, w \models \boxplus\varphi & \textit{iff} & \forall v \in W : \mathscr{M}, v \models \varphi \\ \mathscr{M}, w \models I\alpha & \textit{iff} & R(w) \subseteq [\![\alpha]\!]_{\mathscr{M}} \textit{ and } f(w) \in s(\alpha). \end{array}$$

The notions of logical consequence, validity, soundness, and completeness are defined standardly [7]. While the semantic clauses for the Booleans and $\boxplus$ are standard, the intention operator $I\varphi$ is interpreted in a problem-sensitive way: the agent intends that α at w iff (1) α is true in all conative alternatives at w and (2) α is a partial solution to the agent's decision problem at w. (2) is what invalidates 1-4, making the resulting logic hyperintensional.

Theorem 1. *Principles 1-4 are invalid (also for atomic propositions).*

Proof. We provide a counterexample against 4. The same counterexample can be used to prove the invalidity of the others. Consider the problem-sensitive model $\mathscr{M} = (\{w\}, \{(w,w)\}, (\{a,b,c\}, \oplus, s), f, V)$ such that $a \oplus b = c$, $a \not\leq b$ and $b \not\leq a$, $s(p) = \{a,c\}$, $s(q) = \{b,c\}$, $f(w) = a$, and $V(p) = V(q) = \{w\}$. We then have that $\mathscr{M} \models \boxplus(p \leftrightarrow q) \wedge Ip$, but $\mathscr{M}, w \not\models Iq$, since $f(w) = a \notin s(q)$. □

To state some of the axioms and other relevant principles in the complete system, we use the abbreviation '$\overline{\varphi}$' to denote the tautology $\bigwedge_{p \in Var(\varphi)}(p \vee \neg p)$[11], following a similar idea in [18], where $Var(\varphi)$ is the set of atomic propositions occuring in φ. It is easy to see that $I\overline{\varphi}$ simply expresses that φ *is a partial solution to the agent's decision problem* ($\mathscr{M}, w \models I\overline{\varphi}$ iff $f(w) \in s(\varphi)$). We list the axioms and rules of the logic of problem-sensitive models in Table 1. We briefly comment on the axioms for the intention

Label	Axiom / Rule	Label	Axiom / Rule
(CPL)	All classical propositional taut. and MP	($\mathsf{S5}_{\boxplus}$)	S5 axioms and rules for $\boxplus$
(Ax1)	$I\top$	(Ax4)	$(I\varphi \wedge I\psi) \leftrightarrow I(\varphi \wedge \psi)$
(Ax2)	$I\varphi \rightarrow I\overline{\varphi}$	(Ax5)	$\boxplus(\psi \rightarrow \varphi) \rightarrow ((I\psi \wedge I\overline{\varphi}) \rightarrow I\varphi)$
(Ax3)	$I\varphi \rightarrow \neg I\neg\varphi$		

Table 1: Axiomatization Log for the logic of problem-sensitive models.

operator. Ax1 means that one always has at least the weakest intention possible: to bring it about that $\top$. Ax2 states that one intends that φ only if φ is a partial solution to one's decision problem, reflecting problem-sensitivity of intention. Ax3 says that intentions are consistent and Ax4 states that intentions agglomerate. Finally Ax5 is our restricted closure principle: one intends those a priori consequences of one's intensions which also constitute partial solutions to one's decision problem. We conclude the section by stating our main technical result, a detailed proof of which can be found in Appendix B.[12]

Theorem 2. Log *is a sound and strongly complete axiomatization of $\mathscr{L}$ with respect to the class of all problem-sensitive models.*

[11] In order to have a unique definition of each $\overline{\varphi}$, we set the convention that elements of $Var(\varphi)$ occur in $\bigwedge_{p \in Var(\varphi)}(p \vee \neg p)$ from left-to-right in the order they are enumerated in $\mathsf{Prop} = \{p_1, p_2, \dots\}$.

[12] Unsurprisingly, the axiomatization is similar to the axiomatization of the logic of simple hyperintensional belief in [31]. We do not have an intention counterpart of their axiom $B\varphi \rightarrow \boxplus B\varphi$, since the semantics of I depends on the actual world.

5 Conclusions and Future Work

In this paper, we developed a hyperintensional logic of intention, addressing the problem of side-effects by invalidating a number of closure principles. By integrating elements from inquisitive semantics and topic-sensitive epistemic frameworks, we provided a formal system that better captures the rational commitments of intentional agents. Our approach refines question-sensitive theory of intention by capturing hyperintensional differences between decision problems. Technically, we introduced a bi-modal logic with a sound and strongly complete axiomatization.

Several directions remain open for future work. First, incorporating belief alongside intention would allow for a more nuanced interaction between epistemic states and decision-making. Second, formalizing a notion of control would further clarify what propositions agents in general can intend. Third, allowing nested I operators would raise interesting questions, such as which decision problems could be solved by adopting intention to φ, and how they are connected with problems solvable by φ itself. Moreover, extending our system to dynamic settings could provide insights into intention revision and deliberative reasoning over time. Finally, a multi-agent extension of the proposed framework would allow us to model interesting interactions among intending agents, e.g., lack of control that results from the actions of others.

Acknowledgments

Many thanks to our anonymous referees for their constructive feedback, which helped us significantly improve the final version of this extended abstract. Special thanks to Franz Berto for allowing us to test some of our examples on him.

References

[1] Bruce Aune (1966): *Intention and foresight*. *The Journal of Philosophy* 63(20), pp. 652–654, doi:10.2307/2024263.

[2] Alexandru Baltag, Rachel Boddy & Sonja Smets (2018): *Group knowledge in interrogative epistemology*. *Jaakko Hintikka on knowledge and game-theoretical semantics*, pp. 131–164, doi:10.1007/978-3-319-62864-6_5.

[3] Bob Beddor & Simon Goldstein (2023): *A question-sensitive theory of intention*. *The Philosophical Quarterly* 73(2), pp. 346–378, doi:10.1093/pq/pqac031.

[4] Francesco Berto (2018): *Aboutness in imagination*. *Philosophical Studies* 175(8), pp. 1871–1886, doi:10.1007/s11098-017-0937-y.

[5] Francesco Berto (2022): *Topics of thought: The logic of knowledge, belief, imagination*. Oxford University Press, doi:10.1093/oso/9780192857491.001.0001.

[6] Patrick Blackburn, Johan FAK van Benthem & Frank Wolter (2006): *Handbook of modal logic*. Elsevier.

[7] Patrick Blackburn, Maarten de Rijke & Yde Venema (2001): *Modal Logic*. Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, doi:10.1017/cbo9781107050884.

[8] Michael Bratman (1987): *Intention, Plans, and Practical Reason*, doi:10.2307/2215590.

[9] Michael Bratman (1999): *Faces of intention: Selected essays on intention and agency*. Cambridge University Press, doi:10.1017/cbo9780511625190.

[10] Xiaoping Chen & Guiquan Liu (1999): *A logic of intention*. In: *Proceedings of the 16th international joint conference on Artifical intelligence-Volume 1*, pp. 172–177.

[11] Roderick M Chisholm (1964): *Human freedom and the self*. In: *Free Will*, Blackwell.

[12] Philip R Cohen & Hector J Levesque (1990): *Intention is choice with commitment*. *Artificial intelligence* 42(2-3), pp. 213–261, doi:10.1016/0004-3702(90)90055-5.

[13] Philip R Cohen & Hector J Levesque (1990): *Persistence, intention, and commitment*. *Intentions in communication*, pp. 33–69, doi:10.7551/mitpress/3839.003.0005.

[14] Wiebe van Der Hoek, Wojciech Jamroga & Michael Wooldridge (2007): *Towards a theory of intention revision*. *Synthese* 155, pp. 265–290, doi:10.1007/s11229-006-9145-6.

[15] 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.

[16] Ronald Fagin, Joseph Y. Halpern, Yoram Moses & Moshe Vardi (2003): *Reasoning About Knowledge*. MIT Press, Cambridge.

[17] John Finnis (1991): *Intention and side-effects*, p. 32–64. Cambridge Studies in Philosophy and Law, Cambridge University Press, doi:10.1017/cbo9780511527395.004.

[18] Alessandro Giordani (2019): *Axiomatizing the Logic of Imagination*. *Studia Logica* 107(4), pp. 639–657, doi:10.1007/s11225-018-9810-z.

[19] Peter Hawke (2018): *Theories of aboutness*. *Australasian Journal of Philosophy* 96(4), pp. 697–723, doi:10.1080/00048402.2017.1388826.

[20] Peter Hawke, Aybüke Özgün & Francesco Berto (2020): *The fundamental problem of logical omniscience*. *Journal of Philosophical Logic* 49(4), pp. 727–766, doi:10.1007/s10992-019-09536-6.

[21] Daniel Hoek (2022): *Questions in Action*. *Journal of Philosophy* 119(3), pp. 113–143, doi:10.5840/jphil202211938.

[22] Daniel Kahneman & Amos Tversky (1984): *Choices, values, and frames.* *American psychologist* 39(4), p. 341, doi:10.1017/cbo9780511803475.002.

[23] Mikayla Kelley (forthcoming): *A Control Theory of Action*. *Australasian Journal of Philosophy*.

[24] Kurt Konolige & Martha E Pollack (1993): *A representationalist theory of intention*. In: *IJCAI*, 93, pp. 390–395.

[25] Wessel Kroon (2024): *Knowledge as Issue-Relevant Information*. Master's thesis, University of Amsterdam. Available at `https://eprints.illc.uva.nl/id/eprint/2316/1/MoL-2024-06.text.pdf`.

[26] David Lewis (1988): *Relevant implication*. *Theoria* 54(3), pp. 161–174, doi:10.1111/j.1755-2567.1988.tb00716.x.

[27] David Lewis (1988): *Statements partly about observation*. *Philosophical papers* 17(1), pp. 1–31, doi:10.1080/05568648809506282.

[28] Bernd van Linder, Wiebe van der Hoek & J J Ch Meyer (1996): *Formalising motivational attitudes of agents: On preferences, goals and commitments*. In: *Intelligent Agents II Agent Theories, Architectures, and Languages: IJCAI'95 Workshop (ATAL) Montréal, Canada, August 19–20, 1995 Proceedings 2*, Springer, pp. 17–32, doi:10.1007/3540608052_56.

[29] Stefan Minica (2011): *Dynamic logic of questions*. Ph.D. thesis, University of Amsterdam.

[30] Daniel Nolan (2014): *Hyperintensional metaphysics*. *Philosophical Studies* 171, pp. 149–160, doi:10.1007/s11098-013-0251-2.

[31] Aybüke Özgün & Francesco Berto (2021): *Dynamic hyperintensional belief revision*. *The Review of Symbolic Logic* 14(3), pp. 766–811, doi:10.1017/S1755020319000686.

[32] Matteo Plebani & Giuseppe Spolaore (2021): *Subject Matter: A Modest Proposal*. *Philosophical Quarterly* 71(3), pp. 605–622, doi:10.1093/pq/pqaa054.

[33] Martha E Pollack (1991): *Overloading intentions for efficient practical reasoning*. *Noûs* 25(4), pp. 513–536, doi:10.2307/2216076.

[34] Anand S Rao & Michael P Georgeff (1997): *Modeling rational agents within a BDI-architecture*. Readings in agents, pp. 317–328.

[35] Olivier Roy (2008): *Thinking before acting: intentions, logic, rational choice*. University of Amsterdam.

[36] Benjamin Schnieder (2011): *A logic for 'because'*. The Review of Symbolic Logic 4(3), pp. 445–465, doi:10.1017/s1755020311000104.

[37] Maximilian Siemers (2021): *Hyperintensional Logics for Evidence, Knowledge and Belief*. Master's thesis, ILLC, University of Amsterdam.

[38] Robert Stalnaker (1984): *Inquiry*. MIT Press.

[39] Robert Stalnaker (1991): *The problem of logical omniscience, I*. Synthese, pp. 425–440, doi:10.1007/bf00413506.

[40] Steven Sverdlik (1996): *Consistency among intentions and the 'Simple View'*. Canadian Journal of Philosophy 26(4), pp. 515–522, doi:10.1080/00455091.1996.10717465.

[41] Johan Van Benthem & Ştefan Minică (2012): *Toward a dynamic logic of questions*. Journal of Philosophical Logic 41, pp. 633–669, doi:10.1007/s10992-012-9233-7.

[42] Moshe Y Vardi (1986): *On epistemic logic and logical omniscience*. In: Theoretical aspects of reasoning about knowledge, Elsevier, pp. 293–305, doi:10.1016/b978-0-934613-04-0.50024-7.

[43] Seth Yalcin (2018): *Belief as question-sensitive*. Philosophy and Phenomenological Research 97(1), pp. 23–47, doi:10.1111/phpr.12330.

[44] Jing Zhu (2010): *On the principle of intention agglomeration*. Synthese 175, pp. 89–99, doi:10.1007/s11229-009-9531-y.

A Proofs from Section 4.2

Lemma 2. *Given a problems model $\mathscr{P} = (P, \oplus, s)$, a decision problem $a \in P$, and formulas $\varphi, \psi \in \mathscr{L}_{CPL}$, $a \in s(\varphi \vee \psi)$ iff $a \in s(\varphi \wedge \psi)$.*

Proof. ($\Rightarrow$) Assume $a \in s(\varphi \vee \psi)$. By Defn. 6, this means that $a \in s(\varphi) \cap s(\psi)$, i.e. $a \in s(\varphi)$ and $a \in s(\psi)$. Then, since $a = a \oplus a$, we obtain that $a \in s(\varphi \wedge \psi)$.

($\Leftarrow$) Assume $a \in s(\varphi \wedge \psi)$. By Defn. 6, this means that there is $b, c \in P$ such that $a = b \oplus c, b \in s(\varphi), c \in s(\psi)$. The fact that $a = b \oplus c$ implies that $b \leq a$ and $c \leq a$. Then, by Lemma 1, we obtain that $a \in s(\varphi)$ and $a \in s(\psi)$, that is, $a \in s(\varphi) \cap t(\psi)$. By Defn. 6, we conclude $a \in s(\varphi \vee \psi)$. □

Proposition 1 (Every decision problem yields a partition)**.** *Given any problem-sensitive model $\mathscr{M} = (W, R, \mathscr{P}, f, V)$ any decision problem $a \in P$, and $\varphi \in \mathscr{L}_{CPL}$, there exists a partition Π_a of W such that if $a \in s(\varphi)$, then φ is defined on Π_a, i.e. $sm(\varphi) \sqsubseteq \Pi_a$.*

Proof. Let $s^{-1}(a) = \{\varphi \in \mathscr{L}_{CPL} \mid a \in s(\varphi)\}$ that is , $\varphi \in s^{-1}(a)$ iff bringing it about that φ partially solves a. Let $S^{-1}(a)$ be a collection of maximally satisfiable subsets of $s^{-1}(a)$, i.e. $\Gamma \in S^{-1}(a)$ iff 1) $\Gamma \subseteq s^{-1}(a)$; 2) there exists a world $w \in W$, such that $\mathscr{M}, w \models \varphi$ for all $\varphi \in \Gamma$ (from now on, abbreviated as $\mathscr{M}, w \models \Gamma$) and 3) for any $\Delta \subseteq s^{-1}(a)$, such that $\Gamma \subset \Delta$, there is no world $w \in W$, such that $\mathscr{M}, w \models \Delta$.

For every $\Gamma \in S^{-1}(a)$ we define a set $[\![\Gamma]\!] = \{w \in W \mid \mathscr{M}, w \models \varphi \text{ for all } \varphi \in \Gamma\}$. Then, clearly the set $\Pi_a = \{[\![\Gamma]\!] \mid \Gamma \in S^{-1}(a)\}$ is a partition of W. We show that for any $\varphi \in \mathscr{L}_{CPL}$, if $a \in s(\varphi)$ then φ is defined on Π_a.

By induction on φ. Let $\varphi := p$. Assume $a \in s(p)$ for some proposition $p \in \mathsf{Prop}$. W.l.o.g., let $[\![p]\!] \neq W$. By Definition 3, $sm(p) = \{[\![p]\!], W \setminus [\![p]\!]\}$. Let $\widehat{p} = \{\bigcup [\![\Gamma]\!] \mid \Gamma \in S^{-1}(a), p \in \Gamma\}$. Clearly, $\widehat{p} = [\![p]\!]$: since $p \in s^{-1}(a)$, for every $\Gamma \in S^{-1}(a)$ either $p \in \Gamma$ or $\neg p \in \Gamma$. Hence, for any possible world

$w \in W$, if $\mathscr{M}, w \models p$, then there exists $\Gamma \in S^{-1}(a)$, such that $\mathscr{M}, w \models \Gamma$, from which it follows that $w \in \widehat{p}$. On the other hand, if $w \in \widehat{p}$, then there exists a maximally satisfiable set $\Gamma \in S^{-1}(a)$, such that $p \in \Gamma$. Since $\mathscr{M}, w \models \Gamma$, $\mathscr{M}, w \models p$, from which, by definition, it follows that $w \in [\![p]\!]$. Then, $[\![p]\!] = \bigcup_{i \in I} X_i$ for some $\{X_i\}_{i \in I} \subseteq \Pi_a$, and $W \setminus [\![p]\!] = \bigcup_{j \in J} X_j$, where $\{X_j\}_{j \in J} = \Pi_a \setminus \{X_i\}_{i \in I}$ i.e. $sm(p) \sqsubseteq \Pi_a$, i.e. p is defined on Π_a.

Let $\varphi := \neg\psi$. By IH, if $a \in sm(\psi)$, then ψ is defined on Π_a. Since $sm(\psi) = sm(\neg\psi) = sm(\varphi)$, if $a \in sm(\varphi)$, then $sm(\varphi) \sqsubseteq \Pi_a$.

Let $\varphi := \psi_1 \vee \psi_2$ and $a \in s(\psi_1 \vee \psi_2)$. Since $s(\psi_1 \vee \psi_2) = s(\psi_1) \cap s(\psi_2)$, $a \in s(\psi_1)$ and $a \in s(\psi_2)$. By IH, if $a \in s(\psi_1), s(\psi_2)$, then ψ_1 and ψ_2 are defined on Π_a, i.e. $sm(\psi_1) \sqsubseteq \Pi_a$ and $sm(\psi_2) \sqsubseteq \Pi_a$. By Definition 4, it means that $\forall A \in sm(\psi_1) \exists \{X_i\}_{i \in I} \subseteq \Pi_a : \bigcup_{i \in I} X_i = A$ and $\forall B \in sm(\psi_1) \exists \{X_j\}_{j \in J} \subseteq \Pi_a : \bigcup_{i \in J} X_j = B$. By Definition 3, $sm(\psi_1 \vee \psi_2) = \{A \cap B \mid A \in sm(\psi_1), B \in sm(\psi_2)\} \setminus \{\emptyset\}$. Take any $A \cap B \in sm(\psi_1 \vee \psi_2)$. As we have shown before, $A = \bigcup_{i \in I} X_i$ and $B = \bigcup_{j \in J} X_j$ for some $\{X_k\}_{k \in I \cup J} \subseteq \Pi_a$. Hence, given that all X_i's and X_j's are disjoint[13], $A \cap B = \bigcup_{l \in I \cap J} X_l$. Since $A \cap B$ was an arbitrary element of $sm(\psi_1 \vee \psi_2)$, we can conclude that $\forall O \in sm(\psi_1 \vee \psi_2) \exists \{X_l\}_{l \in L} \subseteq \Pi_a : \bigcup_{l \in L} X_l = O$, so that, $sm(\psi_1 \vee \psi_2) \sqsubseteq \Pi_a$, which means that $\psi_1 \vee \psi_2$ is define don Π_a.

Let $\varphi := \psi_1 \wedge \psi_2$. Analogously to the previous case (cases are analogous in light of Lemma 2). □

Proposition 2. *Given any problem-sensitive model $\mathscr{M} = (W, R, \mathscr{P}, f, V)$ and any two problems $a, b \in P$, if $a \leq b$, then $\Pi_a \sqsubseteq \Pi_b$, where Π_a and Π_b are defined as in Proposition 1.*

Proof. Let $a \leq b$ for some arbitrary decision problems $a, b \in P$. Then, by Lemma 1, if $a \in s(\varphi)$, then $b \in s(\varphi)$ for any $\varphi \in \mathscr{L}_{CPL}$. In other words, $s^{-1}(a) \subseteq s^{-1}(b)$.

Let $X \in \Pi_a$ be an arbitrary complete solution to Π_a. By the definition of Π_a, $X = [\![\Gamma]\!]$, where Γ is a maximally satisfiable subset of $s^{-1}(a)$. Since $s^{-1}(a) \subseteq s^{-1}(b)$, there exists some maximally satisfiable subsets $\Sigma \subseteq s^{-1}(b)$ such that $\Gamma \subseteq \Sigma$. Let $[\Gamma]_b = \{\Sigma \in S^{-1}(b) \mid \Gamma \subseteq \Sigma\}$ be a collection of such maximally satisfiable subsets. As we have just shown, $[\Gamma]_b \neq \emptyset$. Moreover, as we are to show, $\bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!] = [\![\Gamma]\!]$. From left-to-right: assume $w \in \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$. Then, $w \in [\![\Sigma_i]\!]$ for some $\Sigma_i \in [\Gamma_b]$. Then, $\mathscr{M}, w \models \Sigma_i$. Since $\Gamma \subseteq \Sigma_i$, $\mathscr{M}, w \models \Gamma$, i.e. $w \in [\![\Gamma]\!]$. Right-to-left: assume $w \in [\![\Gamma]\!]$. Then, there exists some $\Sigma_i \in [\Gamma_b]$, such that $\mathscr{M}, w \models \Sigma_i$.[14] Then, $w \in [\![\Sigma_i]\!]$ and since $[\![\Sigma_i]\!] \subseteq \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$, $w \in \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$.

Moreover, by definition of Π_b, $\{[\![\Sigma]\!] \mid \Sigma \in [\Gamma]_b\} \subseteq \Pi_b$. Hence, X, a complete solution to Π_a, is a partial solution to Π_b. Since X was arbitrary, we can conclude that any complete solution to Π_a is a partial solution to Π_b, from which, by Definition 4, it follows that $\Pi_a \sqsubseteq \Pi_b$. □

Proposition 3 ($\Pi_a \sqsubseteq \Pi_b$ does not entail $a \leq b$)**.** *Given any problem-sensitive model $\mathscr{M} = (W, R, \mathscr{P}, f, V)$ and two problems $a, b \in P$, the fact that $\Pi_a \sqsubseteq \Pi_b$ does not entail $a \leq b$.*

Proof. We prove this statement by providing a counterexample. Let $\mathscr{M} = (W, R, \mathscr{P}, f, V)$ be a problem-sensitive model, such that for two propositions $p, q \in \mathsf{Prop}$, $V(p) = V(q)$. Then, let a be a problem such that $a \in s(p)$ and for any formula $\psi \in \mathscr{L}_{CPL}$, if ψ contains any propositions other that p, then $a \notin s(\psi)$. Consequently, $a \notin s(q)$. Let b be a decision problem such that $b \in s(q)$ and for any formula $\psi \in \mathscr{L}_{CPL}$, if ψ contains any propositions other that q, then $b \notin s(\psi)$. Consequently, $b \notin s(p)$. Obviously, $\Pi_a = \Pi_b$,

[13] For the contradiction, assume they are not disjoint. Then, there exists $X_i = [\![\Gamma_i]\!]$, $X_j = [\![\Gamma_j]\!]$, such that $X_i \cap X_i \neq \emptyset$. Let $w \in X_i \cap X_j$. Then, $\mathscr{M}, w \models \Gamma_i \cup \Gamma_j$, which contradicts the fact that Γ_i and Γ_j are *maximally satisfiable* subsets of $s^{-1}(a)$.

[14] In order to see why it holds, note that for any problem $x \in P$, $s^{-1}(x)$ is closed under negation, i.e., for any formula $\varphi \in \mathscr{L}_{CPL}$, $\varphi \in s^{-1}(x)$ iff $\neg\varphi \in s^{-1}(x)$ (since $s(\varphi) = s(\neg\varphi)$). Given that and the fact that $s^{-1}(a) \subseteq s^{-1}(b)$, we can complete any maximally satisfiable set $\Gamma \in S^{-1}(a)$ to some maximally satisfiable set $\Sigma \in S^{-1}(b)$ by an analogue of Lindenbaum construction.

hence, $\Pi_a \sqsubseteq \Pi_b$. At the same time, $a \not\leq b$, since it is not the case that for any $\varphi \in \mathscr{L}_{CPL}$, if $a \in s(\varphi)$, then $b \in s(\varphi)$: $a \in s(p)$, $b \notin s(p)$. □

B Proof of Theorem 2: Soundness and Completeness of Log

We prove completeness via a canonical model construction. Our proof closely resembles the completeness proof in [18] except for the construction of the problem sensitive components. We follow [37] for the construction of P^c and $\oplus^c$, and the construction of f^c is similar to the construction of awareness sets in awareness logics [15].

Consistency, maximal consistency, and derivability for Log are defined standardly [7].

Lemma 3. *The following are derivable in* Log*:*

1. $I\overline{\varphi} \leftrightarrow I(\bigwedge_{p \in Var(\varphi)} \overline{p})$
2. $I\overline{\varphi} \rightarrow I\overline{\psi}$, *if* $Var(\psi) \subseteq Var(\varphi)$
3. $(I\varphi \wedge I\psi) \rightarrow I(\varphi \vee \psi)$

Proof. 1 follows from Ax4; 2 follows from Theorem 3.1; and 3 follows from Axioms 7, 4, and 2. □

Lemma 4. *For every maximally consistent set (mcs)* Γ *of* Log *and* $\varphi, \psi \in \mathscr{L}$*, the following hold:*

1. $\Gamma \vdash \varphi$ *iff* $\varphi \in \Gamma$,
2. *if* $\varphi \in \Gamma$ *and* $\varphi \rightarrow \psi \in \Gamma$ *then* $\psi \in \Gamma$,
3. *if* $\vdash \varphi$ *then* $\varphi \in \Gamma$,
4. $\varphi \in \Gamma$ *and* $\psi \in \Gamma$ *iff* $\varphi \wedge \psi \in \Gamma$,
5. $\varphi \in \Gamma$ *iff* $\neg\varphi \notin \Gamma$.

Lemma 5 (Lindenbaum's Lemma)**.** *Every* Log*-consistent set can be extended to a maximally* Log*-consistent one.*

The proofs of Lemmas 4 and 5 are standard and we sometimes use them without explicit mention. We are now ready define our canonical model.

Let W^c be the set of all maximally Log-consistent sets. For each $\Gamma \in W^c$, define

$$\begin{aligned} \Gamma[\boxplus] &:= \{\varphi \in \mathscr{L} : \boxplus\varphi \in \Gamma\}, \\ \Gamma[I] &:= \{\varphi \in \mathscr{L} : I\psi \wedge \boxplus(\psi \rightarrow \varphi) \in \Gamma \text{ for some } \psi \in \mathscr{L}_{CPL}\} \end{aligned}$$

Moreover, we define $\sim_{\boxplus}$ and $\rightarrow^c_{\varphi}$ on W^c, respectively, as (1) $\Gamma \sim_{\boxplus} \Delta$ iff $\Gamma[\boxplus] \subseteq \Delta$; and (2) $\Gamma \rightarrow^c \Delta$ iff $\Gamma[I] \subseteq \Delta$. Since $\boxplus$ is an S5 modality, $\sim_{\boxplus}$ is an equivalence relation [7]. For any maximally Log-consistent set Γ, we denote by $[\Gamma]_{\boxplus}$ the equivalence class of Γ induced by $\sim_{\boxplus}$, i.e., $[\Gamma]_{\boxplus} = \{\Delta \in W^c : \Gamma \sim_{\boxplus} \Delta\}$. The following lemma shows that $\rightarrow^c \subseteq \sim_{\boxplus}$.

Lemma 6. *For all* $\Gamma, \Delta \in W^c$*, if* $\Gamma \rightarrow^c \Delta$*, then* $\Gamma \sim_{\boxplus} \Delta$.

Proof. Let $\Gamma, \Delta \in W^c$ such that $\Gamma \rightarrow^c \Delta$, i.e., that $\Gamma[I] \subseteq \Delta$. Let $\varphi \in \Gamma[\boxplus]$. This means that $\boxplus\varphi \in \Gamma$. Then, by $\mathsf{S5}_{\boxplus}$ (since $\vdash \boxplus\varphi \leftrightarrow \boxplus(\top \rightarrow \varphi)$) and Lemma 4, we have that $\boxplus(\top \rightarrow \varphi) \in \Gamma$. Moreover, by Ax1, we have that $I\top \in \Gamma$. Then, by the definition of $\Gamma[I]$, we conclude that $\varphi \in \Gamma[I]$. By the first assumption that $\Gamma[I] \subseteq \Delta$, we obtain $\varphi \in \Delta$. Therefore, $\Gamma \sim_{\boxplus} \Delta$. □

Lemma 7. *Given a mcs* Γ*, for all finite* $\Phi \subseteq \Gamma[I]$*, we have* $\bigwedge \Phi \in \Gamma[I]$.

Proof. Let $\Phi = \{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma[I]$. This means that, for each φ_j with $1 \leq j \leq n$, there is a $\psi_j \in \mathscr{L}_{CPL}$ such that $I\psi_j \wedge \boxplus(\psi_j \to \varphi_j) \in \Gamma$. Thus, $\bigwedge_{1\leq j\leq n} I\psi_j \wedge \bigwedge_{1\leq j\leq n} \boxplus(\psi_j \to \varphi_j) \in \Gamma$. Then, by Ax4, we obtain that $I(\bigwedge_{j\leq n} \psi_j) \in \Gamma$. By $\mathsf{S5}_{\boxplus}$, we also have $\boxplus(\bigwedge_{j\leq n} \psi_j \to \bigwedge_{j\leq n} \varphi_j) \in \Gamma$. Therefore, $\bigwedge \Phi \in \Gamma[I]$. □

Lemma 8. *Given a mcs Γ of* Log, *$\Gamma[I]$ is consistent.*

Proof. Assume, toward contradiction, that $\Gamma[I]$ is not consistent, i.e., $\Gamma[I] \vdash \bot$. This means that there is a finite subset $\Phi = \{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma[I]$ such that $\vdash \bigwedge \Phi \supset \neg\varphi_j$ for some $j \leq n$. By Lemma 7, we have that $\bigwedge \Phi \in \Gamma[I]$, thus, there is a $\psi \in \mathscr{L}_{CPL}$ such that $I\psi \in \Gamma$ and $\boxplus(\psi \to \bigwedge \Phi) \in \Gamma$. Since $\vdash \bigwedge \Phi \to \neg\varphi_j$, by $\mathsf{S5}_{\boxplus}$, we also have $\boxplus(\psi \to \neg\varphi_j) \in \Gamma$. Hence, $\neg\varphi_j \in \Gamma[I]$ too. As $\varphi_j \in \Gamma[I]$, we also have a $\psi' \in \mathscr{L}_{CPL}$ with $I\psi' \in \Gamma$ and $\boxplus(\psi' \to \varphi_j) \in \Gamma$. From $\boxplus(\psi \to \neg\varphi_j) \in \Gamma$ and $\boxplus(\psi' \to \varphi_j) \in \Gamma$, by $\mathsf{S5}_{\boxplus}$, we obtain that $\boxplus(\psi \to \neg\psi') \in \Gamma$. As $I\psi' \in \Gamma$, by Ax2 and Lemma 3.2, $I\overline{\neg\psi'} \in \Gamma$. Therefore, $I\overline{\neg\psi'} \in \Gamma$, $\boxplus(\psi \supset \neg\psi') \in \Gamma$, $I\psi \in \Gamma$, by Ax5, imply that $I\neg\psi' \in \Gamma$, contradicting the consistency of Γ: $I\psi' \in \Gamma$ implies $\neg I\neg\psi' \in \Gamma$, by Ax3. Therefore, $\Gamma[I]$ is consistent. □

Given a mcs Γ_0 of Log, the canonical model for Γ_0 is a tuple $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$, where

- $[\Gamma_0]_{\boxplus}$ is as described above.
- $R^c = \to^c \cap ([\Gamma_0]_{\boxplus} \times [\Gamma_0]_{\boxplus})$
- $\mathscr{P}^c = (P^c, \oplus^c, s^c)$, where
  - $P^c = \mathscr{P}(\mathsf{Prop})$
  - $\oplus^c : P^c \times P^c \to P^c$ such that for all $A, B \in \mathscr{P}(\mathsf{Prop})$, $A \oplus^c B = A \cup B$.
  - $s^c : \mathscr{L} \to \mathscr{P}(P^c)$ such that $s^c(p) = \{A \in P^c \mid p \in A\}$ for all $p \in \mathsf{Prop}$. s extends to the language $\mathscr{L}_{CPL}$ as in Definition 6.
- $f^c : [\Gamma_0]_{\boxplus} \to P^c$ such that $f^c(\Gamma) = \{p \in \mathsf{Prop} \mid I\overline{p} \in \Gamma\}$.
- $V^c : \mathsf{Prop} \to \mathscr{P}([\Gamma_0]_{\boxplus})$ such that $V^c(p) = \{\Gamma \in [\Gamma_0]_{\boxplus} : p \in \Gamma\}$.

Lemma 9. *For all $\varphi \in \mathscr{L}_{CPL}$, $s^c(\varphi) = \{A \subseteq \mathsf{Prop} \mid Var(\varphi) \subseteq A\}$.*

Proof. The proof follows by induction on the structure of φ. The case of constant $\top$ is trivial: $Var(\top) = \emptyset$, hence, $Var(\top) \subseteq A$ for any $A \in P^c$, i.e. $\{A \subseteq \mathsf{Prop} \mid Var(\top) \subseteq A\} = \mathscr{P}(\mathsf{Prop}) = P^c$, which by Definition 6 is equal to $s^c(\top)$. The case for atomic propositions follows directly by the defn. of s^c. Case $\varphi := \neg\psi$ follows from the induction hypothesis (IH) and the fact that $s^c(\neg\psi) = s^c(\psi)$ and $Var(\neg\psi) = Var(\psi)$. We only present the proof for Case: $\varphi := \psi \vee \chi$ and Case: $\varphi := \psi \wedge \chi$ follows similarly:

Case: $\varphi := \psi \vee \chi$:

$$\begin{aligned}
s^c(\psi \vee \chi) &= s^c(\psi) \cap s^c(\chi) && \text{(by the defn. of } s^c\text{)} \\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi) \subseteq A\} \cap \{A \subseteq \mathsf{Prop} \mid Var(\chi) \subseteq A\} && \text{(by IH)} \\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi) \cup Var(\chi) \subseteq A\} && \text{(simple set theory)} \\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi \vee \chi) \subseteq A\} && (Var(\psi \vee \chi) = Var(\psi) \cup Var(\chi))
\end{aligned}$$

□

Corollary 1. *For any $A, B \in P^c$ and $\varphi \in \mathscr{L}_{CPL}$, if $A \subseteq B$ and $A \in s^c(\varphi)$, then $B \in s^c(\varphi)$.*

Lemma 10. *Given a mcs Γ_0, the canonical model $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$ for Γ_0 is a problem-sensitive model.*

Proof. Obviously functions and operations $s^c, \oplus^c, f^c, g^c$ are well-defined. Since $\oplus^c$ is defined as $\cup$ on $\mathscr{P}(\mathsf{Prop})$, it is idempotent, commutative, associative, and it satisfies unrestricted fusion. By Corollary 1, we know that s^c is upward closed. Seriality of R^c follows from Lemmas 8, 5, 6. □

Lemma 11 (Existence lemma)**.** *Given a mcs Γ_0, the canonical model $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$ for Γ_0, a world $\Delta \in [\Gamma_0]_{\boxplus}$ and a formula $\varphi \in \mathscr{L}_{CPL}$, if $\neg I\varphi \in \Delta$, then either $I\overline{\varphi} \notin \Delta$ or there exists a world $\Gamma \in [\Gamma_0]_{\boxplus}$, such that $\Delta R^c \Gamma$ and $\neg\varphi \in \Gamma$.*

Proof. Suppose $\neg I\varphi, I\overline{\varphi} \in \Delta$. Since Δ is maximal and consistent, $I\varphi \notin \Delta$. Then, consider the set $\Delta' = \{\neg\varphi\} \cup \Delta[I]$. This set is consistent. To show this claim, assume otherwise. Then, $\Delta[I] \vdash \varphi$. In other words, there exists a finitely many formulae $\psi_1, \ldots, \psi_n \in \Delta[I]$ such that $\psi_1, \ldots, \psi_n \vdash \varphi$. This implies that $\boxplus(\bigwedge_{1 \leq i \leq n} \psi_i \to \varphi) \in \Delta$. Moreover, by Lemma 7, we also have that $\bigwedge_{1 \leq i \leq n} \psi_i \in \Delta[I]$. Thus, there is $\chi \in \mathscr{L}_{CPL}$ such that $I\chi \wedge \boxplus(\chi \to \bigwedge_{1 \leq i \leq n} \psi_i) \in \Delta$. Then, by $\mathsf{S5}_{\boxplus}$, we have $I\chi \wedge \boxplus(\chi \to \varphi) \in \Delta$. This, together with $I\overline{\varphi}$ and Ax5, implies that $I\varphi \in \Delta$, contradicting consistency of Δ. Therefore, $\Delta' = \{\neg\varphi\} \cup \Delta[I]$ is consistent. By Lemma 5, it can be extended to a msc Γ. Obviously $\neg\varphi \in \Gamma$. Since $\Delta[I] \subseteq \Gamma$ and $\Gamma \in [\Gamma_0]_{\boxplus}$ (by Lemma 6), we also have $\Delta R^c \Gamma$. □

Lemma 12. *Given a mcs Γ_0, the canonical model $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$ for Γ_0, a world $\Delta \in [\Gamma_0]_{\boxplus}$ and a formula $\varphi \in \mathscr{L}_{CPL}$, $I\overline{\varphi} \in \Delta$ iff $Var(\varphi) \subseteq f^c(\Delta)$.*

Proof. $I\overline{\varphi} \in \Delta$ iff I $(\bigwedge_{p \in Var(\varphi)} \overline{p}) \in \Delta$ (by Ax2) iff $I\overline{p} \in \Delta$ for all $p \in Var(\varphi)$ (by Ax4) iff $p \in f^c(\Delta)$ for all $p \in Var(\varphi)$ (by the defn. of f^c) iff $Var(\varphi) \subseteq f^c(\Delta)$. □

Lemma 13 (Truth lemma)**.** *Let Γ_0 be a mcs of Log and $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$ be the canonical model for Γ_0. Then, for all $\varphi \in \mathscr{L}$ and $\Delta \in [\Gamma_0]_{\boxplus}$, we have $\mathscr{M}^c, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

Proof. By induction on φ. Cases of propositional atoms, Boolean connectives and $\boxplus$ are proven as usual. Case of $\varphi := I\psi$.

($\Leftarrow$) Assume $I\psi \in \Delta$. Then, by Ax2, $I\overline{\psi} \in \Delta$. This means, by Lemma 12 that $Var(\psi) \subseteq f^c(\Delta)$. By Lemma 9, this means that $f^c(\Delta) \in s^c(\varphi)$. Now let $\Gamma \in [\Gamma_0]_{\boxplus}$ such that $\Delta R^c \Gamma$, i.e., that $\Delta[I] \subseteq \Gamma$. Observe that $I\psi \in \Delta$ implies that $\psi \in \Delta[I]$ (since $\boxplus(\psi \to \psi) \in \Delta$, by $\mathsf{S5}_{\boxplus}$). By $\Delta[I] \subseteq \Gamma$, this implies that $\psi \in \Gamma$. Then, by IH, we obtain that $\mathscr{M}^c, \Gamma \models \psi$. As Γ has been chosen arbitrarily, we obtain that $R^c(\Delta) \subseteq [\![\psi]\!]_{\mathscr{M}^c}$. Putting everything together, we conclude $\mathscr{M}^c, \Delta \models I\psi$.

($\Rightarrow$) Assume $\mathscr{M}^c, \Delta \models I\psi$. Then, by the semantics, (1) $f^c(\Delta) \in s^c(\psi)$ and (2) for all $\Gamma \in [\Gamma_0]_{\boxplus}$ such that $\Delta R^c \Gamma$, $\mathscr{M}^c, \Gamma \models \psi$. By (1) and Lemma 9, we know that $Var(\psi) \subseteq f^c(\Delta)$. This implies, by Lemma 12, that $I\overline{\psi} \in \Delta$. Now, toward contradiction, assume $I\psi \notin \Delta$, i.e. $\neg I\psi \in \Delta$. By Lemma 11, there exists a world $\Gamma \in [\Gamma_0]_{\boxplus}$, such that $\Delta R^c \Gamma$ and $\neg\psi \in \Gamma$. By IH, we have $\mathscr{M}^c, \Gamma \models \neg\psi$, implying that $\mathscr{M}^c, \Delta \not\models I\psi$, which contradicts our initial assumption. Hence, $I\psi \in \Delta$. □

Theorem 3 (Soundness and completeness)**.** *Given any set of formulae $\Delta \subseteq \mathscr{L}$ and any formula $\varphi \in \mathscr{L}$, the following holds: $\Gamma \vdash \varphi$ iff $\Gamma \models \varphi$.*

Proof. ($\Rightarrow$) By a routine check of axiom validities and validity preservation under the rules of inference of Log. ($\Leftarrow$) By contraposition. Assume $\Gamma \nvdash \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is Log-consistent. By Lemma 5, there exists a maximal consistent set $\Gamma_0 \subseteq \mathscr{L}$, such that $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma_0$. Then, we can build a canonical model $\mathscr{M}^c = \langle [\Gamma_0]_{\boxplus}, R^c, \mathscr{P}^c, f^c, V^c \rangle$. Since $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma_0$, by Lemma 13, $\mathscr{M}^c, \Gamma_0 \models \Gamma \cup \{\neg\varphi\}$. By Lemma 10, $\mathscr{M}^c$ is a problem-sensitive model. Hence, there exists a pointed problem-sensitive model, where Γ is true and φ is false, i.e. $\Gamma \not\models \varphi$. □

Coordinating Beliefs

Abstract

The idea that your credences should be probabilistic is often justified in terms of the observation that any non-probabilistic credences are accuracy-dominated. I provide an analogous accuracy dominance argument for introspection, that is for being certain what your credences are. This raises a puzzle, since you arguably shouldn't be certain whether you have one credence function or a near-identical one. I sketch an answer, building on the thought that credence coordination is noisy.

Richard Roth
rhr8837@nyu.edu

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 65–65, doi:10.4204/EPTCS.437.9

Common p-Belief with Plausibility Measures Extended Abstract

Eric Pacuit Leo Yang
Department of Philosophy
University of Maryland
epacuit@umd.edu leoyang@umd.edu

1 Introduction

In his seminal 1976 paper, Robert Aumann proved a fascinating result, later called the "Agreeing to Disagree" theorem [2]. Suppose that two agents share the same prior probability distribution and update this probability by conditioning on different private information. Aumann showed that if their posterior probabilities of an event E become common knowledge—meaning each agent knows the other's posterior, knows that the other knows it, and so on—then they must assign identical posterior probabilities to E. In other words, two Bayesian agents who share a common prior cannot agree (in the sense of having common knowledge) to disagree (in the sense of holding different posterior probabilities). This result has important implications across multiple disciplines: it underlies "no-trade" theorems in economics [21, 28], helps to provide an epistemic characterization of Nash equilibria in games with more than two players [1], and reveals subtle issues that arise when rational agents exchange information [12, 24, 20].

The literature on Aumann's theorem has focused on identifying which assumptions are necessary to prove the theorem [25, 11, 19]. Various generalizations have emerged, exploring scenarios where, for instance, agents have ambiguous beliefs [16, 17, 30] or use learning rules other than classical Bayesian conditioning [13, 18]. This paper is inspired by three important generalizations of Aumann's theorem:

1. There are several *qualitative* versions of Aumann's theorem. One approach focuses on reformulating Aumann's insights in multi-agent logics of belief [8, 9]. A second approach, developed by Cave [7], Bacharach [5], and Samet [26], reconceptualizes the core problem by shifting focus from probabilities to *abstract decisions*. In Aumann's original theorem, the posterior probability of an agent for a fixed event can be viewed as a random variable mapping states to real numbers in the $[0,1]$ interval. The core idea is to replace these random variables with decision functions mapping worlds to elements of a set $\mathscr{D}$ of "decisions". [1]

2. A second direction of research extends Aumann's theorem to generalizations of classical Bayesian models, including conditional probability structures [29], lexicographic probability systems [4, 3], and imprecise probabilities [31]. By reformulating the agreement theorem within these more general models of beliefs, researchers have shown that Aumann's core insight—that rational agents cannot agree to disagree—remains robust even when relaxing classical probability assumptions.

[1] The resulting qualitative agreement theorem establishes that "like-minded" agents cannot have common knowledge of different decisions, with the concept of being "like-minded" formalized using a variant of Savage's sure-thing principle: For any pair of agents i and j, if i knows that j is at least as knowledgeable as she is, and also knows that j's decision is d, then her decision is also d [27, p. 171].

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 66–79, doi:10.4204/EPTCS.437.10

© E. Pacuit and L. Yang
This work is licensed under the
Creative Commons Attribution License.

3. The third line of research generalizes Aumann's theorem by replacing knowledge with the weaker notion of *p-belief*. An agent p-believes an event E when assigning a probability of at least p to E. In this setting, we can "approximate" common knowledge with *common p-belief*. Monderer and Samet [22], later refined by Neeman [23], proved that if agents sharing a common prior have common p-belief of their posteriors of some event E, then these posteriors must be "close", in the sense that they cannot differ by more than $1-p$. This elegant result shows that if the agents are close to having common knowledge of their posteriors of an event E, then these posteriors must also be close. Then, Aumann's agreement theorem can be viewed as the special case when $p = 1$.

This paper lies in the intersection of these three approaches, aiming to identify the minimal assumptions that any belief model must satisfy to prove a version of the Monderer-Samet-Neeman agreement theorem. To accomplish this, we use Halpern's [10, 14] (conditional) *plausibility measures* to represent the agents' beliefs. These measures provide a very general framework that encompasses many existing formal models of beliefs, including probability measures, possibility measures, ranking functions, and Dempster-Shafer belief functions (see [15] for a discussion).

Our generalized Monderer-Samet-Neeman theorem using conditional plausibility measures provides a unifying perspective that clarifies which properties of belief are essential for demonstrating that common belief of the posteriors of some event implies these posteriors must be close. This deepens our theoretical understanding of agreement theorems while extending the applicability of the Monderer-Samet-Neeman Theorem to settings where classical probabilistic assumptions may not hold.

This extended abstract is organized as follows. Section 2 provides an overview of Aumann's classic agreeing to disagree theorem. We introduce (common) p-belief and discuss the Monderer-Samet-Neeman generalization of Aumann's theorem in Section 3. Section 4 introduces conditional plausibility measures [10, 14] and states our generalizations of the Monderer-Samet-Neeman agreement theorem. The proofs of the results are available in the full version of the paper. We conclude in Section 6 with a brief discussion of some applications of our results.

2 Background: Aumann's Agreement Theorem

In this section, we briefly introduce the formal definitions needed to state and prove Aumann's agreeing to disagree theorem. We begin with models representing the knowledge and beliefs of a group of agents.

Definition 2.1 (Epistemic Probability Model)**.** Suppose that $\mathscr{A}$ is a set of agents. An **epistemic probability model** for $\mathscr{A}$ is a tuple

$$\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle,$$

where

1. W is a non-empty set of states;
2. For each $i \in \mathscr{A}$, Π_i is a partition of W, where for each $w \in W$ and $i \in \mathscr{A}$, we write $\Pi_i(w)$ for the unique element of Π_i containing w, called agent i's **information set at** w;[2]
3. $\mathscr{F} \subseteq \wp(W)$ is a **σ-algebra** on W, meaning that $\mathscr{F}$ is closed under complement and countable union, such that for all $w \in W$ and $i \in \mathscr{A}$, $\Pi_i(w) \in \mathscr{F}$;
4. For each $i \in \mathscr{A}$, P_i is a probability measure on $\langle W, \mathscr{F}\rangle$.

[2] A partition of W is a collection of pairwise disjoint sets whose union is W.

An epistemic probability model has a **common prior** if $P_i = P_j$ for all $i, j \in \mathscr{A}$. To simplify notation, we write $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, P\rangle$ when P is the common prior.

Remark 2.2. We could also allow a different σ-algebra $\mathscr{F}_i$ for each P_i, but this complexity is unnecessary for the purpose of this paper.

Given a state w and an agent $i \in \mathscr{A}$, the set $\Pi_i(w)$ represents i's information at state w. With this interpretation in mind, we can now define the agents' knowledge and posterior beliefs.

Definition 2.3 (Knowledge)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. The **knowledge operator** for agent $i \in \mathscr{A}$ is the function $K_i : \wp(W) \to \wp(W)$ where: for all $E \subseteq W$,

$$K_i(E) = \{w \mid \Pi_i(w) \subseteq E\}.$$

If $w \in K_i(E)$ we say that agent i knows that E in w.[3]

Definition 2.4 (Posterior Belief)**.** Given an epistemic probability model $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$. The **posterior belief for** i **at state** w is the (possibly partial) function $P_{i,w} : \wp(W) \to [0,1]$ where: for all $E \subseteq W$,

$$P_{i,w}(E) = P_i(E \mid \Pi_i(w)) = \frac{P_i(E \cap \Pi_i(w))}{P_i(\Pi_i(w))}.$$

Note that $P_{i,w}(E)$ is undefined if $E \cap F \notin \mathscr{F}$, $\Pi_i(w) \notin \mathscr{F}$, or $P_i(\Pi_i(w)) = 0$.

Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic-probability model. We write $K(E)$ for the event that everyone knows E, and $C(E)$ for the event that E is common knowledge.[4] Formally, *everyone knows* is a function $K : \wp(W) \to \wp(W)$ where: for all $E \subseteq W$,

$$K(E) = \bigcap_{i\in\mathscr{A}} K_i(E).$$

Common knowledge of E is the event where everyone knows E and this fact is completely transparent to all agents. This can be made precise using the concept of a self-evident event:

Definition 2.5 (Self-Evident Event)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. An event $E \subseteq W$ is said to be **self-evident for** i provided that $E \subseteq K_i(E)$. An event is **self-evident** if it is self-evident for all $i \in \mathscr{A}$.

Thus, the event E is self-evident when E is closed with respect to the agents' information partitions: for all $i \in \mathscr{A}$ and all $w \in W$, if $w \in E$, then $\Pi_i(w) \subseteq E$.

Definition 2.6 (Common Knowledge)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. The **common knowledge operator** is the function $C : \wp(W) \to \wp(W)$ such that, for each $E \subseteq W$,

$$C(E) = \{w \mid \text{there is a self-evident set } Z \text{ such that } w \in Z \subseteq E\}.$$

An event E is common knowledge provided that there is a true self-evident event that implies E. We can now state Aumann's agreeing to disagree theorem.

Theorem 2.7 (Agreeing to Disagree Theorem [2])**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, P\rangle$ is an epistemic probability model with a common prior, and for each $i \in \mathscr{A}$, $r_i \in [0,1]$. For any event $E \subseteq W$, if

$$C(\{w \mid P_{i,w}(E) = r_i \text{ for all } i \in \mathscr{A}\}) \neq \emptyset$$

then $r_i = r_j$ for all $i, j \in \mathscr{A}$.

[3] It is well known that each K_i satisfies the **S5** axioms: For all events $E, F \subseteq W$, $K_i(E \cap F) = K_i(E) \cap K_i(F)$, $K_i(E) \subseteq E$, $K_i(E) \subseteq K_i(K_i(E))$, and $\overline{K_i(E)} \subseteq K_i(\overline{K_i(E)})$ (where $\overline{X}$ denotes the complement of X in W).

[4] These definitions can be relativized to any subset of agents. For $G \subseteq \mathscr{A}$, let $K_G(E)$ be the event that everyone in G knows E, and $C_G(E)$ be the event that E is common knowledge among everyone in G. To keep notation minimal, we do not pursue this more general approach in this paper.

3 Common p-Belief

In this section, we present the generalization of Aumann's agreement theorem (Theorem 2.7) first proven by Monderer and Samet [22] and later refined by Neeman [23]. Before stating the theorem, we introduce several key definitions.

Definition 3.1 (p-Belief)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model and $p \in (0,1]$. Agent i's **p-belief** is the function $B_i^p : \mathscr{F} \to \wp(W)$ where: for all $E \subseteq W$,

$$B_i^p(E) = \{w \mid P_{i,w}(E) \geq p\}.$$

If $w \in B_i^p(E)$ we say that i believes E to degree at least p in w.

Remark 3.2. Note that $w \notin B_i^p(E)$ if $P_{i,w}(E)$ is undefined. Moreover, we assume that for all $E \in \mathscr{F}$ and $i \in \mathscr{A}$ and $p \in (0,1]$, $B_i^p(E) \in \mathscr{F}$. Note that we can prove that $\mathscr{F}$ is closed under B_i^p if we assume that W is finite or countable.

The definition of common p-belief closely follows the definition of common knowledge (Definitions 2.5 and 2.6). We first define the **mutual p-belief operator** $B^p : \wp(W) \to \wp(W)$ as follows: for all $E \subseteq W$,

$$B^p(E) = \bigcap_{i\in\mathscr{A}} B_i^p(E).$$

Mutual p-belief of E is the event where all agents in $\mathscr{A}$ believe E with probability at least p. Common p-belief of E is the event where everyone p-believes E and this fact is "partially transparent" to all agents.

Definition 3.3 (p-Self-Evident Event)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. An event $E \subseteq W$ is **p-self-evident for agent** i if $E \subseteq B_i^p(E)$. An event is **p-self-evident** if it is p-self-evident for all $i \in \mathscr{A}$.

Definition 3.4 (Common p-Belief)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. For $p \in (0,1]$, the **common p-belief operator** is the function $CB^p : \wp(W) \to \wp(W)$ where: for all $E \subseteq W$,

$$CB^p(E) = \{w \mid \text{there is a } p\text{-self-evident event } Z \text{ such that } w \in Z \subseteq B^p(E)\}.$$

So, an event E is common p-believed when there is a true p-self-evident event that implies everyone p-believes E. Next, we state a useful lemma that is needed to prove the Moderer–Samet–Neeman generalization of the agreement theorem.

Lemma 3.5. Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, (P_i)_{i\in\mathscr{A}}\rangle$ is an epistemic probability model. Let $E \subseteq W$ be an event, $p \in (0,1]$, and $r_i \in [0,1]$ for each $i \in \mathscr{A}$. Then, for each $i \in \mathscr{A}$:

1. Π_i partitions $B_i^p(E)$, i.e., $B_i^p(E) = \bigcup\{X \mid X \in \Pi_i \text{ and } X \cap B_i^p(E) \neq \emptyset\}$.
2. $CB^p(E) \subseteq B_i^p CB^p(E)$.
3. If $B_i^p(E) \neq \emptyset$, then $P_i(E \mid B_i^p(E)) \geq p$.
4. Let $X = \{w \mid P_{i,w}(E) = r_i \text{ for all } i \in \mathscr{A}\}$. If $CB^p(X) \neq \emptyset$, then $P_i(E \mid B_i^p CB^p(X)) = r_i$.

Proof (Sketch). 1. For any $E \subseteq W$ and $w, w' \in W$, if $w' \in \Pi_i(w)$, then $P_i(E \mid \Pi_i(w)) = P_i(E \mid \Pi_i(w'))$. Thus, if $X \in \Pi_i$ and $X \cap B_i^p(E) \neq \emptyset$, then $X \subseteq B_i^p(E)$. This immediately implies that $B_i^p(E) = \bigcup\{X \mid X \in \Pi_i \text{ and } X \cap B_i^p(E) \neq \emptyset\}$.

2. Suppose that $w \in CB^p(E)$. Then there is a p-self-evident event Z such that $w \in Z \subseteq B^p(E)$. Since Z is p-self-evident, we have $Z \subseteq B_i^p(Z)$ and so, $P_i(Z \mid \Pi_i(w)) \geq p$. Clearly, $Z \subseteq CB^p(E)$, and so, $P_i(CB^p(E) \mid \Pi_i(w)) \geq P_i(Z \mid \Pi_i(w)) \geq p$. Thus, $w \in B_i^p CB^p(E)$.

3. By part 1., $B_i^p(E) = \bigcup\{X \mid X \in \Pi_i \text{ and } X \cap B_i^p(E) \neq \emptyset\}$. First of all, note that since $P_i(\Pi_i(w)) > 0$ for all $w \in W$, we have that if $B_i^p(E) \neq \emptyset$, then $P_i(B_i^p(E)) > 0$. Let $\mathscr{E}_i = \{X \mid X \in \Pi_i \text{ and } X \cap B_i^p(E) \neq \emptyset\}$. Then, for each $Y \in \mathscr{E}_i$, we have $P_i(E \mid Y) \geq p$. Hence,

$$P_i(E \mid B_i^p(E)) = \sum_{Y \in \mathscr{E}_i} P(E \mid Y) \cdot P(Y \mid B_i^p(E)) \geq \sum_{Y \in \mathscr{E}_i} p \cdot P(Y \mid B_i^p(E)) = p \sum_{Y \in \mathscr{E}_i} P(Y \mid B_i^p(E)) = p.$$

4. Let $w \in CB^p(X)$. Since $p > 0$, we have $\Pi_i(w) \cap X \neq \emptyset$ for all $i \in \mathscr{A}$. So $P_i(E \mid \Pi_i(w)) = r_i$ for all $i \in \mathscr{A}$. Thus, $w \in X$. Therefore, $CB^p(X) \subseteq X$. By similar reasoning, we obtain that $P_{i,w}(E) = r_i$ for all $w \in B_i^p CB^p(X)$. Then, we obtain the desired result using reasoning similar to part 3. since Π_i partitions $B_i^p CB^p(X)$.

□

We can now state the Monderer-Samet-Neeman agreement theorem which, for brevity, we call the **MSN-Theorem**.

Theorem 3.6 (MSN-Theorem [22, 23])**.** Suppose that $\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, P \rangle$ is an epistemic probability model with a common prior, $p \in (0,1]$, and for each $i \in \mathscr{A}$, $r_i \in [0,1]$. For any event $E \subseteq W$, if

$$CB^p(\{w \mid P_{i,w}(E) = r_i \text{ for all } i \in \mathscr{A}\}) \neq \emptyset$$

then $|r_i - r_j| \leq 1 - p$ for all $i, j \in \mathscr{A}$.

The original proof of Monderer, Samet, and Neeman's uses Lemma 3.5 and the following fact about probability measures P: For all events A, B, C, if $A \subseteq B \subseteq C$, $P(B) > 0$, and $P(C) > 0$, then $P(A \mid B) \cdot P(B \mid C) = P(A \mid C)$. In the full version of the paper, we give a second proof that avoids explicit mention of multiplication, appealing instead to the following Lemma:

Lemma 3.7. Suppose that $\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, P \rangle$ is an epistemic probability model with a common prior. For all events $E, X \subseteq W$ and $p \in (0,1]$, if $CB^p(X) \neq \emptyset$, then for all $i, j \in \mathscr{A}$, either

1. $P(CB^p(X) - E \mid B_i^p CB^p(X)) \leq P(CB^p(X) - E \mid B_j^p CB^p(X))$, and
2. $P(E \cap CB^p(X) \mid B_i^p CB^p(X)) \leq P(E \cap CB^p(X) \mid B_j^p CB^p(X))$.

or

3. $P(CB^p(X) - E \mid B_i^p CB^p(X)) \geq P(CB^p(X) - E \mid B_j^p CB^p(X))$, and
4. $P(E \cap CB^p(X) \mid B_i^p CB^p(X)) \geq P(E \cap CB^p(X) \mid B_j^p CB^p(X))$.

Proof. Suppose that $\langle W, (\Pi_i)_{i \in \mathscr{A}}, P \rangle$ is an epistemic probability model with a common prior. Let $E, X \subseteq W$ and $p \in (0,1]$. We first note that for all $i, j \in \mathscr{A}$, either

$$P(CB^p(X) \mid B_i^p CB^p(X)) \geq P(CB^p(X) \mid B_j^p CB^p(X)) \text{ or}$$
$$P(CB^p(X) \mid B_j^p CB^p(X)) \geq P(CB^p(X) \mid B_i^p CB^p(X)).$$

Without loss of generality, let $0 \neq P(CB^p(X) \mid B_i^p CB^p(X)) \geq P(CB^p(X) \mid B_j^p CB^p(X))$. Since $CB^p(X) \subseteq B^p CB^p(X)$, this means that:

$$0 \neq \frac{P(CB^p(X))}{P(B_i^p CB^p(X))} \geq \frac{P(CB^p(X))}{P(B_j^p CB^p(X))}.$$

This implies that:

$$P(B_i^p CB^p(X)) \leq P(B_j^p CB^p(X))$$

Thus, for all $F \subseteq B_i^p CB^p(X) \cap B_j^p CB^p(X)$, we have $P(F \mid B_i^p CB^p(X)) \geq P(F \mid B_j^p CB^p(X))$. In particular, this holds for $F = CB^p(X) - E$ and $F = CB^p(X) \cap E$. This completes the proof. □

As we will see in Section 5, these different proofs suggest two ways to generalize the MSN-Theorem.

4 Epistemic Plausibility Models

In this section, we introduce **epistemic plausibility models**. These models generalize epistemic probability models (Definition 2.1) by replacing probabilities with the more general notion of a *conditional plausibility measure*.

To define a conditional plausibility measure, we first fix a non-empty set D of **plausibility values** satisfying the following two conditions. First, it is partially ordered by a relation $\leq_D \subseteq D \times D$. That is, $\leq_D$ is reflexive (for all $d \in D$, $d \leq_D d$), transitive (for all $d, d', d'' \in D$, if $d \leq_D d'$ and $d' \leq_D d''$, then $d \leq_D d''$), and antisymmetric (for all $d, d' \in D$, if $d \leq_D d'$ and $d' \leq_D d$, then $d = d'$). Second, D contains minimal and maximal elements $\bot_D$ and $\top_D$ respectively, such that $\bot_D \leq_D d \leq_D \top_D$ for all $d \in D$. As usual, we write $d <_D d'$ when $d \leq_D d'$ and $d \neq d'$. When D is clear from context, we omit subscripts and write $\leq$, $<$, $\top$, and $\bot$.

A conditional plausibility measure for an agent i assigns an element from $d \in D$ to a pair of events X and Y, denoted $Pl_i(X \mid Y) = d$ meaning that according to i, the plausibility of X conditional on Y is d. The domain of a conditional plausibility measure is a **Popper algebra**. Suppose that W is a nonempty set. A collection of subsets $\mathscr{F} \subseteq \wp(W)$ (where $\wp(W)$ is the powerset of W) is called an **algebra** when $W \in \mathscr{F}$, $\mathscr{F}$ is closed under union (if $X, Y \in \mathscr{F}$, then $X \cup Y \in \mathscr{F}$), and closed under complement (if $X \in \mathscr{F}$ then $\overline{X} \in \mathscr{F}$, where $\overline{X}$ denotes the complement of X in W). An algebra $\mathscr{F}$ is called a **σ-algebra** when it is closed under countable unions. Then, a set $\mathscr{F} \times \mathscr{F}'$ is a **Popper algebra** if:

1. $\mathscr{F}$ is an σ-algebra;
2. $\mathscr{F}' \subseteq \mathscr{F}$; and
3. for $X \in \mathscr{F}'$, if $X \subseteq Y$ and $Y \in \mathscr{F}$, then $Y \in \mathscr{F}'$.

The intended interpretation is that $\mathscr{F}$ is the set of events that are assigned plausibility values, and $\mathscr{F}'$ is the set of events that can be learned. A conditional plausibility measure satisfies very minimal properties:

Definition 4.1 (Conditional Plausibility Structure)**.** Suppose that D is a partially ordered set of plausibility values (with ordering $\leq_D$). A **conditional plausibility structure** on D is a tuple $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i \in \mathscr{A}} \rangle$, where $W \neq \emptyset$, $\mathscr{F} \times \mathscr{F}'$ is a Popper algebra on W, and for each $i \in \mathscr{A}$, $Pl_i : \mathscr{F} \times \mathscr{F}' \to D$ satisfies: For all $X, Y \in \mathscr{F}$ and $E \in \mathscr{F}'$:

(CP1) $Pl_i(E \mid E) = \top$,

(CP2) $Pl_i(\emptyset \mid E) = \bot$,

(CP3) If $X \subseteq Y$, then $Pl_i(X \mid E) \leq Pl_i(Y \mid E)$, and

(CP4) $Pl_i(X \mid E) = Pl_i(X \cap E \mid E)$.

A conditional plausibility structure is **acceptable** if for any $E \in \mathscr{F}$ and $F \in \mathscr{F}'$, we have $E \cap F \in \mathscr{F}'$ whenever $Pl_i(E \mid F) \neq \bot$ for some $i \in \mathscr{A}$. A conditional plausibility structure is said to have a **common prior** if $Pl_i = Pl_j$ for all $i, j \in \mathscr{A}$. In this case, we write it as $\langle W, \mathscr{F}, \mathscr{F}', Pl \rangle$.

The properties **CP1**–**CP4** are minimal conditions satisfied by most formal models of belief [**?**, 15]. Property **CP1** states that any event conditioned on itself receives the maximal plausibility value, while **CP2** requires assigning the minimal plausibility value to the empty set $\emptyset$. Property **CP3** ensures that the plausibility measures are monotonic, and **CP4** captures a minimal condition that any sensible notion of conditional belief should satisfy. **Acceptability** generalizes the following property of standard probability measures P: if $P(E \mid F) \neq 0$, then $P(E \cap F) \neq 0$; and so, conditioning on $E \cap F$ is well-defined.

We are now ready to define analogues of epistemic probability models (Definition 2.1), posterior belief (Definition 2.4), and p-belief operators (Definition 3.1), replacing probability measures with conditional plausibility measures. So, for instance, we write $B_i^d(E)$ for the event that i assigns the plausibility value at least d (according to the ordering $\geq_D$) to the event E.

Definition 4.2 (Epistemic Plausibility Model)**.** [5] An **epistemic plausibility model** is a tuple

$$\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, \mathscr{F}', (Pl_i)_{i \in \mathscr{A}} \rangle,$$

where $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i \in \mathscr{A}} \rangle$ is a conditional plausibility model such that for all $w \in W$ and $i \in \mathscr{A}$, $\Pi_i(w) \in \mathscr{F}'$, and for each $i \in \mathscr{A}$, Π_i is a partition of W. We say that an epistemic plausibility model has a common prior if its conditional plausibility structure has a common prior.

Definition 4.3 (Posterior Belief)**.** Suppose that $\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, \mathscr{F}', (Pl_i)_{i \in \mathscr{A}} \rangle$ is an epistemic plausibility model. The **posterior belief for** i **at state** w is the function $Pl_{i,w} : \mathscr{F} \to D$, where, for all $X \in \mathscr{F}$,

$$Pl_{i,w}(X) = Pl(X \mid \Pi_i(w)).$$

Remark 4.4. For each agent $i \in \mathscr{A}$ and state $w \in W$, the posterior belief function $Pl_{i,w}$ is a **plausibility measure**: a function μ from a σ-algebra to a partially ordered set D satisfying $\mu(W) = \top$ (cf. **CP1**), $\mu(\emptyset) = \bot$ (cf. **CP2**), and monotonicity (cf. **CP3**): if $X \subseteq Y$, then $\mu(X) \leq \mu(Y)$.

Definition 4.5 (d-Belief Operator)**.** Suppose that $\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, \mathscr{F}', (Pl)_{i \in \mathscr{A}} \rangle$ is an epistemic plausibility model for a set D of plausibility values. For each $d \in D$ and $i \in \mathscr{A}$, a d**-belief operator** for i is the function $B_i^d : \mathscr{F} \to \wp(W)$, where for all $X \subseteq W$,

$$B_i^d(X) = \{w \mid Pl_i(X \mid \Pi_i(w)) \geq d\}.$$

Remark 4.6. In the remainder of this paper, we assume the following: In any epistemic plausibility model $\langle W, (\Pi_i)_{i \in \mathscr{A}}, \mathscr{F}, \mathscr{F}', (Pl_i)_{i \in \mathscr{A}} \rangle$, for all $E \in \mathscr{F}$, we have that $B_i^d(E) \in \mathscr{F}'$.

With these definitions in place, the notion of common p-belief (Definition 3.4) extends straightforwardly to epistemic plausibility models; we omit the explicit definition here for brevity.

[5] Cf. the epistemic probability models defined in [6].

5 MSN-Theorems in Epistemic Plausibility Structures

In this section, we present two distinct generalizations of the MSN-Theorem (Theorem 3.6) using epistemic plausibility models (Definition 4). Our goal is to identify the minimal properties necessary to prove these generalized versions of the MSN-Theorem.

The first step is to equip plausibility values with a notion of *addition*. Indeed, both Aumann's Theorem (Theorem 2.7) and the MSN-Theorem (Theorem 3.6) crucially rely on conditions involving the addition of plausibility values. Formally, this involves specifying a function $\oplus : D \times D \to D$ that maps each pair $\langle d, f\rangle \in D \times D$ to an element $d \oplus f \in D$, satisfying at least the following minimal conditions:

Definition 5.1 (Additive Conditional Plausibility Structure)**.** A conditional plausibility structure $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ is said to be an **additive conditional plausibility structure** when there is a (partial) function $\oplus : D \times D \to D$ such that for each $i \in \mathscr{A}$,[6]

(A1) For all $a, b, c \in D$, if $\langle a, b\rangle \in dom(\oplus)$, then $c \leq b$ iff $\langle a, c\rangle \in dom(\oplus)$ and $a \oplus c \leq a \oplus b$. Similarly, if $\langle b, a\rangle \in dom(\oplus)$, then $c \leq b$ iff $\langle c, a\rangle \in dom(\oplus)$ and $c \oplus a \leq b \oplus a$.

(A2) For all $a \in D$, there exists $b \in D$ such that $a \oplus b = \top$.

(A3) For all $X, Y \in \mathscr{F}$ and $Z \in \mathscr{F}'$, if $X \cap Y = \emptyset$, then $Pl_i(X \cup Y \mid Z) = Pl_i(X \mid Z) \oplus Pl_i(Y \mid Z)$,

(A4) For all $E \in \mathscr{F}$ and sets of disjoint sets $\mathscr{S} \subseteq \mathscr{F}'$, if $a \leq Pl_i(E \mid S) \leq b$ for all $S \in \mathscr{S}$, then

$$a \leq Pl_i(E \mid \bigcup \mathscr{S}) \leq b.$$

We say that the epistemic plausibility model $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ is **additive** when the underlying conditional plausibility structure $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ is additive.

Properties **A1** and **A2** concern the binary operation $\oplus$. Property **A1** states that $\oplus$ is monotonic in both arguments, while **A2** requires each element to have an additive inverse. Properties **A3** and **A4** describe additive properties of the conditional plausibility measures. Specifically, **A3** requires that each conditional plausibility measure Pl_i is finitely additive. Property **A4** captures a crucial requirement needed to prove the agreement theorems (Theorems 2.7 and 3.6).

Example 5.2. Let $\langle W, \mathscr{F}, \mathscr{F}', Pl\rangle$ be a conditional plausibility structure where:

- $W = \{w_1, w_2, w_3\}$;
- $D = \{\frac{n}{6} \mid n \in \{0, 1, 2, \ldots, 6\}\}$;
- $\oplus = +$;
- $\mathscr{F} = \wp(W)$ and $\mathscr{F}' = \wp(W) - \{\emptyset\}$;
- For all $E \in \mathscr{F}$ and $F \in \mathscr{F}'$, $Pl(E \mid F) = \frac{|E\cap F|}{|F|}$.

It is not difficult to see that this is an additive conditional plausibility structure by checking that it satisfies **CP1** - **CP4** and **A1** - **A4**.

There are two important consequences of additive conditional plausibility models. First, we obtain an analogue of Lemma 3.5. The statement is omitted since it is an easy variation of Lemma 3.5, and its proof is an immediate consequence of **A4**. The second consequence is that we can introduce a notion of subtraction. This is essential because the MSN-Theorem is stated in terms of differences between plausibility values. The existence of a subtraction is a direct consequence of **A1**:

[6] If F is a (partial) function, we write $dom(F)$ for the domain of F.

Lemma 5.3 (Existence of a Subtraction)**.** Suppose that $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ is an additive conditional plausibility structure on a set D of plausibility values. For all $a,b \in D$, if there is a $c \in D$, such that $a \oplus c = b$, then this c is unique. When such a c exists, we denoted it by $b \ominus a$.

We first note some useful algebraic properties that immediately follow from Lemma 5.3 and the definition of an additive conditional plausibility structure (Definition 5.1).

Lemma 5.4. Suppose that $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ is an additive conditional plausibility structure on a set D of plausibility values. Then,

1. For all $a,b \in D$, $\top \ominus a \leq \top \ominus b$ iff $a \geq b$,
2. For all $a,b,c,d \in D$, if $a \leq b \leq c \leq d$ and both $c \ominus b$ and $d \ominus a$ exist, then $c \ominus b \leq a \ominus d$,
3. If $\oplus$ is **associative** (for all $a,b,c \in D$, $a \oplus (b \oplus c) = (a \oplus b) \oplus c$), then for all $a,b \in D$,

$$(\top \ominus a) \ominus b = \top \ominus (a \oplus b).$$

To motivate our generalizations, we first present a new characterization of the original MSN-Theorem.

Theorem 5.5. Given an additive conditional plausibility structure with common prior $\mathscr{M} = \langle W, \mathscr{F}, \mathscr{F}', Pl\rangle$, the following are equivalent.

1. For all additive epistemic plausibility models based on $\mathscr{M}$, if the agents commonly d-believe their posteriors of an event $E \in \mathscr{F}$, then their posteriors differ by less than $\top \ominus d$ (if the difference exists).
2. For all $A,B \in \mathscr{F}'$, if $Pl(A \mid B) \geq d$ and $Pl(B \mid A) \geq d$, then for any event $E \in \mathscr{F}$, if $\top \ominus d$ exists, then
$$Pl(E \mid A) \ominus Pl(E \mid B) \leq \top \ominus d.$$

Condition 1 is the MSN-theorem. Our strategy to generalize the MSN theorem is to ensure that condition 2 holds for appropriate choices of A and B.

Our first generalization of the MSN-Theorem follows the original proof given in [22] and [23]. This proof makes essential use of a property involving the *multiplication* of plausibility values. Therefore, before stating the theorem, we introduce a multiplication operation on the set of plausibility values.

Definition 5.6 (Conditional Plausibility Structure with Addition and Multiplication)**.** An additive conditional plausibility structure $\langle W, \mathscr{F}, \mathscr{F}', (Pl_i)_{i\in\mathscr{A}}\rangle$ has a multiplication when there is a (partial) function $\otimes : D \times D \to D$ such that:

(M1) For all $a,b,c \in D$, if $\langle a,b\rangle \in dom(\otimes)$, $\langle a,c\rangle \in dom(\otimes)$, and $b \geq c$, then $a \otimes b \geq a \otimes c$,

(M2) For all $a,b,c \in D$, if $a \oplus b$ is defined, then $(a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c)$

(M3) For all $i \in \mathscr{A}$, $X \in \mathscr{F}$ and $Y,Z \in \mathscr{F}'$, if $X \subseteq Y \subseteq Z$, then $Pl_i(X \mid Z) = Pl_i(X \mid Y) \otimes Pl_i(Y \mid Z)$

Property **M1** requires that the multiplication operation $\otimes$ is monotonic, while property **M2** requires that $\otimes$ distributes over the addition operation $\oplus$. The essential property required for the proof is **M3**. We can now state our generalized version of the MSN-Theorem.

Theorem 5.7 (Generalized MSN-Theorem with Multiplication)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, \mathscr{F}', Pl\rangle$ is an additive epistemic plausibility model with a common prior. Further, suppose that the model is acceptable, the addition operator $\oplus$ is associative, and there is a multiplication $\otimes$. Let $E \in \mathscr{F}$, for each $i \in \mathscr{A}$, $r_i \in D$, and $d \in D$. If

$$CB^d(\{w \mid Pl_{i,w}(E) = r_i \text{ for all } i \in \mathscr{A}\}) \neq \emptyset$$

then for all $i,j \in \mathscr{A}$, if $r_i \ominus r_j$ exists, then $r_i \ominus r_j \leq \top \ominus (\top \otimes d)$.

Remark 5.8. The antecedent of this theorem implies Condition 2 of Theorem 5.5.

There is a second generalization of the MSN-Theorem that does not require a multiplication operation. Instead, it relies on two additional properties of an additive conditional plausibility model with a common prior $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, \mathscr{F}', Pl\rangle$.[7]

(CP6) For all $E_1, E_2 \in \mathscr{F}$ and $A_1, A_2 \in \mathscr{F}'$, if $E_1, E_2 \subseteq A_1 \cap A_2$ and $\bot \neq Pl(E_1 \mid A_1) \geq Pl(E_1 \mid A_2)$, then $Pl(E_2 \mid A_1) \geq Pl(E_2 \mid A_2)$,

(CP7) For all $E \in \mathscr{F}$ and $i, j \in \mathscr{A}$, either

$$Pl(E \mid B_i^p(E)) \leq Pl(E \mid B_j^p(E)) \text{ or } Pl(E \mid B_j^p(E)) \leq Pl(E \mid B_i^p(E)).$$

Property **CP6** captures the following idea: if $i \in \mathscr{A}$, $E_1 \in A_1 \cap A_2$ and $\bot \neq Pl_i(E_1 \mid A_1) \geq Pl_i(E_1 \mid A_2)$, then A_1 must be considered "less plausible" than A_2. Moreover, if A_1 is less plausible than A_2, then it follows that $Pl(E_2 \mid A_1) \geq Pl(E_2 \mid A_2)$ for all events $E_2 \in A_1 \cap A_2$. Property **CP7** requires that the reliabilities of beliefs held by any two agents can always be compared. The relationship between **M1**-**M3** and **CP6**-**CP7** is clarified in Appendices A and B.

We are now ready to state our main theorem: a version of the MSN-Theorem that does not rely on multiplication.

Theorem 5.9 (Generalized MSN-Theorem without Multiplication)**.** Suppose that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, \mathscr{F}', Pl\rangle$ is an additive epistemic plausibility model with a common prior that satisfies **CP6** and **CP7**. Let $E \in \mathscr{F}$, for each $i \in \mathscr{A}$, $r_i \in D$, and $d \in D$. If

$$CB^d(\{w \mid Pl_{i,w}(E) = r_i \text{ for all } i \in \mathscr{A}\}) \neq \emptyset$$

then for all $i, j \in \mathscr{A}$, if $r_i \ominus r_j$ exists, then $(r_i \ominus r_j) \leq (\top \ominus d)$.

Remark 5.10. **CP6** - **CP7** guarantee that Condition 2 of Theorem 5.5 is satisfied for suitable A and B.

Remark 5.11. The proofs of Theorems 5.7 and 5.9 do not depend on whether condition **A3** holds when $Pl_i(X \mid Z) = Pl_i(Y \mid Z) = \bot$. This observation is useful in some applications.

Both versions of the MSN-Theorem are important. Some additive epistemic plausibility structures satisfy conditions **CP6** - **CP7** but do not allow for a multiplication operation satisfying **M1** - **M3**. Conversely, other structures have a suitable multiplication operation $\otimes$ satisfying **M1** - **M3** but fail to satisfy conditions **CP6** - **CP7**. Nevertheless, certain natural conditions on multiplication imply property **CP6**. The proofs of both generalizations of the MSN-Theorem are available in the full version of the paper.

6 Conclusion and Applications

This extended abstract presents two generalizations of the Monderer–Samet–Neeman (MSN) agreement theorem, using additive conditional plausibility measures. These measures provide a unified framework that incorporates many existing formal models of belief. The first generalization closely follows Monderer, Samet, and Neeman's original proof, relying on a multiplication operation, while the second generalization avoids this requirement. Together, these two results clarify the minimal assumptions required to establish MSN-type agreement theorems in various formal belief models.

To illustrate the significance of these assumptions, we briefly examine several existing models of belief (a complete discussion is left to the full version of the paper). Our generalized MSN-theorems apply, in particular, to conditional probability structures [29], lexicographic probability structures [4, 3], and order models of belief [8]. When these conditions are not satisfied, the corresponding MSN-theorem typically fails or becomes trivial for the given belief model.

[7] We are numbering them as (**CP6**) and (**CP7**) to avoid confusion with Halpern's (**CP5**) from [14, 15].

References

[1] Robert Aumann & Adam Brandenburger (1995): *Epistemic Conditions for Nash Equilibrium*. Econometrica 63(5), pp. 1161–1180, doi:10.2307/2171725.

[2] Robert J. Aumann (1976): *Agreeing to Disagree*. Annals of Statistics 4(6), pp. 1236–1239, doi:10.1214/aos/1176343654.

[3] Christian W. Bach & Jérémie Cabessa (2023): *Lexicographic Agreeing to Disagree and Perfect Equilibrium*. Journal of Mathematical Economics 109, p. 102908, doi:10.1016/j.jmateco.2023.102908.

[4] Christian W. Bach & Andrés Perea (2013): *Agreeing to Disagree with Lexicographic Prior Beliefs*. Mathematical Social Sciences 66(2), pp. 129–133, doi:10.1016/j.mathsocsci.2013.03.004.

[5] Michael Bacharach (1985): *Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge*. Journal of Economic Theory 37(1), pp. 167–190, doi:10.1016/0022-0531(85)90035-3.

[6] Alexandru Baltag & Sonja Smets (2008): *Probabilistic dynamic belief revision*. Synthese 165(2), pp. 179–202, doi:10.1007/s11229-008-9369-8.

[7] Jonathan A.K. Cave (1983): *Learning to agree*. Economics Letters 12(2), pp. 147–152, doi:10.1016/0165-1765(83)90126-X.

[8] Cédric Dégremont & Olivier Roy (2012): *Agreement Theorems in Dynamic-Epistemic Logic*. Journal of Philosophical Logic 41(4), pp. 735–764, doi:10.1007/s10992-012-9236-4.

[9] Lorenz Demey (2014): *Agreeing to disagree in probabilistic dynamic epistemic logic*. Synthese 191, pp. 409–438, doi:10.1007/s11229-013-0280-6.

[10] Nir Friedman & Joseph Y. Halpern (1995): *Plausibility Measures: A User's Guide*. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., p. 175–184, doi:10.5555/2074158.2074179.

[11] John Geanakoplos (1994): *Common Knowledge*. In: Handbook of Game Theory with Economic Applications, chapter 40, 2, Elsevier, pp. 1437–1496, doi:10.1016/S1574-0005(05)80072-4.

[12] John D. Geanakoplos & Heraklis M. Polemarchakis (1982): *We Can't Disagree Forever*. Journal of Economic Theory 28(1), pp. 192–200, doi:10.1016/0022-0531(82)90099-0.

[13] Itzhak Gilboa, Larry Samuelson & David Schmeidler (2022): *Learning (to Disagree?) in Large Worlds*. Journal of Economic Theory 199, p. 105166, doi:10.1016/j.jet.2020.105166.

[14] Joseph Y. Halpern (2001): *Plausibility Measures: A General Approach for Representing Uncertainty*. In: International Joint Conference on Artificial Intelligence, 17, pp. 1474–1483. Available at `https://www.cs.cornell.edu/home/halpern/papers/plaus.pdf`.

[15] Joseph Y. Halpern (2017): *Reasoning about Uncertainty*. The MIT press, doi:10.7551/mitpress/10951.001.0001.

[16] Joseph Y. Halpern & Willemien Kets (2014): *A Logic for Reasoning about Ambiguity*. Artificial Intelligence 209, pp. 1–10, doi:10.1016/j.artint.2013.12.003.

[17] Joseph Y. Halpern & Willemien Kets (2015): *Ambiguous Language and Common Priors*. Games and Economic Behavior 90, pp. 171–180, doi:10.1016/j.geb.2015.01.007.

[18] Matthias Hild, Richard Jeffrey & Mathias Risse (1997): *Agreeing to Disagree: Harsanyi and Aumann*. Vienna Circle Institute Yearbook 5, pp. 109–115, doi:10.1007/978-94-017-1654-3_9.

[19] Harvey Lederman (2015): *People with Common Priors Can Agree to Disagree*. The Review of Symbolic Logic 8(1), pp. 11–45, doi:10.1017/S1755020314000380.

[20] Lucie Ménager (2006): *Consensus, Communication and Knowledge: An Extension with Bayesian Agents*. Mathematical Social Sciences 51(3), pp. 274–279, doi:10.1016/j.mathsocsci.2006.01.003.

[21] Paul R. Milgrom & Nancy Stokey (1982): *Information, Trade and Common Knowledge*. Journal of Economic Theory 26(1), pp. 17–27, doi:10.1016/0022-0531(82)90046-1.

[22] Dov Monderer & Dov Samet (1989): *Approximating Common Knowledge with Common Beliefs*. Games and Economic Behavior 1(2), pp. 170–190, doi:10.1016/0899-8256(89)90017-1.

[23] Zvika Neeman (1996): *Approximating Agreeing to Disagree Results with Common p-Beliefs*. Games and Economic Behavior 12(1), pp. 162–164, doi:10.1006/game.1996.0011.

[24] Rohit Parikh & Paul Krasucki (1990): *Communication, Consensus, and Knowledge*. Journal of Economic Theory 52(1), pp. 178–189, doi:10.1016/0022-0531(90)90073-S.

[25] Ariel Rubinstein & Asher Wolinsky (1990): *On the Logic of "Agreeing to Disagree" Type Results*. Journal of Economic Theory 51(1), pp. 184–193, doi:10.1016/0022-0531(90)90057-Q.

[26] Dov Samet (2010): *Agreeing to Disagree: The Non-Probabilistic Case*. Games and Economic Behavior 69(1), pp. 169–174, doi:10.1016/j.geb.2008.09.032.

[27] Dov Samet (2022): *The Impossibility of Agreeing to Disagree: An Extension of the Sure-Thing Principle*. Games and Economic Behavior 132, pp. 390–399, doi:10.1016/j.geb.2022.01.016.

[28] James K. Sebenius & John Geanakoplos (1983): *Don't Bet on It: Contingent Agreements with Asymmetric Information*. Journal of the American Statistical Association 78(382), pp. 424–426, doi:10.2307/2288651.

[29] Elias Tsakas (2018): *Agreeing to Disagree with Conditional Probability Systems*. The BE Journal of Theoretical Economics 18(2), pp. 1–7, doi:10.1515/bejte-2016-0182.

[30] Vassili Vergopoulos & Antoine Billot (2024): *Agreement, Dialogue and Trade Under Ambiguity*, doi:10.2139/ssrn.4745253. Available at `https://ssrn.com/abstract=4745253`. Available at SSRN.

[31] Jiji Zhang, Hailin Liu & Teddy Seidenfeld (2018): *Agreeing to Disagree and Dilation*. International Journal of Approximate Reasoning 101, pp. 150–162, doi:10.1016/j.ijar.2018.07.006.

A Multiplication Implies (CP6)

Proposition A.1. Given an additive epistemic plausibility model with common prior $\langle W, (\Pi_i)_{i\in\mathscr{A}}, \mathscr{F}, \mathscr{F}', Pl\rangle$. If there exists a function $\otimes : D \times D \to D$ that satisfies (**M1**), (**M3**) and

(M4) For all $a,b,c \in D$, if $a \neq \bot$ and $a \otimes b \geq a \otimes c$, then $b \geq c$.

Then (**CP6**) is satisfied.

Proof. Let $E_1, E_2 \in \mathscr{F}$, $A_1, A_2 \in \mathscr{F}'$, and $E_1, E_2 \subseteq A_1 \cap A_2$. Assume that $\bot \neq Pl(E_1 \mid A_1) \geq Pl(E_1 \mid A_2)$. Given (**M3**), this amounts to

$$\bot \neq Pl(E_1 \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_1) \geq Pl(E_1 \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_2)$$

Now we have $Pl(E_1 \mid A_1 \cap A_2) \neq \bot$, since otherwise $Pl(E_1 \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_1) = Pl(\emptyset \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_1) = Pl(\emptyset \mid A_1) = \bot$. Then, by (**M4**), we can obtain

$$Pl(A_1 \cap A_2 \mid A_1) \geq Pl(A_1 \cap A_2 \mid A_2)$$

which, by (**M1**) implies that

$$Pl(E_2 \mid A_1) = Pl(E_2 \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_1) \geq Pl(E_2 \mid A_1 \cap A_2) \otimes Pl(A_1 \cap A_2 \mid A_2) = Pl(E_2 \mid A_2)$$

□

B Independence of (CP6)-(CP7) and (M1)-(M3)

First, we show that even (**M1**)-(**M4**) do not imply (**CP7**).

Example B.1. Given two standard epistemic probability structures with common prior $\langle W, (\Pi_i)_{i\in\{1,2\}}, p\rangle$ and $\langle W, (\Pi_i)_{i\in\{1,2\}}, q\rangle$, where

- W=$\{w_1, w_2, w_3\}$
- $\Pi_1 = \{\{w_1, w_2\}, \{w_3\}\}$
- $\Pi_2 = \{\{w_1\}, \{w_2, w_3\}\}$
- $p(w_1) = p(w_2) = 0.2$, $p(w_3) = 0.6$
- $q(w_1) = q(w_2) = 0.4$, $q(w_3) = 0.2$

Then, we let $D = [0,1] \times [0,1]$, and $\leq_D, \oplus, \otimes, Pl$ be such that: For all $a_1, a_2, b_1, b_2 \in [0,1]$ and $A, B \subseteq W$

$$\langle a_1, b_1\rangle \leq_D \langle a_2, b_2\rangle \iff a_1 \leq a_2 \text{ and } b_1 \leq b_2$$

$$\langle a_1, b_1\rangle \oplus \langle a_2, b_2\rangle = \langle a_1 + a_2, b_1 + b_2\rangle$$

$$\langle a_1, b_1\rangle \otimes \langle a_2, b_2\rangle = \langle a_1 \cdot a_2, b_1 \cdot b_2\rangle$$

$$Pl(A \mid B) = \langle p(A \mid B), q(A \mid B))\rangle$$

It can be verified that $\langle W, (\Pi_i)_{i\in\mathscr{A}}, Pl\rangle$ is an additive epistemic plausibility structure with a common prior, that $\oplus$ is associative, and that $\otimes$ satisfies (**M1**)-(**M4**). Now, suppose that $d = \langle 0.1, 0.1\rangle$ and $E = \{w_2\}$. It can be verified that $Pl(E \mid B_1^p(E)) = \langle \frac{1}{2}, \frac{1}{2}\rangle$ and $Pl(E \mid B_2^d(E)) = \langle \frac{1}{4}, \frac{2}{3}\rangle$, which violates (**CP7**).

Now we show that (**CP6**)-(**CP7**) do not imply that (**M1**)-(**M3**) can be satisfied.

Example B.2. Given a single-agent standard epistemic probability structures with common prior $\langle W, \Pi, p\rangle$ where

- W=$\{w_1, w_2, w_3\}$
- $\Pi = \{W\}$
- $p(w_1) = p(w_2) = 0.25$ and $p(w_3) = 0.5$

Then we define

$$p'(A \mid B) = \begin{cases} 0.26 & \text{if } A = \{w_1\} \text{ and } B = W \\ 0.24 & \text{if } A = \{w_2\} \text{ and } B = W \\ 0.76 & \text{if } A = \{w_1, w_3\} \text{ and } B = W \\ 0.74 & \text{if } A = \{w_2, w_3\} \text{ and } B = W \\ p(A \mid B) & \text{otherwise} \end{cases}$$

It can be verified that $\langle W, \Pi, p'\rangle$ is an additive epistemic plausibility structure with a common prior that satisfies (**CP6**)-(**CP7**). However, $p'(\{w_1\} \mid \{w_1, w_2\}) = p'(\{w_2\} \mid \{w_1, w_2\})$ but $p'(\{w_1\} \mid \{w_1, w_2, w_3\}) \neq p'(\{w_2\} \mid \{w_1, w_2, w_3\})$, which implies that no function $\otimes$ can satisfy (**M3**).

Distributed Knowing How

Bin Liu
Department of Philosophy, Peking University
liub85@stu.pku.edu.cn

Yanjing Wang
Department of Philosophy, Peking University
y.wang@pku.edu.cn

Distributed knowledge is a key concept in the standard epistemic logic of knowledge-that. In this paper, we propose a corresponding notion of distributed knowledge-how and study its logic. Our framework generalizes two existing traditions in the logic of know-how: the individual-based multi-step framework and the coalition-based single-step framework. In particular, we assume a group can accomplish more than what its individuals can jointly do. The distributed knowledge-how is based on the distributed knowledge-that of a group whose multi-step strategies derive from distributed actions that subgroups can collectively perform. As the main result, we obtain a sound and strongly complete proof system for our logic of distributed knowledge-how, which closely resembles the logic of distributed knowledge-that in both the axioms and the proof method of completeness.

1 Introduction

One charming feature of epistemic logic is the formalization of group knowledge, such as common knowledge and distributed knowledge [7]. In particular, distributed knowledge plays an important role in various applications of epistemic logic, which also brings technical challenges that go beyond the bisimulation-invariant modal logic (cf. e.g., [1]).

While epistemic logic mainly focuses on propositional *de dicto* knowledge (knowledge-that), there is an increasing interest in studying *de re* knowledge, such as knowledge-how/why/what, and so on [20]. It is natural to ask what the suitable notions of group knowledge-wh are [23]. In [21, 8], various common/distributed notions of *knowing whether* have been proposed and studied. In this work, we try to propose a notion of distributed *knowing how* and study its logic.

We build our work by first combining two logical frameworks of knowing how. Initiated in [22, 24], the planning-based multi-step framework was inspired by the philosophical discussion of knowledge-how and the theory of automated planning in AI. In such a logic, one knows how to achieve φ means there is a *plan* consisting of her own actions such that the agent knows that the plan can be executed and will always terminate and guarantee φ. The plans can be linear, branching, or even a program written in a programming language [15]. The second framework is a coalition-based one-step setting initiated in [18, 19], where a coalition knows how to achieve φ means there is a *joint action* of the coalition such that the coalition knows that this joint action can make sure φ no matter what others do. To some extent, the second framework captures a simple form of distributed knowledge-how as [14] remarked, but in this work we aim for a much more general notion which requires the power of both frameworks. The differences between these two existing approaches and our new "mixed" approach can be summarized below:

approach	know-that	know-how	actions
planning-based	individual	multi-step	individual
coalition-based	distributed	one-step	joint
mixed	distributed	multi-step	group

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 80–97, doi:10.4204/EPTCS.437.11

© Liu & Wang
This work is licensed under the
Creative Commons Attribution License.

The last column specifies what actions are used. The first approach assumes that the actions by individuals can already move the states. The second approach presupposes that the transitions only happen due to the joint actions of *all* the agents, which is suitable for game-theoretical settings. In contrast, actions in our approach are somehow the generalization of the two, featuring what we call *group actions* that may or may not be decomposable into individual actions. To see the need for our mixed approach, let us see a few examples of knowledge-how intuitively distributed in a group based on various types of group actions, such that only the group knows how to achieve a certain goal without each member knowing how to do it.

The first example is about actions that agents can only do together as a group.

Example 1 (Couch moving) *No one can single-handedly move a heavy couch through a narrow doorway around a corner. However, if two people work together, they know how to do it through a synchronized group action of one person lifting and maneuvering one end of the couch and the other handling the opposite end of the couch.*

The group action here in the example belongs to the group as a whole and is not decomposable into *independent* individual actions. We will take such actions as *atomic* in the framework. In contrast with the joint actions in the coalition-based game-like framework (and ATL-like epistemic frameworks [11, 12]), other agents outside the group do not affect the result of such a group action, which is similar to the individual actions in the planning-based approach.

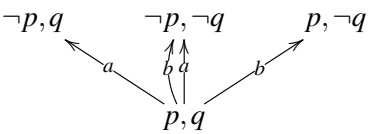
Moreover, distributed knowledge-how arise when a group can use actions from its members or subgroups.

Example 2 (Theorem proving) *In collaborations on mathematical work, it often happens that two researchers do not know how to prove a theorem independently, but together they do. For example, i knows how to prove a lemma, and j knows how to prove a theorem given the lemma is true, then together they should know how to prove the theorem by first proving the lemma using i's knowledge-how and then proving the theorem using j's knowledge-how. Intuitively, the knowledge-how to prove the theorem is (sequentially) distributed in the group of these two researchers.*

This example shows that the distributed knowledge-how of a group may be based on a plan making use of actions inherited from its members or subgroups.

Yet another source of distributed knowledge-how is based on actions of a group that are *decomposable* into *independent* actions of its members and subgroups. It resembles the joint actions in the coalition-based approach, but not necessarily for the full set of agents.

Example 3 (Joint treatment) *Suppose a patient has two health problems represented by p and q. Doctor* 1 *knows how to cure p by treatment a but is not sure whether it can cure q, and Doctor* 2 *knows how to treat q by treatment b but is not sure whether it can cure p. Then, assuming the treatments are independent (not affecting each other), by doing both a and b respectively, the doctors distributedly know how to cure both problems (making sure* $\neg p \wedge \neg q$*).*



Compared to the group action in Example 1, the effect of the joint action in Example 3 is determined and decomposable by the effects of the independent composing actions. As we will see later, the set of outcomes induced by such a joint action is the *intersection* of the outcomes caused by each composing action.

From these three examples, we can summarize the following observations respectively:

O1 A group may have some *extra* irreducible actions that none of its proper subgroups can do.

O2 The available actions of a group at least include all the actions of the subgroups.

O3 The group can perform decomposable *joint actions*, whose outcomes are the intersection of the outcomes of the composing actions (by subgroups).

Thus, there are three types of group actions relevant for distributed knowledge-how: (I) irreducible group actions belonging to the whole group but not less; (II) actions inherited from the subgroups; (III) the joint actions computed as the "intersections" of the composing actions.

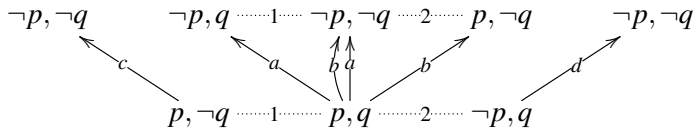
We will formalize the above idea by first giving "atomic" group actions A_G of type (I) for each group G (including singleton groups as individuals). By O1, if $H \subsetneq G$ then $A_H \cap A_G = \emptyset$. Next, we compute the closure A_G^* (call it the set of distributed actions) recursively with respect to all the inherited actions of type (II) and the joint actions of type (III). In particular, the joint actions of group G are composed of actions belonging to any non-singleton partial partition of the group, not just a tuple of actions from each member. This reflects the fact that a task force may be divided into subteams with their own know-how. We assume that for joint actions, each person can remain idle but cannot be used twice *at the same time*[1], thus requiring the partial partition (partition of a subset).

Below is an example to illustrate the definition of group actions, where we write $A_{\{i\}}$ as A_i for brevity and present the joint actions according to a fixed ordering of agents and groups.

Example 4 *Suppose there are two agents* $1,2$*, the atomic group actions of type (I) are defined as* $A_1 = \{a,b\}$*,* $A_2 = \{c\}$*,* $A_{\{1,2\}} = \{d\}$*. Then the closures* A_G^* *are computed as:* $A_i^* = A_i$ *for singleton sets* $\{i\}$*;* $A_{\{1,2\}}^* = A_{\{1,2\}} \cup (A_1 \cup A_2) \cup (A_1^* \times A_2^*) = \{d, a, b, c, \langle a,c\rangle, \langle b,c\rangle\}$ *where* $A_{\{1,2\}}$ *is the set of type (I) group actions, and* $A_1 \cup A_2$ *is the set of inherited actions of type (II), and* $(A_1^* \times A_2^*)$ *is the set of intersecting joint actions of type (III). Note that if there are more than two agents, then the actions of type (III) become complicated, e.g.,* $A_{\{1,2,3\}}^*$ *should include* $A_{\{1,2\}}^* \times A_3^*$*,* $A_1^* \times A_{\{2,3\}}^*$*,* $A_{\{1,3\}}^* \times A_2^*$*,* $A_1^* \times A_2^*$*,* $A_2^* \times A_3^*$*,* $A_1^* \times A_3^*$*, and* $A_1^* \times A_2^* \times A_3^*$*, i.e., it includes the joint actions of any non-singleton partial partition of the group representing how the work can be distributed among subgroups.*

Distributed knowledge-how is then based on multi-step strategies that make use of actions in A_G^* and distributed knowledge-that. We say a (possibly singleton) group G (distributedly) knows how to achieve φ if there is a strategy σ making use of A_G^* such that it is distributed knowledge of G that σ will terminate and guarantee φ. Consider the following example extending Example 3:

Example 5 (Distributed knowledge-that and -how) *On top of Example 3, we further assume that the two doctors have some uncertainty regarding the patient's problems: Doctor 1 is sure the patient has p but not sure about q, and Doctor 2 is sure the patient has q but not sure about p. Furthermore, if* $p \wedge \neg q$ *then Doctor 1 should do treatment c (not a) to cure p, and if* $\neg p \wedge q$ *Doctor 2 should do treatment d (not b) to cure q. The situation can be formalized in the following model where* $A_1 = \{a,c\}$*,* $A_2 = \{b,d\}$ *and* $A_{\{1,2\}} = \emptyset$*. The dotted lines represent the epistemic indistinguishability relations of the agents, and we omit the reflexive arrows.*



[1]Otherwise, an agent may have the unrealistic ability to execute two actions at the same time. For example, if we assume a and b in Example 3 are both actions for Doctor 1, then Doctor 1 can team up with herself, to execute $\langle a,b\rangle$ to guarantee the outcome $\neg p \wedge \neg q$, which may not be possible.

Intuitively, at world (p,q), Doctor 1 does not know how to achieve $\neg p$ as she does not have a uniform move to do so over the indistinguishable worlds $(p,\neg q)$ and (p,q). Similarly, Doctor 2 does not know how to achieve $\neg q$. However, the group of the two doctors 1 and 2 distributedly knows how to make sure $\neg p \wedge \neg q$ by doing the joint action $\langle a,b\rangle \in A^*_{\{1,2\}}$ based on their distributed knowledge, which leaves them with a sole possible world (p,q).

In this paper, we propose such a mixed framework of distributed knowing how, based on the group actions explained above and the notion of strategies proposed in [9]. The main technical contribution is a sound and complete proof system of the logic with intuitive axioms, whose completeness proof is based on the unraveling of both the epistemic relations and the action transitions. One crucial difference from the systems of coalition-based know-how logics is that the *cooperation axiom* that plays a key role there is no longer valid here due to the multi-step setting. In the following, Section 2 lays out the language and semantics for our logic of distributed knowledge-how. Section 3 gives a proof system. In Section 4, we prove the completeness. Finally, we conclude and discuss future work in Section 5.

2 Language and semantics

We present the following syntax and semantics given the set P of proposition letters and the finite set $I = \{i_0, \cdots, i_n\}$ of agents.[2]

Definition 6 (Language) *The language* **DKH** *is defined by the following BNF where $p \in P$ and G is a subset of I called group:*

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{K}_G\varphi \mid \mathsf{Kh}_G\varphi$$

Definition 7 (Model) *A model $\mathscr{M}$ is a tuple $\left\langle S, \{\sim_i\}_{i\in I}, \{A_G\}_{G\subseteq I}, \{\xrightarrow{a} \mid a \in \bigcup_{G\subseteq I} A_G\}, V\right\rangle$, where*

- *S is a set of states;*
- *$\sim_i$ is an equivalence relation on S for each $i \in I$;*
- *A_G is a set of atomic group actions for each $G \subseteq I$ such that: $G \subsetneq H$ implies $A_G \cap A_H = \emptyset$; $A_\emptyset = \emptyset$;*
- *$\xrightarrow{a}$ is a binary relation on S for each $a \in \bigcup_{G\subseteq I} A_G$, called the transition relation of a;*
- *$V : P \to \mathscr{P}(S)$ is a valuation function.*

We use A^+_G to denote $\bigcup_{G'\subseteq G} A_{G'}$ for any $G \subseteq I$, thus A^+_I is the set of all atomic group actions and $G \subseteq H$ implies $A^+_G \subseteq A^+_H$. Note that the actions only appear in the models, but not in the language, and the transitions can be non-deterministic. The models in this paper closely align with those in [16] with some notable distinctions. First, instead of the set of actions of each agent, we have the set of actions for each group. Moreover, as mentioned in the introduction, we will compute the closures of these atomic action sets, which serve as the basis for distributed knowledge-how. Second, the requirement that $G \subsetneq H$ implies $A_G \cap A_H = \emptyset$, reflects the intuition behind the type (I) group actions that are only available to the whole group and not less.[3]

Definition 8 (Distributed indistinguishability) *For each group G, $\sim_G$ is a binary relation on S such that $s \sim_G t$ iff $s \sim_i t$ for each $i \in G$.*

[2] The requirement for the finiteness of the set I of agents is solely for the convenience of denoting distributed actions and has no effect on the logic (valid formulas).

[3] Although this requirement is conceptually important, technically, we think it can be dropped without changing the logic, because the closures of action sets are the same with or without the requirement. To tell the difference in logic, we may need to talk about concrete actions in the language.

It is routine to show that distributed indistinguishability relations are equivalence relations.

Proposition 9 *For any group G, $\sim_G$ is an equivalence relation on S.*

Therefore, we can define equivalence classes w.r.t. $\sim_G$.

Definition 10 *For any group G and $s \in S$, we use $[s]_G$ to denote the equivalence class $\{t \in S \mid s \sim_G t\}$, and use $[S]_G$ to denote the collection of all the equivalence classes on S w.r.t. $\sim_G$.*

Note that by definition $s \sim_\emptyset t$ for any $s,t \in S$, which implies that $[s]_\emptyset = S$ for any $s \in S$.

We will now define the set of *distributed actions* A^*_G as a closure of A_G recursively, which includes the three types of group actions as discussed in the introduction and illustrated in Example 4. To make the technical definition concise, we divide the actions in A^*_G into two parts: (1) atomic group actions of G and its subgroups, and (2) joint actions in the form of $\langle d_0, \cdots, d_n\rangle$. Since the order of $d_0, \cdots, d_n$ is insignificant in our setting, we fix an ordering $\prec$ below on *mutually disjoint groups* to avoid generating redundant joint actions, which will not affect the logic technically. Recall that the set of all agents is $I = \{i_0, \cdots, i_n\}$. For any nonempty group $G = \{i_{n_0}, \cdots, i_{n_k}\}$, let $\min G = \min\{n_0, \cdots, n_k\}$ and let $G \prec H$ iff $\min G < \min H$. This will always give us a strict ordering over a set of mutually disjoint groups of agents.

Definition 11 (Distributed actions) *For each group G, its distributed action set $A^*_G := A^+_G \cup \{\langle d_0, \cdots, d_n\rangle \in A^*_{G_0} \times \cdots \times A^*_{G_n} \mid \{G_0, \cdots, G_n\}$ is a non-trivial partial partition of G and $G_0 \prec \cdots \prec G_n\}$, where a non-trivial partial partition of a set X is a partition of a (not necessarily proper) subset of X such that it is not a singleton set. It follows that $A^*_\emptyset = \emptyset$ and $A^*_i = A_i$ for each $i \in I$.*

Note that joint actions of subgroups of a group G are also inherited because a non-trivial partial partition of a subgroup is also a non-trivial partial partition of G. Thus we have the following monotonicity for both distributed indistinguishability and distributed action sets.

Proposition 12 *$G \subseteq H$ implies $\sim_H \subseteq \sim_G$ and $A^*_G \subseteq A^*_H$.*

PROOF Suppose that $G \subseteq H$. $\sim_H \subseteq \sim_G$ is straightforward by definition. We show $A^*_G \subseteq A^*_H$ as follows. If $G = \emptyset$, then $A^*_G = \emptyset$, then we have $A^*_G \subseteq A^*_H$. If $G = \{i\}$ where $i \in I$, then $A^*_G = A_i \subseteq A^+_H \subseteq A^*_H$. If G is a multi-agent group, then suppose that $d \in A^*_G$. If $d \in A^+_G$, then since $A^+_G \subseteq A^+_H$, we have $d \in A^+_H \subseteq A^*_H$. If $d = \langle d_0, \cdots, d_n\rangle \in A^*_{G_0} \times \cdots \times A^*_{G_n}$ where $\{G_0, \cdots, G_n\}$ is a non-trivial partial partition of G and $G_0 \prec \cdots \prec G_n$, since $G \subseteq H$, $\{G_0, \cdots, G_n\}$ is also a non-trivial partial partition of H. Then $\langle d_0, \cdots, d_n\rangle \in A^*_H$. Therefore $A^*_G \subseteq A^*_H$. □

By Proposition 12, A^*_I is the set of all distributed actions and $A^+_I \subseteq A^*_I$. We now define distributed transitions according to the observation O3 in the introduction.

Definition 13 (Distributed transition) *For each $d = \langle d_0, \cdots, d_n\rangle \in A^*_I$, we define the distributed transition relation $\xrightarrow{d} := \bigcap_{0 \le k \le n} \xrightarrow{d_k}$.*

For any nonempty $X, Y \subseteq S$ and $d \in A^*_I$, we use $X \xrightarrow{d} Y$ to indicate that there is some $s \in X$ and some $t \in Y$ such that $s \xrightarrow{d} t$. A distributed transition relation could be empty, which indicates that there are conflicting actions among subgroups. For example, the actions of opening and closing the same door cannot constitute an executable distributed action.

Although a joint action $d = \langle d_0, \cdots, d_n\rangle \in A^*_I$ may well use other joint actions as d_i, they can be in effect reduced to the intersections of atomic transitions, e.g., the transitions for $\langle\langle a, b\rangle, c\rangle$ are the same as the transitions for $\langle a, b, c\rangle$, as they are computed by taking the intersection of $\xrightarrow{a}$, $\xrightarrow{b}$, and $\xrightarrow{c}$. The following observation plays an important role in the completeness proof.

Proposition 14 *For any nonempty group G and $d \in A_G^*$, there exist $a'_0 \in A_{G'_0}, \cdots, a'_m \in A_{G'_m}$ such that $\{G'_0, \cdots, G'_m\}$ is a partial partition of G and $\xrightarrow{d} = \bigcap_{0 \le k \le m} \xrightarrow{a'_k}$.*

PROOF We show by induction on the number of agents in G. If $G = \{i\}$ where $i \in I$, then $d \in A_i$ and $\{\{i\}\}$ is a partial partition of $\{i\}$. Assume that G is a multi-agent group. If $d \in A_G^+$, then there exists $G' \subseteq G$ such that $d \in A_{G'}$, then $\{G'\}$ is a partial partition of G. If $d = \langle d_0, \cdots, d_n \rangle \in A_{G_0}^* \times \cdots \times A_{G_n}^*$ where $\{G_0, \cdots, G_n\}$ is a non-trivial partial partition of G and $G_0 \prec \cdots \prec G_n$, then $n > 0$ and $\emptyset \neq G_k \subsetneq G$ for $0 \le k \le n$. By induction hypothesis, for $0 \le k \le n$ there exist $a_0^k \in A_{G_0^k}, \cdots, a_{m_k}^k \in A_{G_{m_k}^k}$ such that $\{G_0^k, \cdots, G_{m_k}^k\}$ is a partial partition of G_k and $\xrightarrow{d_k} = \bigcap_{0 \le j \le m_k} \xrightarrow{a_j^k}$. Then $\{G_0^0, \cdots, G_{m_0}^0, \cdots, G_0^n, \cdots, G_{m_n}^n\}$ is a partial partition of G. Since $\xrightarrow{\langle d_0, \cdots, d_n \rangle} = \bigcap_{0 \le k \le n} \xrightarrow{d_k}$, the proposition holds. □

Observe that in the last case of the proof, $|\{G_0^0, \cdots, G_{m_0}^0, \cdots, G_0^n, \cdots, G_{m_n}^n\}| > 1$ because $n > 0$. Therefore, we have the following corollary, which will be helpful for the completeness proof:

Corollary 15 *For any multi-agent group G and $d \in A_G^* \setminus A_G^+$, there exist $a_0 \in A_{G_0}, \cdots, a_m \in A_{G_m}$ such that $m > 0$, $\{G_0, \cdots, G_m\}$ is a partial partition of G and $\xrightarrow{d} = \bigcap_{0 \le k \le m} \xrightarrow{a_k}$.*

The following definitions are generalizations of the counterparts in [9].

Definition 16 (Executability) *For any nonempty $X \subseteq S$, we say that $d \in A_I^*$ is executable on X, if for each $s \in X$ there exists t such that $s \xrightarrow{d} t$.*

Definition 17 (Strategy) *A strategy of group G is a partial function $\sigma_G : [S]_G \to A_G^*$ such that $\sigma_G([s]_G)$ is executable on $[s]_G$. Particularly, for each group, the empty function is also a strategy, called the empty strategy.*

By definition, the only strategy for the empty group is the empty strategy.

Definition 18 (Execution) *Given a strategy σ_G of group G w.r.t. a model $\mathscr{M}$, a* possible execution *of σ_G is a possibly infinite nonempty sequence of equivalence classes $\delta = [s_0]_G[s_1]_G \cdots$ such that $[s_j]_G \xrightarrow{\sigma_G([s_j]_G)} [s_{j+1}]_G$ for all $0 \le j < |\delta| - 1$. If the execution is a finite sequence $[s_0]_G \cdots [s_n]_G$, we call $[s_n]_G$ the leaf-node, and $[s_j]_G (0 \le j < n)$ an inner-node w.r.t. this execution. If it is infinite, then all $[s_j]_G (j \in \mathbb{N})$ are inner-nodes. A possible execution of σ_G is* complete *if it is infinite or its leaf-node is not in $\mathtt{dom}(\sigma_G)$. We use $\mathtt{CELeaf}(\sigma_G, s)$ to denote the set of all leaf-nodes of all complete executions of σ_G starting from $[s]_G$, and $\mathtt{CEInner}(\sigma_G, s)$ to denote the set of all inner-nodes of all complete executions of σ_G starting from $[s]_G$.*

Now we formally define the semantics of know-how $\mathsf{Kh}_G\varphi$ which intuitively says that there is a strategy σ_G such that G distributedly knows that σ will terminate and guarantee φ.

Definition 19 (Semantics) *Given a model $\mathscr{M}$, for any state $s \in S$ and any formula $\varphi \in$ **DKH***

$\mathscr{M}, s \vDash \top$		*always holds*
$\mathscr{M}, s \vDash p$	*iff*	$s \in V(p)$, *where* $p \in P$
$\mathscr{M}, s \vDash \neg\varphi$	*iff*	$\mathscr{M}, s \nvDash \varphi$
$\mathscr{M}, s \vDash \varphi \wedge \psi$	*iff*	$\mathscr{M}, s \vDash \varphi$ *and* $\mathscr{M}, s \vDash \psi$
$\mathscr{M}, s \vDash \mathsf{K}_G\varphi$	*iff*	$\mathscr{M}, s' \vDash \varphi$ *for all* $s' \in [s]_G$
$\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$	*iff*	*there is a strategy σ_G of G such that:*
		1. $[t]_G \subseteq [\![\varphi]\!]$ *for all* $[t]_G \in \mathtt{CELeaf}(\sigma_G, s)$, *and*
		2. all its complete executions starting from $[s]_G$ are finite.

where $[\![\varphi]\!] = \{s \in S \mid \mathscr{M}, s \vDash \varphi\}$.

Note that when G is a singleton, the semantics of Kh_G is exactly as in [9]. It is also worth noting that the key axiom Cooperation:[4] $\mathsf{Kh}_G\varphi \wedge \mathsf{Kh}_H\varphi \to \mathsf{Kh}_{G\cup H}(\varphi \wedge \psi)$ (G and H are disjoint) in the coalition-based approach such as [19] is *not* valid in our setting, which demonstrates that our mixed framework is by no means a trivial combination. There are two reasons for the invalidity. First of all, as our know-how is based on *multi-step* strategies, the strategies behind the know-how of G and H may not be synchronizable in the number of steps of reaching the goals, which is fundamentally different from the one-step coalition-based setting where joint actions always end at the same time. Moreover, even if we restrict to one-step actions, the distributed transition relations (as intersections) can be empty, that is, joint actions may be non-executable, which is different from the setting in [19] and makes it uncertain whether the two actions can be combined into an executable joint action.

Moreover, the invalidity also shows that distributed knowledge-how cannot be conceptualized as simply putting together individual/subgroup knowledge-how (and their consequences), which separates it from distributed knowledge-that, where Cooperation is valid.

As a technical remark, the empty group provides us with additional expressive power of a *universal modality*, which will play a vital role in the axiomatization proposed in the next section.

Proposition 20 $\mathscr{M}, s \vDash \mathsf{K}_\emptyset\varphi$ *iff* $\mathscr{M}, s \vDash \mathsf{Kh}_\emptyset\varphi$ *iff* $\mathscr{M}, t \vDash \varphi$ *for any* $t \in S$.

PROOF It is obvious that $\mathscr{M}, s \vDash \mathsf{K}_\emptyset\varphi$ iff $\mathscr{M}, t \vDash \varphi$ for any $t \in S$. We show $\mathscr{M}, s \vDash \mathsf{Kh}_\emptyset\varphi$ iff $\mathscr{M}, t \vDash \varphi$ for any $t \in S$.

($\Rightarrow$) Since $\mathscr{M}, s \vDash \mathsf{Kh}_\emptyset\varphi$, then there is a strategy $\sigma_\emptyset$ for the empty group such that

1. $[t]_\emptyset \subseteq [\![\varphi]\!]$ for all $[t]_\emptyset \in \mathtt{CELeaf}(\sigma_\emptyset, s)$, and
2. all its complete executions starting from $[s]_\emptyset$ are finite.

By definition, the only strategy for the empty group is the empty strategy $\emptyset$. Since $[s]_\emptyset \notin \mathtt{dom}(\emptyset)$, the only execution of the empty strategy $\emptyset$ starting from $[s]_\emptyset$ is $\delta_\emptyset = [s]_\emptyset$, then $\mathtt{CELeaf}(\sigma_\emptyset, s) = \{[s]_\emptyset\}$. Therefore, $S = [s]_\emptyset \subseteq [\![\varphi]\!]$.

($\Leftarrow$) Since $\mathscr{M}, t \vDash \varphi$ for any $t \in S$, we have $\mathscr{M}, s \vDash \mathsf{K}_\emptyset\varphi$. We show a stronger result that $\mathscr{M}, s \vDash \mathsf{K}_G\varphi$ implies $\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$ for any group G. Consider the empty strategy $\emptyset$ for G. Since $[s]_G \notin \mathtt{dom}(\emptyset)$, the only execution of the empty strategy $\emptyset$ starting from $[s]_G$ is $\delta_G = [s]_G$, then $\mathtt{CELeaf}(\emptyset, s) = \{[s]_G\}$. Since $\mathscr{M}, s \vDash \mathsf{K}_G\varphi$, we have $[s]_G \subseteq [\![\varphi]\!]$. Therefore, $\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$. □

3 Axiomatization

Proof system $\mathbb{SDKH}$[5]

Axioms			
TAUT	axioms of propositional logic[6]	AxKtoKh	$\mathsf{K}_G\varphi \to \mathsf{Kh}_G\varphi$
DISTK	$\mathsf{K}_G\varphi \wedge \mathsf{K}_G(\varphi \to \psi) \to \mathsf{K}_G\psi$	AxEmpKhtoK	$\mathsf{Kh}_\emptyset\varphi \to \mathsf{K}_\emptyset\varphi$
T	$\mathsf{K}_G\varphi \to \varphi$	AxKhtoKKh	$\mathsf{Kh}_G\varphi \to \mathsf{K}_G\mathsf{Kh}_G\varphi$
4	$\mathsf{K}_G\varphi \to \mathsf{K}_G\mathsf{K}_G\varphi$	AxEmpMono	$\mathsf{K}_\emptyset(\varphi \to \psi) \to \mathsf{K}_\emptyset(\mathsf{Kh}_G\varphi \to \mathsf{Kh}_G\psi)$

[4] An alternative form is $\mathsf{Kh}_G(\varphi \to \psi) \to (\mathsf{Kh}_H\varphi \to \mathsf{Kh}_{G\cup H}\psi)$.

[5] Formulas φ, ψ are from the language defined in Definition 6.

[6] All instantiations in language **DKH** of axioms of propositional logic, to be more precise.

5	$\neg \mathsf{K}_G\varphi \to \mathsf{K}_G\neg \mathsf{K}_G\varphi$	AxKhbot	$\mathsf{Kh}_G\bot \to \bot$
AxKMono	$\mathsf{K}_G\varphi \to \mathsf{K}_H\varphi$, where $G \subseteq H$	AxKhtoKhK	$\mathsf{Kh}_G\varphi \to \mathsf{Kh}_G\mathsf{K}_G\varphi$
AxKhMono	$\mathsf{Kh}_G\varphi \to \mathsf{Kh}_H\varphi$, where $G \subseteq H$	AxKhKh	$\mathsf{Kh}_G\mathsf{Kh}_G\varphi \to \mathsf{Kh}_G\varphi$
Rules			
MP	$\dfrac{\varphi, \varphi \to \psi}{\psi}$	NECK	$\dfrac{\vdash \varphi}{\vdash \mathsf{K}_G\varphi}$

$\mathbb{SDKH}$ contains all $\mathbb{S}5$ axioms for K. AxKMono and AxKhMono are monotonicity axioms for distributed knowledge-that and distributed knowledge-how. They express that a group knows/knows how φ whenever its subgroup knows/knows how φ. Their validity is due to the monotonicity of distributed indistinguishability and distributed action sets shown by Proposition 12. We will provide the validity of AxKhMono later. AxKhKh is a key axiom w.r.t. multi-step planning, revealing the compositional nature of knowing how. Its validity is highly non-trivial as in [9], and we will provide a sketch of the proof later.

Note that the logic is *not* normal, for example, $\mathsf{Kh}_G\varphi \wedge \mathsf{Kh}_G\psi \to \mathsf{Kh}_G(\varphi \wedge \psi)$ is invalid. In [9] the monotonicity rule MONOKh is proposed to fill the gap left by the absence of the K axiom for Kh, which states that if $\varphi \to \psi$ is provable then $\mathsf{Kh}\varphi \to \mathsf{Kh}\psi$ is provable. In $\mathbb{SDKH}$ we use a stronger *axiom* AxEmpMono which reflects the monotonicity of Kh_G at the *model level*. The rule MONOKh can be derived from AxEmpMono, NECK and T.

As for other important axioms, AxKtoKh says if a group already knows φ then it knows how to achieve φ by doing nothing (the empty strategy). AxEmpKhtoK helps us reduce $\mathsf{Kh}_\emptyset$ to $\mathsf{K}_\emptyset$. AxKhtoKKh is the positive introspection axiom for Kh_G, whose validity comes from uniformity of strategies, and the negative introspection of know-how is derivable. AxKhbot encodes part of the termination condition within the semantics, for it excludes strategies that have no terminating executions. AxKhtoKhK says that φ is known after executing the know-how strategy, as required in the semantics, which is usually also assumed in discussions of contingent planning (cf. [17]).

Proposition 21 (Axiom AxKhMono) *$\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$ implies $\mathscr{M}, s \vDash \mathsf{Kh}_H\varphi$, where $G \subseteq H$.*

PROOF Suppose $\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$. Then there exists a strategy σ_G such that:

1. $[t]_G \subseteq [\![\varphi]\!]$ for all $[t]_G \in \mathtt{CELeaf}(\sigma_G, s)$, and
2. all its complete executions starting from $[s]_G$ are finite.

If $[s]_G \notin \mathtt{dom}(\sigma_G)$, the only execution of σ_G starting from $[s]_G$ is $\delta_G = [s]_G$, which is also complete and finite. Then $\mathtt{CELeaf}(\sigma_G, s) = \{[s]_G\}$, implying $[s]_G \subseteq [\![\varphi]\!]$. Since $G \subseteq H$, it follows that $[s]_H \subseteq [s]_G \subseteq [\![\varphi]\!]$, then by empty strategy we have $\mathscr{M}, s \vDash \mathsf{Kh}_H\varphi$.

If $[s]_G \in \mathtt{dom}(\sigma_G)$, then A^*_G is non-empty. Since $G \subseteq H$, by Proposition 12 we have $A^*_G \subseteq A^*_H$, so A^*_H is also non-empty. Define partial function $\sigma_H : [S]_H \to A^*_H$ as follows: for each $v \in S$ such that $[v]_G \in \mathtt{dom}(\sigma_G)$, let $\sigma_H([v]_H) = \sigma_G([v]_G)$. We first show that σ_H is well-defined. Suppose $s_1 \sim_H s_2$. Since $G \subseteq H$, we have $s_1 \sim_G s_2$, so $[s_1]_G = [s_2]_G$. Thus $[s_1]_G \in \mathtt{dom}(\sigma_G)$ if and only if $[s_2]_G \in \mathtt{dom}(\sigma_G)$. If $[s_1]_G = [s_2]_G \in \mathtt{dom}(\sigma_G)$, then $\sigma_H([s_1]_H) = \sigma_G([s_1]_G) = \sigma_G([s_2]_G) = \sigma_H([s_2]_H)$. Additionally, for any $[v]_H \in \mathtt{dom}(\sigma_H)$, by definition $[v]_G \in \mathtt{dom}(\sigma_G)$ and $\sigma_G([v]_G)$ is executable on $[v]_G$. Since $\sigma_H([v]_H) = \sigma_G([v]_G)$ and $[v]_H \subseteq [v]_G$, $\sigma_H([v]_H)$ is executable on $[v]_H$. Therefore, σ_H is well-defined.

Let $\delta_H = [s_0]_H[s_1]_H \ldots$ be an arbitrary complete execution of σ_H starting from $[s]_H = [s_0]_H$, with $|\delta_H| > 1$. We show that δ_H is finite and its leaf nodes are contained in $[\![\varphi]\!]$. For $[s_j]_H$ in δ_H such that $[s_j]_H \in \mathtt{dom}(\sigma_H)$, we have $[s_j]_H \xrightarrow{\sigma_H([s_j]_H)} [s_{j+1}]_H$. Thus there exist $s'_j \in [s_j]_H$ and $s'_{j+1} \in [s_{j+1}]_H$ such that $s'_j \xrightarrow{\sigma_H([s_j]_H)} s'_{j+1}$. By the definition of σ_H, $\sigma_H([s_j]_H) = \sigma_G([s_j]_G)$, so $s'_j \xrightarrow{\sigma_G([s_j]_G)} s'_{j+1}$. This implies

$[s'_j]_G \xrightarrow{\sigma_G([s_j]_G)} [s'_{j+1}]_G$. Since $s_j \sim_H s'_j$, $s_{j+1} \sim_H s'_{j+1}$, and $G \subseteq H$, we have $s_j \sim_G s'_j$, $s_{j+1} \sim_G s'_{j+1}$, hence $[s_j]_G = [s'_j]_G$, $[s_{j+1}]_G = [s'_{j+1}]_G$. Therefore, $[s_j]_G \xrightarrow{\sigma_G([s_j]_G)} [s_{j+1}]_G$. Let $\delta_G = [s_0]_G[s_1]_G\ldots$. Then δ_H must be finite; otherwise, δ_G would be an infinite execution of σ_G starting from $[s]_G$. Let $|\delta_G| = |\delta_H| = n+1$. Since δ_H is finite and complete, $[s_n]_H \notin \mathtt{dom}(\sigma_H)$, implying $[s_n]_G \notin \mathtt{dom}(\sigma_G)$. Thus δ_G is a finite complete execution starting from $[s]_G$. As $\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$, we have $[s_n]_G \subseteq [\![\varphi]\!]$, hence $[s_n]_H \subseteq [\![\varphi]\!]$. Therefore, $\mathscr{M}, s \vDash \mathsf{Kh}_H\varphi$. □

Proposition 22 (Axiom AxKhKh) *$\mathscr{M}, s \vDash \mathsf{Kh}_G\mathsf{Kh}_G\varphi$ implies $\mathscr{M}, s \vDash \mathsf{Kh}_G\varphi$.*

PROOF [Sketch] Axiom AxKhKh states that for any pointed model $\mathscr{M}, s$, if there exists a strategy σ_G of G such that all complete executions starting from $[s]_G$ are finite, and for each leaf node $[s']_G$ of these executions, there exists another strategy σ'_G of G such that all complete executions starting from $[s']_G$ are finite and φ is satisfied on their leaf nodes, then there exists a strategy τ_G of G such that all complete executions starting from $[s]_G$ are finite and φ is satisfied on their leaf nodes. To prove the validity, we need to combine σ_G and each σ'_G into τ_G, while resolving three critical issues:

First, the domains of strategies σ'_G may contain irrelevant equivalence classes that are not visited during executions starting from its corresponding leaf node; however, when combined with other strategies, new executions might pass through these equivalence classes. This may result in the new execution failing to achieve φ. Second, overlapping domains of different strategies could result in the combined strategy not being a function. Third, the domain of some strategies might include leaf nodes of other strategies' complete executions. This can cause executions that are originally planned to terminate to continue, even leading to infinite executions.

To address these, we will proceed as follows:

1. Restrict the domain of each strategy to the inner nodes of its complete executions starting from the corresponding leaf node, eliminating irrelevant equivalence classes.
2. Use the Well-ordering Theorem to order the leaf nodes of complete executions of σ_G (this also orders their corresponding strategies σ'_G), then combine these strategies according to the ordering while skipping defined equivalence classes and leaf nodes, constructing τ^γ_G.
3. Finally, combine σ_G and τ^γ_G into τ_G, also skipping defined equivalence classes and leaf nodes.

We present a sketch of the proof (cf. [9] for a similar proof). Suppose that $\mathscr{M}, s \vDash \mathsf{Kh}_G\mathsf{Kh}_G\varphi$, then there exists a strategy σ_G of G such that all complete executions starting from $[s]_G$ are finite and $\mathsf{Kh}_G\varphi$ is satisfied on their leaf nodes. By the Well-ordering Theorem, let $\mathtt{CELeaf}(\sigma_G, s) = \{S_i \mid i < \gamma\}$, where γ is an ordinal number. Pick $s_i \in S_i$, then $[s_i]_G = S_i$. Since $\mathscr{M}, s_i \vDash \mathsf{Kh}_G\varphi$ for each $i < \gamma$, there exists a strategy σ^i_G of G for each $i < \gamma$ such that all complete executions starting from $[s_i]_G$ are finite and φ is satisfied on their leaf nodes. We define τ^γ_G as follows:

Define τ^i_G inductively where $0 \le i < \gamma$:

- If $i = 0$, $\tau^0_G = \sigma^0_G|_{\mathrm{CEInner}(\sigma^0_G, s_0)}$;
- If $i > 0$, $\tau^i_G = f_i \cup (\sigma^i_G|_{D_i})$,

where $f_i = \bigcup_{j<i} \tau^j_G$, $D_i = \mathtt{CEInner}\left(\sigma^i_G, s_i\right) \setminus (\mathtt{dom}(f_i) \cup \{[v]_G \in \mathtt{CELeaf}(f_i, t) \mid [t]_G \in \mathtt{dom}(f_i)\})$. Let $\tau^\gamma_G = \bigcup_{i<\gamma} \tau^i_G$.

We can see that the domain of each τ^i_G is restricted to the inner nodes of its complete executions starting from the corresponding leaf node, and defined equivalence classes and leaf nodes are skipped

(this requires further proof). Moreover, we can verify that τ_G^γ is indeed a partial function. With these results, we can show that on any equivalence class of $\mathtt{dom}(\tau_G^\gamma)$, τ_G^γ is a witness of $\mathsf{Kh}_G\varphi$.

Finally, we define strategy τ_G:

$$\tau_G = \tau_G^\gamma \cup (\sigma_G|_E),$$

where $E = \mathtt{CEInner}(\sigma_G, s) \setminus (\mathtt{dom}(\tau_G^\gamma) \cup \{[v]_G \in \mathtt{CELeaf}(\tau_G^\gamma, t) \mid [t]_G \in \mathtt{dom}(\tau_G^\gamma)\})$. The combination of τ_G^γ and σ_G follows a similar method as before, also skipping defined equivalence classes and leaf nodes. With the above results, we can show that τ_G is the desired strategy.

□

The above propositions give us the soundness of the logic.

Theorem 23 (Soundness) $\vdash \varphi$ *implies* $\vDash \varphi$.

4 Completeness

In this section, we prove the completeness theorem for $\mathbb{SDKH}$. Inspired by [19], we adopt the unraveling technique to handle distributed knowledge-that. On the other hand, when dealing with distributed knowledge-how, we face a similar problem to that with distributed knowledge-that. To find witness strategies for Kh_G-formulas, we make use of atomic group actions and define them as pairs (φ, G), and then transition relations can be defined as $s \xrightarrow{(\varphi,G)} t$ iff $\mathsf{Kh}_G\varphi \in s$ and $\mathsf{K}_G\varphi \in t$, meaning that the witness for $\mathsf{Kh}_G\varphi$ can be found in one step by (φ, G). However, $\mathsf{Kh}_{\{i,j\}}(\varphi \wedge \psi)$ could be satisfied at a state s accidentally by $\langle(\varphi, i), (\psi, j)\rangle$ in the absence of $\mathsf{Kh}_{\{i,j\}}(\varphi \wedge \psi) \in s$, since $\mathsf{K}_{\{i,j\}}(\varphi \wedge \psi) \in t$ always holds whenever $s \xrightarrow{\langle(\varphi,i),(\psi,j)\rangle} t$. To avoid this, we use the unraveling technique again, splitting distributed transitions into transitions of atomic group actions, thereby ensuring that all joint actions are not executable in the canonical model. Moreover, our multi-step setting introduces complications when refuting Kh_G-formulas in the canonical model, where AxKhKh and AxEmpMono will play crucial roles instead of the Cooperation axiom in [19].

Based on the ideas discussed above, we fix a maximal consistent set X_0 and define the canonical model $\mathscr{M}^c(X_0)$ as follows:

Definition 24 (Canonical model) *Given a maximal consistent subset X_0 of* **DKH***, the canonical model* $\mathscr{M}^c(X_0) = \left\langle S^c, \{\sim_i^c\}_{i\in I}, \{A_G^c\}_{G\subseteq I}, \{\xrightarrow{a} \mid a \in A_I^{c+}\}, V^c \right\rangle$ *is defined as follows:*

- *S^c is the set of all* mixed sequences *$X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$ such that*
  - *X_j is a maximal consistent subsets of* **DKH** *for each $j \geq 0$,*
  - *$\varphi_j \in$* **DKH** *for each $j \geq 1$,*
  - *$G_j \neq \emptyset$ and $G_j, H_j \subseteq I$ for each $j \geq 1$,*
  - *$\mathsf{Kh}_{G_j}\varphi_j \in X_{j-1}$ and $\mathsf{K}_{G_j}\varphi_j \in X_j$ for each $j \geq 1$,*
  - *$\{\varphi \mid \mathsf{K}_{H_j}\varphi \in X_{j-1}\} \subseteq X_j$ for each $j \geq 1$;*

  for any state $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$, we use $ed(s)$ to denote X_n;
- *For each $i \in I$, $\sim_i^c$ is defined as follows: for any states $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$ and $s' = X_0\langle(\varphi'_1, G'_1), H'_1\rangle X'_1 \cdots \langle(\varphi'_m, G'_m), H'_m\rangle X'_m$, $s \sim_i^c s'$ iff there exists an integer k such that:*
  - *$0 \leq k \leq \min\{n, m\}$;*
  - *$X_j = X'_j$ and $H_j = H'_j$ for each $1 \leq j \leq k$;*

- $i \in H_j$ *for each* $k < j \leq n$;
- $i \in H'_j$ *for each* $k < j \leq m$;

- *For each nonempty* $G \subseteq I$, $A^c_G = \{(\varphi, G) \mid \varphi \in \mathbf{DKH}\}$ *and* $A^c_\emptyset = \emptyset$;
- *For any* $(\varphi, G) \in A^{c+}_I$ *and states* s,t, $s \xrightarrow{(\varphi,G)} t$ *iff* $t = s\langle(\varphi, G), G'\rangle X$ *where* $G' \subseteq I$ *and* X *is an MCS;*
- *For any* $p \in P$, $V^c(p) = \{s \in S^c \mid p \in ed(s)\}$.

Before explaining the construction of the canonical model, we first verify that $\mathscr{M}^c(X_0)$ is well-defined, i.e., $\sim^c_i$ is an equivalence relation and A^c_G satisfies the requirement in Definition 7.

Proposition 25 *$\sim^c_i$ is an equivalence relation on S^c for any $i \in I$.*

PROOF

Let $i \in I$. $\sim^c_i$ clearly satisfies reflexivity and symmetry.

For any states $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$, $s' = X_0\langle(\varphi'_1, G'_1), H'_1\rangle X'_1 \cdots \langle(\varphi'_m, G'_m), H'_m\rangle X'_m$, and $s'' = X_0\langle(\varphi''_1, G''_1), H''_1\rangle X''_1 \cdots \langle(\varphi''_r, G''_r), H''_r\rangle X''_r$, suppose $s \sim^c_i s'$ and $s' \sim^c_i s''$. Then there exist integers k and k' such that:

- $0 \leq k \leq \min\{n, m\}$, satisfying:
  - $X_j = X'_j$ and $H_j = H'_j$ for $1 \leq j \leq k$;
  - $i \in H_j$ for $k < j \leq n$;
  - $i \in H'_j$ for $k < j \leq m$;
- $0 \leq k' \leq \min\{m, r\}$, satisfying:
  - $X'_j = X''_j$ and $H'_j = H''_j$ for $1 \leq j \leq k'$;
  - $i \in H'_j$ for $k' < j \leq m$;
  - $i \in H''_j$ for $k' < j \leq r$.

Let $k'' = \min\{k, k'\}$. Then for all j:

- If $1 \leq j \leq k''$, then $X_j = X''_j$ and $H_j = H''_j$;
- If $k'' < j \leq n$:
  - Case $k \leq k'$: $k'' = k$, and then $i \in H_j$ for $k'' < j \leq n$;
  - Case $k' < k$: $k'' = k'$, and then:
    - For $k < j \leq n$, $i \in H_j$;
    - For $k' < j \leq k$, since $k \leq m$, we have $H_j = H'_j$ and $i \in H'_j$, and then $i \in H_j$.
- If $k'' < j \leq r$, similarly $i \in H''_j$.

Therefore, $s \sim^c_i s''$. This proves that $\sim^c_i$ is an equivalence relation on S^c. □

By the definition of A^c_G it is obvious:

Proposition 26 *$G \subsetneq H$ implies $A^c_G \cap A^c_H = \emptyset$.*

Now we are ready to explain how we construct the canonical model. Recall that we would like to unravel both the (group) epistemic relations and the (group) action relations in the model. Instead of building a model and then unraveling it, we do it in one go by considering paths of MCSs. Note that between two given MCSs, there can be both the epistemic relations and the action relations, and this is the reason for considering the so-called mixed sequences $X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$, where

H_i represents the epistemic relation and (φ_i, G_i) represents the action transition. Moreover, although we need to unravel the epistemic relations and the action transitions respectively, we do not need to unravel the "cross-type" relations, i.e., it is fine to have transitions of both types between two MCSs. The detailed definition of the mixed sequences resembles the usual conditions of canonical epistemic relation and the canonical relations for the actions in the literature of knowing how logics [9].

Interested readers may wonder why in mixed sequences, we can always require the double (epistemic+action) transitions between the MCSs. Intuitively, there can be two MCSs with only epistemic relations between them, without any action transition (or the other way around). Thanks to the epistemic relation of the empty group and the trivial action $(\top, G)$, such cases are not ruled out, as there are always such trivial epistemic and action relations between any two relevant MCSs: the empty group cannot distinguish any two states and any group always knows how to achieve $\top$. The trivial epistemic relation $\sim_\emptyset$ and the trivial action transition $\xrightarrow{(\top,G)}$ will also play an important role in the existence lemmas for both K_G and Kh_G.

Finally, the definition of $\sim^c_G$ is essentially the same as the canonical epistemic relation in [19], if we ignore the action transitions in the mixed sequences. Intuitively, two sequences are indistinguishable by i if they are two extensions of the same initial history (ignoring the action information) such that i is always in the groups that cannot distinguish the adjacent MCSs from the departing point onward of the two extensions. The action transitions between the mixed sequences are defined similarly as the canonical action relation used in [9].

We now show that all joint actions are not executable in the canonical model, as discussed in the beginning of this section.

Proposition 27 *For any $d \in A^{c*}_I \setminus A^{c+}_I$, d is not executable anywhere in $\mathscr{M}^c$.*

PROOF By Corollary 15, there exist $(\varphi_0, G_0) \in A^c_{G_0}, \cdots, (\varphi_m, G_m) \in A^c_{G_m}$ such that $m > 0$, $\{G_0, \cdots, G_m\}$ is a partial partition of I and $\xrightarrow{d} = \bigcap_{0 \le k \le m} \xrightarrow{(\varphi_k, G_k)}$, then $G_0, \cdots, G_m$ are distinct. Suppose towards a contradiction that there is $s, t \in S^c$ such that $s \xrightarrow{d} t$. Then $s \xrightarrow{(\varphi_k, G_k)} t$ for $0 \le k \le m$. Then $t = s\langle(\varphi_0, G_0), G'_0\rangle Y_0 = \cdots = s\langle(\varphi_m, G_m), G'_m\rangle Y_m$ where $G'_0, \cdots, G'_m \subseteq I$ and $Y_0, \cdots, Y_m$ are MCSs. But $G_0, \cdots, G_m$ are distinct, contradiction. □

For the knowledge-that part in the Truth Lemma, we firstly need the following propositions:

Proposition 28 *For any $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n \in S^c$ and $k \le n$, if $\mathsf{K}_H\varphi \in X_n$ and $H \subseteq H_j$ holds for all $k < j \le n$, then $\mathsf{K}_H\varphi \in X_k$.*

PROOF Suppose there exists $k \le n$ such that $\neg\mathsf{K}_H\varphi \in X_k$. Let m be the maximal such k. Since $\mathsf{K}_H\varphi \in X_n$, we have $m < n$. By Axiom 5, we get $\mathsf{K}_H\neg\mathsf{K}_H\varphi \in X_m$. By Axiom AxKMono, we have $\mathsf{K}_{H_{m+1}}\neg\mathsf{K}_H\varphi \in X_m$. By the definition of states in the canonical model, it follows that $\neg\mathsf{K}_H\varphi \in X_{m+1}$, contradicting the maximality of m. □

Proposition 29 *For any $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n \in S^c$ and $k \le n$, if $\mathsf{K}_H\varphi \in X_k$ and $H \subseteq H_j$ holds for all $k < j \le n$, then $\varphi \in X_n$.*

PROOF We show by induction on $n-k$. If $n-k=0$, then $\mathsf{K}_H\varphi \in X_n$. By Axiom T, we have $\varphi \in X_n$. If $n-k>0$, by Axiom 4 and Axiom AxKMono, we get $\mathsf{K}_{H_{k+1}}\mathsf{K}_H\varphi \in X_k$. By the definition of states in the canonical model, it follows that $\mathsf{K}_H\varphi \in X_{k+1}$. Since $n-(k+1) < n-k$, $k+1 \le n$, $\mathsf{K}_H\varphi \in X_{k+1}$, and

$H \subseteq H_j$ holds for all $k+1 < j \leq n$, by induction hypothesis we have $\varphi \in X_n$. □

With Proposition 28 and 29, we can show the following key lemmas that are essential for the knowledge-that part in the Truth Lemma. Note that we will need to make use of the trivial action transition $\xrightarrow{(\top,G)}$ in the proof of the existence lemma, as mentioned earlier in this section.

Lemma 30 *If $\mathsf{K}_H\varphi \in ed(s)$ and $s \sim^c_H s'$, then $\varphi \in ed(s')$.*

PROOF Note that $\sim^c_H$ is defined by Definition 8 via indistinguishability relations of agents. Let $s = X_0\langle(\varphi_1, G_1), H_1\rangle X_1 \cdots \langle(\varphi_n, G_n), H_n\rangle X_n$ and $s' = X_0\langle(\varphi'_1, G'_1), H'_1\rangle X'_1 \cdots \langle(\varphi'_m, G'_m), H'_m\rangle X'_m$. If $H = \emptyset$, then since $\emptyset \subseteq H_j$ holds for all $0 < j \leq n$, by Proposition 28 we have $\mathsf{K}_\emptyset\varphi \in X_0$. Since $\emptyset \subseteq H'_j$ holds for all $0 < j \leq m$, by Proposition 29 we have $\varphi \in X'_m = ed(s')$.

If $H \neq \emptyset$, since $s \sim^c_H s'$, for each $i \in H$ there exists an integer k_i satisfying:

- $0 \leq k_i \leq \min\{n, m\}$;
- $X_j = X'_j$ and $H_j = H'_j$ for $1 \leq j \leq k_i$;
- $i \in H_j$ for $k_i < j \leq n$;
- $i \in H'_j$ for $k_i < j \leq m$.

Let $k = \max\{k_i\}_{i\in H}$. Then $X_k = X'_k$ and for each $i \in H$, $i \in H_j$ for $k < j \leq n$ and $i \in H'_j$ for $k < j \leq m$. Then $H \subseteq H_j$ holds for all $k < j \leq n$ and $H \subseteq H'_j$ holds for all $k < j \leq m$. Since $\mathsf{K}_H\varphi \in ed(s)$, by Proposition 28, $\mathsf{K}_H\varphi \in X_k = X'_k$. By Proposition 29, $\varphi \in X'_m = ed(s')$. □

Lemma 31 *If $\mathsf{K}_H\varphi \notin ed(s)$, then there exists a state $s' \in S^c$ such that $s \sim^c_H s'$ and $\varphi \notin ed(s')$.*

PROOF Let $Y_0 = \{\psi \mid \mathsf{K}_H\psi \in ed(s)\}$. First, we prove that $Y = Y_0 \cup \{\neg\varphi\}$ is consistent. Suppose not. Then there exist $\psi_1, \ldots, \psi_n \in Y_0$ with $\mathsf{K}_H\psi_1, \ldots, \mathsf{K}_H\psi_n \in ed(s)$ such that

$$\vdash (\psi_1 \wedge \cdots \wedge \psi_n) \to \varphi.$$

By NEC and Axiom DISTK, we get:

$$\vdash (\mathsf{K}_H\psi_1 \wedge \cdots \wedge \mathsf{K}_H\psi_n) \to \mathsf{K}_H\varphi.$$

Since $\mathsf{K}_H\psi_1, \ldots, \mathsf{K}_H\psi_n \in ed(s)$, it follows that $\mathsf{K}_H\varphi \in ed(s)$, contradicting the assumption $\mathsf{K}_H\varphi \notin ed(s)$. Thus Y is consistent. By Lindenbaum's Lemma, there exists a maximal consistent set Y^* such that $Y \subseteq Y^*$. Let $s' = s\langle(\top, H), H\rangle Y^*$. To verify $s' \in S^c$, we only need to check that $\mathsf{K}_H\psi \in ed(s)$ implies $\psi \in Y^*$ and that $\mathsf{Kh}_H\top \in ed(s)$ and $\mathsf{K}_H\top \in Y^*$, which is guaranteed by the construction of Y and the maximality of $ed(s)$ and Y^*. By definition, $s \sim^c_H s'$, making s' the desired state. □

The following lemma is a crucial observation, which is in effect the existence lemma for Kh_G. Note that we will make use of the trivial epistemic relation $\sim_\emptyset$ as mentioned earlier.

Lemma 32 *Let s be a state in S^c, G, H be groups such that $\emptyset \neq H \subseteq G$, (ψ, H) be an action executable on $[s]_G$. If $\mathsf{Kh}_G\varphi \in ed(s')$ for any s' such that $[s]_G \xrightarrow{(\psi,H)} [s']_G$, then $\mathsf{Kh}_G\varphi \in ed(s)$.*

PROOF Let $Y_0 = \{\chi \mid \mathsf{K}_\emptyset\chi \in ed(s)\}$. We first show that $Y = Y_0 \cup \{\mathsf{K}_H\psi, \neg\mathsf{Kh}_G\varphi\}$ is inconsistent. Suppose towards a contradiction that Y is consistent, then by Lindenbaum's Lemma, there is an MCS Y^* such that $Y \subseteq Y^*$. Let $s' = s\langle(\psi,H),\emptyset\rangle Y^*$. By the definition of Y_0, for any $\mathsf{K}_\emptyset\chi \in ed(s)$, we have $\chi \in ed(s')$. To verify $s' \in S^c$, we still need to show that $\mathsf{Kh}_H\psi \in ed(s)$ and $\mathsf{K}_H\psi \in Y^*$. Since $\mathsf{K}_H\psi \in Y \subseteq Y^*$, we only need to show that $\mathsf{Kh}_H\psi \in ed(s)$. Since (ψ,H) is executable on $[s]_G$, there exists $t \in S^c$ such that $s \xrightarrow{(\psi,H)} t$, then $t = s\langle(\psi,H),H'\rangle X$ where $H' \subseteq I$ and X is an MCS. Then $\mathsf{Kh}_H\psi \in ed(s)$ and $\mathsf{K}_H\psi \in X$. Therefore $s' \in S^c$. It follows that $s \xrightarrow{(\psi,H)} s'$, then $[s]_G \xrightarrow{(\psi,H)} [s']_G$, which is contradictory with the assumption that $\mathsf{Kh}_G\varphi \in ed(s')$ for any s' such that $[s]_G \xrightarrow{(\psi,H)} [s']_G$. Therefore $Y = Y_0 \cup \{\mathsf{K}_H\psi, \neg\mathsf{Kh}_G\varphi\}$ is inconsistent. Then there are $\chi_1,\cdots,\chi_n \in Y_0$ such that $\mathsf{Kh}_\emptyset\chi_j \in ed(s)$ for each $1 \leq j \leq n$ and $\{\chi_1,\ldots,\chi_n,\mathsf{K}_H\psi,\neg\mathsf{Kh}_G\varphi\}$ is inconsistent. Then

$$\vdash (\chi_1 \wedge \cdots \wedge \chi_n) \to (\mathsf{K}_H\psi \to \mathsf{Kh}_G\varphi).$$

By NECK and Axiom DISTK,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to \mathsf{K}_\emptyset(\mathsf{K}_H\psi \to \mathsf{Kh}_G\varphi).$$

By Axiom AxEmpMono,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to \mathsf{K}_\emptyset(\mathsf{Kh}_H\mathsf{K}_H\psi \to \mathsf{Kh}_H\mathsf{Kh}_G\varphi).$$

By Axiom T,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_H\mathsf{K}_H\psi \to \mathsf{Kh}_H\mathsf{Kh}_G\varphi).$$

By Axiom AxKhtoKhK,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_H\psi \to \mathsf{Kh}_H\mathsf{Kh}_G\varphi).$$

By Axiom AxKhMono and Axiom AxKhKh,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_H\psi \to \mathsf{Kh}_G\varphi).$$

Since $\mathsf{K}_\emptyset\chi_1,\ldots,\mathsf{K}_\emptyset\chi_n,\mathsf{Kh}_H\psi \in ed(s)$, we have $\mathsf{Kh}_G\varphi \in ed(s)$. □

We are now ready to prove the Truth Lemma.

Lemma 33 (Truth Lemma) *For any* $\varphi \in$ **DKH** *and* $s \in S^c$, $\mathscr{M}^c(X_0), s \vDash \varphi$ *iff* $\varphi \in ed(s)$.

PROOF We show by induction on φ. We only show the case of $\mathsf{Kh}_G\varphi$, the other cases are straightforward.

Right to Left: Suppose that $\mathsf{Kh}_G\varphi \in ed(s)$, we show that $\mathscr{M}^c, s \vDash \mathsf{Kh}_G\varphi$. There are two cases: $\mathsf{K}_G\varphi \in ed(s)$ or $\mathsf{K}_G\varphi \notin ed(s)$. If $\mathsf{K}_G\varphi \in ed(s)$, then by Lemma 30, $\varphi \in ed(s')$ for each $s' \in [s]_G$. By IH, $\mathscr{M}^c, s' \vDash \varphi$ for each $s' \in [s]_G$. Therefore, $\mathscr{M}^c, s \vDash \mathsf{K}_G\varphi$. By Axiom AxKtoKh and soundness, we have $\mathscr{M}^c, s \vDash \mathsf{Kh}_G\varphi$.

If $\mathsf{K}_G\varphi \notin ed(s)$, then $G \neq \emptyset$, otherwise $\mathsf{K}_\emptyset\varphi \in ed(s)$ by Axiom AxEmpKhtoK. Consider partial function $\sigma_G = \{[s]_G \mapsto (\varphi, G)\}$. We need to show that (φ, G) is executable on $[s]_G$. Let v be a state in $[s]_G$. Since $\mathsf{Kh}_G\varphi \in ed(s)$, then $\mathsf{K}_G\mathsf{Kh}_G\varphi \in ed(s)$ by Axiom AxKhtoKKh. By Lemma 30 and Lemma 31, $\mathsf{K}_G\mathsf{Kh}_G\varphi \in ed(v)$. By Axiom T, $\mathsf{Kh}_G\varphi \in ed(v)$. Let $Y_0 = \{\chi \mid \mathsf{K}_\emptyset\chi \in ed(v)\}$. We need to show that

$Y = Y_0 \cup \{\mathsf{K}_G\varphi\}$ is consistent. Suppose not, then there are $\chi_1, \ldots, \chi_n \in Y_0$ such that $\mathsf{Kh}_\emptyset\chi_j \in ed(v)$ for each $1 \leq j \leq n$ and $\{\chi_1, \ldots, \chi_n, \mathsf{K}_G\varphi\}$ is inconsistent. Then

$$\vdash (\chi_1 \wedge \cdots \wedge \chi_n) \to (\mathsf{K}_G\varphi \to \bot).$$

By NECK and Axiom DISTK,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to \mathsf{K}_\emptyset(\mathsf{K}_G\varphi \to \bot).$$

By Axiom AxEmpMono,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to \mathsf{K}_\emptyset(\mathsf{Kh}_G\mathsf{K}_G\varphi \to \mathsf{Kh}_G\bot).$$

By Axiom T,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_G\mathsf{K}_G\varphi \to \mathsf{Kh}_G\bot).$$

By Axiom AxKhtoKhK,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_G\varphi \to \mathsf{Kh}_G\bot).$$

By Axiom AxKhbot,

$$\vdash (\mathsf{K}_\emptyset\chi_1 \wedge \cdots \wedge \mathsf{K}_\emptyset\chi_n) \to (\mathsf{Kh}_G\varphi \to \bot).$$

Since $\mathsf{K}_\emptyset\chi_1, \ldots, \mathsf{K}_\emptyset\chi_n, \mathsf{Kh}_G\varphi \in ed(v)$, we have $\bot \in ed(v)$, which is in contradiction with the consistency of $ed(v)$. Therefore, Y is consistent.

By Lindenbaum's Lemma, there is an MCS Y^* such that $Y \subseteq Y^*$. Let $s' = v\langle(\varphi, G), \emptyset\rangle Y^*$. By the definition of Y, $\mathsf{Kh}_G\varphi \in ed(v)$, $\mathsf{K}_G\varphi \in ed(s')$ and $\chi \in ed(s')$ for any $\mathsf{K}_\emptyset\chi \in ed(v)$, therefore $s' \in S^c$. Then $v \xrightarrow{(\varphi,G)} s'$. Therefore (φ, G) is executable on $[s]_G$. Let $[s'']_G$ be an equivalence class such that $[s]_G \xrightarrow{(\varphi,G)} [s'']_G$, then there are $t' \in [s]_G$ and $t'' \in [s'']_G$ such that $t' \xrightarrow{(\varphi,G)} t''$. It follows that $t'' = t'\langle(\varphi, G), H\rangle X$ where $H \subseteq I$ and X is an MCS. Then $\mathsf{Kh}_G\varphi \in ed(t')$ and $\mathsf{K}_G\varphi \in ed(t'')$. By Lemma 30 and Lemma 31, $\mathsf{K}_G\varphi, \varphi \in ed(t)$ for each $t \in [s'']_G$. By IH, $\mathscr{M}^c(X_0), t \vDash \varphi$. Moreover, since $\neg\mathsf{K}_G\varphi \in ed(s)$, it is not the case that $[s]_G \xrightarrow{(\varphi,G)} [s]_G$, then all complete executions of $\sigma_G = \{[s]_G \mapsto (\varphi, G)\}$ starting from $[s]_G$ are finite. Therefore, $\mathscr{M}^c(X_0), s \vDash \mathsf{Kh}_G\varphi$.

Left to Right: Suppose that $\mathscr{M}^c(X_0), s \vDash \mathsf{Kh}_G\varphi$, we show that $\mathsf{Kh}_G\varphi \in ed(s)$. By semantics, there is a strategy σ_G such that

1. $[t]_G \subseteq [\![\varphi]\!]$ for all $[t]_G \in \mathtt{CELeaf}(\sigma_G, s)$, and
2. all its complete executions starting from $[s]_G$ are finite.

For each $t' \in [t]_G$, by IH, $\varphi \in ed(t')$. By Lemma 31, $\mathsf{K}_G\varphi \in ed(t)$. By Axiom AxKtoKh, $\mathsf{Kh}_G\varphi \in ed(t)$.

If $[s]_G \notin \mathtt{dom}(\sigma_G)$, then $[s]_G \in \mathtt{CELeaf}(\sigma_G, s)$, then $\mathsf{Kh}_G\varphi \in ed(s)$. If $[s]_G \in \mathtt{dom}(\sigma_G)$, then $G \neq \emptyset$ and $[s]_G \in \mathtt{CEInner}(\sigma_G, s)$. In order to show that $\mathsf{Kh}_G\varphi \in ed(s)$, we will show a stronger result, that is, $\mathsf{Kh}_G\varphi \in ed(s')$ for all $[s']_G \in \mathtt{CEInner}(\sigma_G, s)$. We firstly show the following claim:

Claim 33.1 *If there exists $[s']_G \in \mathtt{CEInner}(\sigma_G, s)$ such that $\neg\mathsf{Kh}_G\varphi \in ed(s')$, then there exists an infinite execution of σ_G starting from $[s]_G$.*

Proof of Claim 33.1: Suppose that there exists $[s']_G \in \texttt{CEInner}(\sigma_G, s)$ such that $\neg \mathsf{Kh}_G\varphi \in ed(s')$. Let $X = \{[v]_G \in \texttt{CEInner}(\sigma_G, s) \mid \neg \mathsf{Kh}_G\varphi \in ed(v)\}$, then $[s']_G \in X$, and for any $[v]_G \in X$, since $[v]_G \in \texttt{CEInner}(\sigma_G, s)$, we have $[v]_G \in \texttt{dom}(\sigma_G)$, then $\sigma_G([v]_G)$ is executable on $[v]_G$. By Proposition 27, $\sigma_G([v]_G) \in A_G^{c+}$, so it has the form of (ψ, H) where $\emptyset \neq H \subseteq G$.

Define binary relation on $[S^c]_G$ as $R = \{([v]_G, [v']_G) \mid [v]_G \xrightarrow{\sigma_G([v]_G)} [v']_G\}$. For any $[v]_G \in X$, since $\neg \mathsf{Kh}_G\varphi \in ed(v)$, by Lemma 32, there exists $v' \in S^c$ such that $[v]_G \xrightarrow{\sigma_G([v]_G)} [v']_G$ and $\neg \mathsf{Kh}_G\varphi \in ed(v')$. Since $[v]_G \in X$, then $[v]_G \in \texttt{CEInner}(\sigma_G, s)$. Since $\mathsf{Kh}_G\varphi \in ed(t)$ for all $[t]_G \in \texttt{CELeaf}(\sigma_G, s)$, then $[v']_G \notin \texttt{CELeaf}(\sigma_G, s)$. Since $[v]_G \xrightarrow{\sigma_G([v]_G)} [v']_G$, then $[v']_G \in \texttt{CEInner}(\sigma_G, s)$, then $[v']_G \in X$, then $([v]_G, [v']_G) \in R$. Therefore, R is an entire binary relation on X, that is, for each $[v]_G \in X$, there is $[v']_G \in X$ such that $([v]_G, [v']_G) \in R$. By Axiom of Dependent Choice, there exists an infinite sequence $[v_0]_G[v_1]_G \cdots$ such that $([v_n]_G, [v_{n+1}]_G) \in R$ for all $n \in \mathbb{N}$.

By definition of R, $[v_0]_G[v_1]_G \cdots$ is a complete execution of σ_G starting from $[v_0]$. Since all complete execution of σ_G starting from $[s]_G$ are finite and $[v_0]_G \in \texttt{CEInner}(\sigma_G, s)$, there exists an execution of σ_G $[s_0]_G \cdots [s_j]_G$ $(j \in \mathbb{N})$ such that $[s_0]_G = [s]_G$ and $[s_j]_G = [v_0]_G$. Then $[s_0]_G \cdots [s_j]_G[v_1]_G \cdots$ is an infinite execution of σ_G starting from $[s]_G$. ■

For any $[s']_G \in \texttt{CEInner}(\sigma_G, s)$, suppose that $\neg \mathsf{Kh}_G\varphi \in ed(s')$. By Claim 33.1, there exists an infinite execution of σ_G starting from $[s]_G$, which is in contradiction with all complete execution of σ_G starting from $[s]_G$ are finite. Therefore, $\mathsf{Kh}_G\varphi \in ed(s')$ for all $[s']_G \in \texttt{CEInner}(\sigma_G, s)$. Since $[s]_G \in \texttt{CEInner}(\sigma_G, s)$, we have $\mathsf{Kh}_G\varphi \in ed(s)$. □

Theorem 34 $\mathbb{SDKH}$ *is strongly complete.*

5 Conclusions and future work

In this paper, we propose a framework of group knowledge-how, featuring the distributed actions of groups that can come from three sources. The framework can be viewed as a generalization of both the planning-based approach and the coalition-based approach to the logic of knowing how. We give a sound and complete proof system of the logic with intuitive axioms. We leave the discussions on the decidability and model-theoretical properties to a future occasion.

There are many further directions to explore based on this very general framework. One interesting direction is to consider subclasses of the models and study their corresponding logics. For example, one may want to consider the case when the atomic group actions are all empty except the singleton ones, which will bring our framework closer to the setting of coalition logic and (epistemic) ATL such as [18, 19, 11, 12]. Conceptually, this also makes sense if one really wants the distributed knowledge-how to be always decomposable into individual know-how. It will also help us to better understand the differences and connections with other related existing frameworks.

One may notice that in our current setting, a joint action consisting of multiple subgroups doing the same action is equivalent to a single subgroup doing it, [7] e.g., the transition relation of $\langle a, a\rangle$ is the intersection of $\xrightarrow{a}$ and itself, which is exactly $\xrightarrow{a}$. This is different from the ATL-like frameworks, where the effect of the joint action $\langle a, a\rangle$ can be different from a [13, 5]. This invites us to extend our work in a non-trivial way to handle resource-sensitive settings where the number of individuals performing the same action matters. We may also add more structure to groups or consider intensional groups as

[7] It is similar in the setting of STIT logic, where "actions" are formalized as possible outcomes [4].

discussed in [3]. Also, the members of the groups may not be equal in their roles as in [6]. In particular, an interesting direction to explore is the possibility that each group has a leader who can plan the actions of the members. The commitments and communication among the leaders may help the groups synchronize better.

Given our discussions on extra group actions which cannot be reduced to the individual actions, it may be interesting to find out whether this can be applied to distributed knowledge-that as well: the epistemic relation $\sim_G$ is not defined by the intersection of the individual ones but just a subset of it. More precisely, in addition to distributed knowledge where $\sim_G = \bigcap_{i\in G} \sim_i$, through brainstorming, a group can be smarter than the mere union of individuals, so $\bigcap_{i\in G} \sim_i \subseteq \sim_G$ does not necessarily hold. However, we think the logic stays the same.

Finally, inspired by various versions of distributed knowledge-how, e.g., [10, 2], we may also consider other definitions of distributed knowledge-how that incorporate dynamics as well.

Acknowledgment The authors thank Yanjun Li and Pavel Naumov for their comments and discussions about earlier versions of the paper and related topics. The authors are also grateful to the anonymous reviewers, whose comments improved the presentation of the paper.

References

[1] Thomas Ågotnes & Yì N. Wáng (2017): *Resolving distributed knowledge*. *Artif. Intell.* 252, pp. 1–21, doi:10.1016/j.artint.2017.07.002.

[2] Philippe Balbiani & Hans van Ditmarsch (2024): *Towards Dynamic Distributed Knowledge*. In Agata Ciabattoni, David Gabelaia & Igor Sedlár, editors: *Proceedings of AiML 2024*, College Publications, pp. 125–146. Available at `https://hal.science/hal-04712434`.

[3] Marta Bílková & Igor Sedlár (2023): *Epistemic Logics of Structured Intensional Groups*. In Rineke Verbrugge, editor: *Proceedings of TARK2023*, *EPTCS* 379, pp. 113–130, doi:10.4204/EPTCS.379.11.

[4] Jan Broersen & Andreas Herzig (2015): *Using STIT Theory to Talk About Strategies*, pp. 137–173. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-662-48540-8_5.

[5] Nils Bulling, Valentin Goranko & Wojciech Jamroga (2015): *Logics for Reasoning About Strategic Abilities in Multi-player Games*, pp. 93–136. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-662-48540-8_4.

[6] Barbara Dunin-Keplicz & Rineke Verbrugge (2004): *A Tuning Machine for Cooperative Problem Solving*. *Fundam. Inf.* 63(2–3), p. 283–307. Available at `https://www.researchgate.net/publication/220444551_A_Tuning_Machine_for_Cooperative_Problem_Solving`.

[7] R Fagin, J Halpern, Y Moses & M Vardi (1995): *Reasoning about knowledge*. MIT Press, Cambridge, MA, USA, doi:10.7551/mitpress/5803.001.0001.

[8] Jie Fan, Davide Grossi, Barteld Kooi, Xingchi Su & Rineke Verbrugge (2021): *Commonly Knowing Whether*. arXiv:2001.03945.

[9] Raul Fervari, Andreas Herzig, Yanjun Li & Yanjing Wang (2017): *Strategically Knowing How*. In Carles Sierra, editor: *Proceedings of IJCAI 2017*, pp. 1031–1038, doi:10.24963/ijcai.2017/143.

[10] Rustam Galimullin & Louwe B. Kuijer (2024): *Varieties of Distributed Knowledge*. In: *Proceedings of AiML 2024*, College Publications, pp. 379–400. Available at `https://rgalimullin.gitlab.io/AIML24/DK.pdf`.

[11] Wiebe van der Hoek & Michael Wooldridge (2003): *Cooperation, Knowledge, and Time: Alternating-Time Temporal Epistemic Logic and its Applications*. *Studia Logica* 75(1), pp. 125–157, doi:10.1023/a:1026185103185.

[12] Wojciech Jamroga & Thomas Ågotnes (2007): *Constructive knowledge: what agents can achieve under imperfect information*. *J. Appl. Non Class. Logics* 17(4), pp. 423–475, doi:10.3166/JANCL.17.423-475.

[13] Wojciech Jamroga & Wiebe van der Hoek (2004): *Agents that Know How to Play*. *Fundamenta informaticae* 63(2-3), pp. 185–219. Available at `https://www.researchgate.net/publication/220444694`.

[14] Yanjun Li & Yanjing Wang (2019): *Multi-agent Knowing How via Multi-step Plans: A Dynamic Epistemic Planning Based Approach*. In: *Proceedings of LORI2019*, *Lecture Notes in Computer Science* 11813, Springer, pp. 126–139, doi:10.1007/978-3-662-60292-8_10.

[15] Yanjun Li & Yanjing Wang (2021): *Planning-based knowing how: A unified approach*. *Artif. Intell.* 296, p. 103487, doi:10.1016/j.artint.2021.103487.

[16] Yanjun Li & Yanjing Wang (2024): *Knowing how to plan about planning: Higher-order and meta-level epistemic planning*. *Artif. Intell.* 337, p. 104233, doi:10.1016/j.artint.2024.104233.

[17] Yanjun Li, Quan Yu & Yanjing Wang (2017): *More for free: a dynamic epistemic framework for conformant planning over transition systems*. *Journal of Logic and Computation* 27(8), pp. 2383–2410, doi:10.1093/logcom/exx020.

[18] Pavel Naumov & Jia Tao (2017): *Together We Know How to Achieve: An Epistemic Logic of Know-How (Extended Abstract)*. In Jérôme Lang, editor: *Proceedings of TARK 2017*, 251, pp. 441–453, doi:10.4204/EPTCS.251.32.

[19] 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.

[20] Rasmus Rendsvig, John Symons & Yanjing Wang (2024): *Epistemic Logic*. In Edward N. Zalta & Uri Nodelman, editors: *The Stanford Encyclopedia of Philosophy*, Summer 2024 edition, Metaphysics Research Lab, Stanford University. Available at `https://plato.stanford.edu/archives/sum2024/entries/logic-epistemic/`.

[21] Xingchi Su (2017): *Distributed Knowing Whether*. In Alexandru Baltag, Jeremy Seligman & Tomoyuki Yamada, editors: *Proceedings of LORI 2017*, pp. 653–657, doi:10.1007/978-3-662-55665-8_47.

[22] Yanjing Wang (2015): *A Logic of Knowing How*. In: *Proceedings of LORI-V*, pp. 392–405, doi:10.1007/978-3-662-48561-3_32.

[23] Yanjing Wang (2018): *Beyond knowing that: a new generation of epistemic logics*. In: *Jaakko Hintikka on knowledge and game theoretical semantics*, Springer, Cham, pp. 499–533, doi:10.1007/978-3-319-62864-6_21.

[24] Yanjing Wang (2018): *A logic of goal-directed knowing how*. *Synthese* 195(10), pp. 4419–4439, doi:10.1007/s11229-016-1272-0.

Maximum Entropy and Bayesian Conditioning Under Extended Space

Boning Yu
The Department of Philosophy
University of Maryland-College Park, United States
yub@umd.edu

This paper examines the conditions under which Bayesian conditioning aligns with Maximum Entropy. Specifically, I address cases in which newly learned information does not correspond to an event in the probability space defined on the sample space of outcomes. To facilitate Bayesian conditioning in such cases, one must therefore extend the probability space so that the new information becomes an event in this expanded space. Skyrms (1985) argues that Bayesian conditioning in an extended probability space on a product space of outcomes aligns precisely with the solution from Maximum Entropy. In contrast, Seidenfeld (1986) uses Friedman and Shimony's (1971) result to criticize Skyrms' approach as trivial, suggesting that alignment holds only under a degenerate probability model. Here, I argue that Friedman and Shimony's result must either (1) be a benign consequence of Skyrms' approach, or (2) pose a universal challenge to any method of extending spaces. Accepting (2) would imply that Bayesian conditioning is incapable of accommodating information beyond the probability space defined on the original outcome space.
Key Words: maximum entropy, Bayesian conditioning, law of total expectation

1 Introduction

Bayesian conditioning requires a well-defined probability space. However, it is not always clear what constitutes an appropriate probability space to incorporate newly learned information. For instance, consider a six-sided die. If one initially believes that the die is fair and subsequently learns that the number rolled is odd, they can straightforwardly condition on this event within a probability space defined on a sample space of the six possible outcomes. In contrast, if they instead learn that the die is biased in such a way that the expectation of a roll is 5, Bayesian conditioning cannot take place within the same probability space defined on the six possible outcomes. An immediate solution is to extend the probability space, thereby ensuring that the new information corresponds to an event within this extended framework. Nevertheless, given that there are infinitely many ways of extending the probability space, it remains unclear whether one method of extension ought to be preferred over another.

Apart from Bayesian conditioning, another mainstream approach to updating probabilistic beliefs is Maximum (Shannon's) Entropy[1] or Minimizing Relative Entropy.[2] Shannon's Entropy measures uncertainty within a probability distribution. Maximizing this entropy therefore entails adopting the least committed (or, most conservative) stance consistent with all known information, thus imposing no un-

[1] Shannon's Entropy is defined as $H(X) = -\sum_{x\in\chi} p(x)\log p(x)$, where X is a random variable such that $X : \chi \to R$, with a probability distribution $p : \chi \to [0,1]$.

[2] Minimizing Relative Entropy is grounded in relative entropy (Kullback-Leibler divergence), a measure of divergence between two probability distributions.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 98–111, doi:10.4204/EPTCS.437.12

© Boning Yu
This work is licensed under the
Creative Commons Attribution License.

warranted assumptions.[3] Unlike Bayesian conditioning, Maximum Entropy does not require that the learned information corresponds to an event in a certain probability space.

It is unsurprising that identifying the conditions under which the results produced by Bayesian conditioning coincide with those produced by Maximum Entropy have attracted considerable attention among both mathematicians and philosophers. Philosophers such as Skyrms (1985) and Seidenfeld (1986) have notably contributed to the discussion by examining and interpreting the interplay between the two updating methods under varying conditions. Particularly intriguing scenarios arise when newly learned information does not correspond to an event in the original outcome space. Because there is no standard way to extend the original probability space to handle such information, questions about how to do so have become especially important in philosophical debates. Drawing on Martin-Löf's (1974) results, Skyrms (1985) proposes that Bayesian conditioning on an extended probability space that is defined on a product space of the set of potential outcomes yields results consistent with those produced by Maximum Entropy.[4] Seidenfeld, despite being largely in agreement with Skyrms in other ways, challenges Skyrms' space-extending approach. Relying on Friedman and Shimony's (1971) results, Seidenfeld demonstrates that Skyrms' approach works only under a degenerate probability model and argues that Maximum Entropy should therefore not be considered an extension of Bayesian conditioning.

In this paper, I investigate this disagreement between Skyrms and Seidenfeld. Specifically, my analysis focuses on situations in which an agent (1) has no prior information about outcomes, and thus initially adopts a uniform prior,[5] and (2) learns information that does not correspond to an event in the original outcome space.[6] Under these conditions, Skyrms' space-extending approach appears effective in a way that warrants careful reconsideration of Seidenfeld's criticism. To that end, I evaluate two potential interpretations of Friedman and Shimony's (1971) findings. The first interpretation accepts their results at face value but fails to adequately reflect the Skyrms' procedure in space extension. The second interpretation, though mathematically intricate, accurately reflects a key aspect of Skyrms' approach. However, under this second interpretation, a new difficulty arises: If Friedman and Shimony's result poses a problem for Maximum Entropy, then the problems necessarily generalizes to all space-extending approaches. Ultimately, I argue that either (1) Friedman and Shimony's result does not constitute any real challenge for Maximum Entropy, or (2) this challenge, insofar as it exists, is a universal one affecting all methods of space extension. Accepting (2) would imply that Bayesian conditioning is incapable of accommodating information beyond the original outcome space.

The remainder of this paper proceeds as follows. In Section 2, I demonstrate the mechanics and advantages of Skyrms' space-extending approach. Then, in Section 3, I introduce Friedman and Shimony's results and Seidenfeld's argument against Skyrms. In Section 4, I evaluate two interpretations of Friedman and Shimony's findings. In Section 5, I generalize Friedman and Shimony's reasoning to broader

[3]Accordingly, Minimizing Relative Entropy involves selecting a posterior distribution that minimally diverges from one's prior, which reflects a conservative approach that avoids unnecessary departures from previous beliefs. Notably, when the prior distribution is uniform across outcomes, Minimizing Relative Entropy yields the same results as Maximum Entropy. For simplicity and closer alignment with the literature, this paper exclusively discusses the case of Maximum Entropy.

[4]This general fact is known as the Gibbs conditioning principle (Csiszár, 1984; Dembo & Zeitouni, 1996) in statistics.

[5]Admittedly, condition (1) restricts the scope of my analysis. Seidenfeld emphasizes that the problem identified by Friedman and Shimony also arises in contexts involving non-uniform priors, which is where Minimizing Relative Entropy comes into play. However, my defense of Skyrms' approach does not rely on assuming a uniform prior. So, it does no harm to follow Skyrms' initial framing of the issue here.

[6]Strictly speaking, an event is said to be in a probability space rather than in a sample space of outcomes. However, for the purpose of this paper, I assume each probability space's sigma-algebra is "full" in the sense that the sigma-algebra equals the power set of the corresponding sample space. So, any subset of a sample space is measurable, and for ease of illustration, it does no harm to say "an event in a sample space" when the event is a subset of the sample space.

context of space-extension. Finally, in Section 6, I conclude that Friedman and Shimony's results either are not a problem for Maximum Entropy or pose a problem for all space-extending approaches.

2 Skyrms' Space-Extending Approach

I begin with an overview of Skyrms' space-extending approach in order to establish a clear context for evaluating Friedman and Shimony's theorem (hereafter, the F-S theorem). Bayesian conditioning requires extending the probability space when newly learned information does not correspond to an event in the space of outcomes. According to Skyrms (1985), Bayesian conditioning in a product probability space that consists of sequences of random variables (e.g., mapping from outcomes to real values) produces results that align with those produced by Maximum Entropy. Below is an example.

Consider a six-sided die whose faces are numbered 1 through 6. The set $\Omega_0 = \{1,2,3,4,5,6\}$ serves as the outcome space and represents the number that lands face-up after a roll. A random variable X maps each outcome of a roll to an integer from 1 to 6, such that $X : \Omega_0 \to \{1,2,3,4,5,6\}$. Suppose that the agent initially possesses no additional knowledge about the die and therefore adopts a uniform prior, assigning probability $\frac{1}{6}$ to each face, yielding a prior expectation of 3.5. If, later, the agent learns that the expected outcome of this die is 5 (i.e., $E(X) = 5$), this information cannot be incorporated through Bayesian conditioning in the prior probability space $(\Omega_0, \mathscr{F}_0, P_0)$,[7] as $E(X) = 5$ does not correspond to any event in that space.

Skyrms' approach suggests an extension to a space of N-length sequences $\{x_i\}_{i=1}^N$ that record outcomes of N independent rolls. The extension creates a product space $\Omega_E = \Omega_0^N$, which contains 6^N distinct sequences (Figure 1). In the absence of prior knowledge about the die's bias, the agent adopts a uniform measure on Ω_E. The corresponding probability space $(\Omega_E, \mathscr{F}_E, P_E)$ also implies that the expectation of X is 3.5. Intuitively, as the length of sequences grows, the empirical frequencies in a sequence increasingly accurately reflect the die's true bias (or lack thereof) toward each face. In a space composed of infinitely long sequences, Bayesian conditioning rules out the sequences whose long-run averages differ from 5, thereby retaining only those compatible with the learned information. The surviving sequences form an updated space $(\Omega_c, \mathscr{F}_c, \mu_c)$, where $\Omega_C \subset \Omega_E$ and, by Bayes' rule, μ_c remains uniform after renormalization. Projecting this updated measure back onto the original outcome space Ω_0 yields a new distribution, in which each face's relative frequency in $(\Omega_c, \mathscr{F}_c, \mu_c)$ determines the posterior probability distribution on Ω_0.

Formally, in this extended space, when N is large, Bayesian conditioning on $E(X) = 5$ can be approximated by conditioning on the event that the sample mean of the sequence is close to 5, i.e.,

$$A_{N,\varepsilon} = \{x \in \Omega_E : |\bar{x} - 5| < \varepsilon\}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$.[8] Provided that the true expectation is 5, the law of large numbers implies that, as N grows, sequences whose empirical average deviates from 5 become increasingly rare, while large deviation theory (Sanov, 1957/1958) indicates that the measure of $A_{N,\varepsilon}^c$ decays exponentially fast in N. Furthermore, as N approaches infinity, for an arbitrarily small ε (approacheing 0), sequences will almost surely lie inside $A_{N,\varepsilon}$; therefore, in $(\Omega_E, \mathscr{F}_E, \mu_E)$,[9] conditioning on $A_{N,\varepsilon}$ is equivalent to conditioning on the set of infinite sequences whose limiting average is 5.

[7] Here I adopt the standard notation for a probability space, where $\mathscr{F}_i$ is the sigma algebra, and P_i is the probability measure.

[8] Grove and Halpern (1997, pp. 211-212) discuss conditioning on events of zero measure with respect to a different approach to extending space.

[9] As Ω_E is countably infinite, a uniform measure on the sample space is not a probability measure. So, I adopt μ_E to denote a uniform measure on a countably infinite sample space.

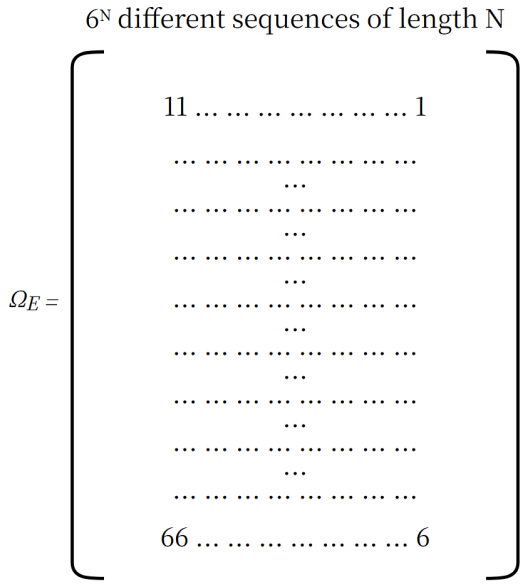


Figure 1: The constitution of the extended sample space Ω_E. The space consists of 6^N different sequences of outcomes, and each sequence has length N.

Crucially, as N approaches infinity, this Bayesian procedure yields a limiting distribution on Ω_0 that exactly matches the solution from Maximum Entropy (van Campenhout & Cover, 1981; Csiszár, 1984; Skyrms, 1985; Csiszár & Shields, 2004). In other words, conditioning on $E(X) = 5$ in $(\Omega_E, \mathscr{F}_E, \mu_E)$ yields an updated probability distribution,

$$\{0.02053, 0.03854, 0.07232, 0.13574, 0.25475, 0.47812\}$$

on the outcomes 1 through 6, respectively. This is the distribution that maximizes Shannon's Entropy, or minimizes relative entropy with respect to the uniform prior, subject to the constraint $E(X) = 5$. A conceptual motivation for this approach will be discussed in Section 4.

Another concrete application of the space-extending approach appears in Vasudevan's (2020) justification for the Maximum Entropy solution in the context of the Judy Benjamin Problem (van Fraassen, 1981). In that scenario, Vasudevan addresses information expressed as a ratio between two outcomes. The results obtained via the space-extending approach effectively align with the solution from Maximum Entropy in both cases.

3 The Problem Posed by the Friedman and Shimony's Theorem

Friedman and Shimony highlight a problematic consequence of aligning the results of Bayesian conditioning with results obtained from Maximum Entropy, particularly when newly learned information does not correspond to an event within the original outcome space.[10] Instead of discussing any one specific extended probability space, the F-S theorem focuses on the alignment between Bayesian conditioning

[10]Under the assumption that a probability space's sigma-algebra is "full," I use the phrase "an event within the original outcome (sample) space" to mean "an event in the probability space defined on the original outcome space." This simplification is valid because every subset of the sample space is measurable."

and Maximum Entropy. Consequently, Friedman and Shimony (1971) - as well as Seidenfeld (1986) - interpret the F-S theorem as a challenge to the claim that Maximum Entropy may be justified as an extension of Bayesian conditioning.

In order to clearly illustrate Friedman and Shimony's result, I will continue the above-mentioned example involving a six-sided die.[11] Initially, the agent holds no prior knowledge of the die beyond the fact that there are six potential outcomes, and thus assigns a uniform prior distribution over those six outcomes. Formally, the prior probability for each outcome is $P(X = x_i) = \frac{1}{6}, x_i = i$, and $i = 1, 2, \ldots, 6$. Based on these prior probabilities, then, the agent's expected value of a single roll is initially given by $E_0(X) = \sum_{i=1}^{6} \frac{x_i}{6} = 3.5$.[12] Now, let $\hat{d_\varepsilon}$ represent the evidence that the posterior expected value of X is ε, formally $\sum_{i=1}^{6} x_i P\left(x_i \mid \hat{d_\varepsilon}\right) = \varepsilon$. If the results of Bayesian conditioning align with those from Maximum Entropy, then the posterior probabilities must take the form:

$$P\left(x_i \mid \hat{d_\varepsilon}\right) = Z^{-1} e^{-\beta x_i} \tag{1}$$

where $Z = \Sigma_{i=1}^{6} e^{-\beta x_i}$, and β decreases monotonically with ε.[13] Given that each value of ε corresponds to a unique value of $\beta, \hat{d_\varepsilon}$ can be redesignated as d_β.[14] Thus:

$$P\left(x_i \mid d_\beta\right) = \frac{1}{e^{\beta(x_i - 1)} + \ldots + e^{\beta(x_i - 6)}} = \frac{e^{-\beta x_i}}{e^{-\beta} + \ldots + e^{-6\beta}} \tag{2}$$

The critical step involves subsequently applying the law of total expectation, yielding:

$$\frac{1}{6} = \int_{-\infty}^{+\infty} P\left(x_i \mid \hat{d_\varepsilon}\right) dF\left(\hat{d_\varepsilon}\right) = \int_{-\infty}^{+\infty} \frac{e^{-\beta x_i}}{e^{-\beta} + \ldots + e^{-6\beta}} dF\left(d_\beta\right) \tag{3}$$

for $i = 1, 2, \ldots, 6$, where $F\left(d_\beta\right)$ is the cumulative distribution associated with d_β.[15] The differential $dF\left(d_\beta\right)$ can be interpreted as the prior probability that the evidence d_β will occur. (An equivalent explanation holds for $dF\left(\hat{d_\varepsilon}\right)$.)

By analyzing the right-hand side of Eq. (3) as a function of x_i, one can see that its second-order derivative (with respect to x_i) is strictly positive unless $\beta = 0$.[16] Consequently, the integral is strictly convex in x_i when $\beta \neq 0$. However, the equation also implies that the integral must equal $\frac{1}{6}$ for each of the six distinct outcomes x_i. This condition can only be satisfied if $\beta = 0$, because, otherwise, the integral is a convex function of x_i and so cannot yield the same value at more than two values of x_i.

In terms of $\hat{d_\varepsilon}, \beta = 0$ corresponds to $\varepsilon = 3.5$. Thus, the crucial result is

$$P\left(\hat{d_\varepsilon}\right) = \delta(\varepsilon - 3.5) \tag{4}$$

where δ returns a value of 1 at 0 and otherwise returns 0.[17] Or, equivalently, (4) may be rewritten as

$$P(E_1(X) = 3.5) = 1. \tag{5}$$

[11] The following presentation is a combination of Friedman and Shimony's 1971 proof and Shimony's 1973 proof. For simplicity, I focus the arguments I present here on an example involving a six-sided die.

[12] I use $E_0(X)$ and $E_1(X)$ to denote the prior and posterior expectations, respectively, while $E(X)$ denotes generic information about expectations.

[13] In Eq. (1), the right-hand side represents the posterior derived via Maximum Entropy, whereas the left-hand side represents Bayesian conditioning on the evidence $\hat{d_\varepsilon}$.

[14] Accordingly, the cumulative distributions satisfy $F\left(\hat{d_\varepsilon}\right) = 1 - F\left(d_\beta\right)$.

[15] Eq. (3) follows directly from Eq. (1) and (2), i.e., the integrand of Eq. (3) aligns with the results from Bayesian conditioning.

[16] The second-order derivative of $e^{-\beta x_i}$ is larger than 0, except when $\beta = 0$. Moreover, $\frac{1}{e^{-\beta} + \ldots + e^{-6\beta}}$ is larger than 0.

[17] In the literature, people (Hobson, 1972; Shimony, 1973) take δ to be a Dirac Delta function, in order to emphasize that the probability mass clusters around a single point (at 0).

This means that the probability distribution for $\hat{d}_\varepsilon$ is degenerate, with probability one assigned exclusively to the value $\varepsilon = 3.5$ and probability zero assigned elsewhere. This result (Friedman & Shimony, 1971, p. 383) implies that the alignment between Bayesian conditioning and Maximum Entropy holds only when the evidence fixes the posterior expected value of X to the value of the prior $E_0(X)$, i.e., 3.5. According to Seidenfeld (1986, p. 482), this result - known as the F-S theorem - can be phrased as follows:

> The probability space defined over the outcomes can be extended if and only if the extension makes the constraint, $\hat{d}_\varepsilon = E_0(X) =$ *the average of outcomes*, practically certain.

Friedman and Shimony (1971, p. 383) view this as an unacceptable consequence for the attempt to align Bayesian conditioning with Maximum Entropy, arguing that it contradicts Jaynes' (1957) goal of deriving probability assignments that accurately reflect one's state of knowledge. Hobson (1972, p. 191) similarly claims that the result is seemingly absurd, noting that mere descriptions of experimental setups should not predetermine long-run averages. Seidenfeld (1986, p. 483) argues that an alignment between Maximum Entropy and Bayesian conditioning may hold only within a degenerate probability model.

4 Two Interpretations of the F-S Theorem

A central tension emerges here: Skyrms' space-extending approach appears effective in aligning the results of Bayesian conditioning with those from Maximum Entropy; yet, the F-S theorem sheds light on a seemingly problematic consequence of such an attempt. To clarify this tension, I will examine two potential interpretations of the F-S theorem.

The first interpretation simply takes the F-S theorem at face value.

> **Interpretation (1):** In the prior state, one must be certain that future evidence will establish a posterior expectation precisely equal to the prior expectation (e.g., 3.5 in the die example).

At first glance, Interpretation (1) suggests a seemingly unacceptable consequence of the F-S theorem, reflecting precisely the absurdity identified by Hobson. Given the initial setup, there is clearly no reason for an agent to anticipate with certainty that evidence will yield an exact posterior expectation. However, this interpretation results from a naïve reading of Eq. (4) or (5). Assigning a probability of one to an event in an infinite space does not necessarily imply absolute certainty or the logical exclusion of alternatives.[18] Within Skyrms' framework, once the agent learns that $E(X) = 5$, they update their belief accordingly, abandon their prior expectation $E_0(X) = 3.5$. In other words, Skyrms' approach to space extension does not imply that the agent must believe that any posterior expectation other than 3.5 is impossible. Interpretation (1), therefore, reflects an oversimplified understanding of the F-S result rather than an inherent flaw in Skyrms' method.

Recognizing that the agent's adoption of a probability of one in an infinite space does not imply absolute certainty motivates a second, more sophisticated interpretation.

> **Interpretation (2):** In the prior state, one is nearly certain (with probability one) that future evidence will indicate a posterior expected value close to the prior expectation (e.g., in the context of the die example, almost all (probability one) elements (sequences) in the extended space have an average sufficiently close to 3.5).

[18] This is a mathematical fact. In a countable infinite space, measure on each single point is 0. Correspondingly, all the points but one single point has measure of 1. Thus, probability 1 mean extremely high certainty but not necessarily absolute certainty.

This second interpretation aligns directly with the mathematical structure of the extended probability space Ω_E. Let us return to the die example. As N approaches infinity, nearly all of the sequences in Ω_E have an average that is sufficiently close to 3.5. Figure 2 illustrates the measure assigned to subsets of sequences in Ω_E that correspond to various average values. Intuitively, given the agent's uniform prior on Ω_0, $(\Omega_E, \mathscr{F}_E, P_E)$ enumerates the results of rolling a fair die up to N times (i.e., selecting elements uniformly and at random from Ω_0). Consequently, as N approaches infinity, a sequence's average converges to 3.5 by the law of large numbers. Furthermore, with an infinitely large N, because each sequence's average converges to 3.5 in limit, nearly all sequences in Ω_E will have averages sufficiently close to that value. Hence, in this extended space, $P\left(\hat{d}_{3.5\pm\varepsilon}\right) = 1$ for an arbitrarily small ε as N approaches infinity; this result reflects the mathematical fact that sequences deviating from 3.5 are extremely rare, i.e., have zero measure.

Interpretation (2) acknowledges the distinction between an agent's assigning probability one and their asserting absolute certainty. Skyrms' approach to space extension clusters the majority of the probability around $E(X) = 3.5$ but does not logically rule out any type of potential future evidence that could lead the agent's belief elsewhere. The problem highlighted by the F-S theorem is indeed a necessary, but conceptually explainable, consequence of Skyrms' approach. Provided the agent's uniform prior on the six outcomes, it is reasonable for them to believe that extremely biased sequences (e.g., sequences whose empirical average deviates far from 3.5) are improbable, yet not impossible. The clustering of averages around 3.5 embodies the agent's intention to fully utilize the information encoded in his prior beliefs. The acquisition of new information, such as $E(X) = 5$, nevertheless motivates the agent to revise their beliefs by referring to sequences compatible with the newly learned information. More importantly, Skyrms' approach has its own motivation which is independent of the question of whether Bayesian conditioning under Skyrms' framework aligns with Maximum Entropy. Hence, the initially concerning implication of the F-S theorem is conceptually explainable and arguably less problematic than it first appears.

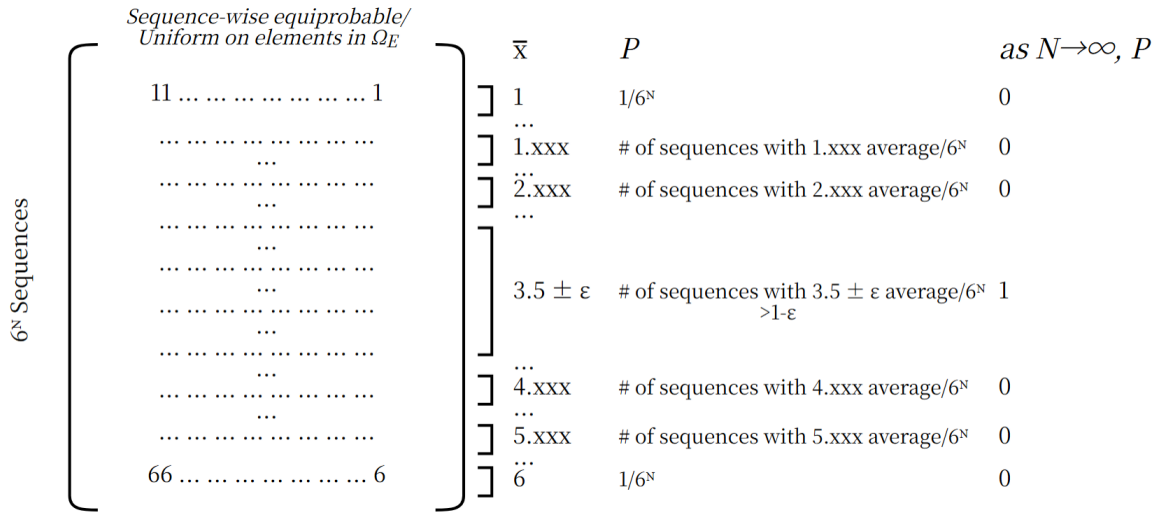


Figure 2: Measures of subsets of sequences with specified average values. Sequences in Ω_E are sorted according to their average values. The $\bar{x}$ column lists these average values; the P column shows the probability measures of these averages, calculated via frequency counts; and the final column shows the limits these measures converge to as N approaches infinity.[18]

More generally, Friedman and Shimony's identification of the consequence of aligning Bayesian conditioning and Maximum Entropy (as discussed in Section 3) and the reasoning illustrated in Figure 2 represent two different aspects of the same underlying fact. In their derivation of the F-S theorem, Freidman and Shimony rely on the alignment between the two methodologies, rather than basing their arguments on any particular extended probability space.[19] In contrast, Figure 2 demonstrates that the specific extended probability space proposed by Skyrms, unsurprisingly, exhibits precisely the same issue identified by the F-S theorem. In fact, adopting any extended probability space necessarily induces a corresponding prior distribution on future evidence.

For example, rather than considering sequences of outcomes from rolling a die, one could instead consider a set of all probability distributions defined over the six outcomes. Enumerating these distributions and assigning a uniform measure to them generates another extended space, denoted as $(\Omega'_E, \mathscr{F}'_E, \mu'_E)$. Each element (probability distributions) within this space corresponds to a possible value of $E(X)$, thus enabling Bayesian conditioning on the space. Furthermore, the extended space $(\Omega'_E, \mathscr{F}'_E, \mu'_E)$ is also consistent with a uniform prior on the six outcomes.[20] Of course, the reasoning captured by Figure 2 still applies within this alternative space. Figure 3 compares the probability distributions over future evidence $\hat{d}_\varepsilon$ induced by $(\Omega_E, \mathscr{F}_E, \mu_E)$ and $(\Omega'_E, \mathscr{F}'_E, \mu'_E)$.[21] Thus, if a derived probability distribution over $\hat{d}_\varepsilon$ were to prove problematic, this would represent a universal problem for the space-extension approach, rather than an issue unique to any one particular method of space-extension.

Moreover Hobson (1972) argues for a non-committal reading of Eq. (4) or Eq. (5), which is consistent with my Interpretation (2). He appeals to a distinction between inductive predictions and deductive implications:

> *"[I]nductive predictions (even when they are 'certain,' i.e., true with probability one) are only the best predictions possible on the basis of the given data. The predictions are not deduced from the data, they are only induced from the data. Thus, even if the data are true, the predictions... may turn out to be experimentally wrong." (Hobson, 1972, pp. 191-2)*

Accordingly, Hobson interprets the implication of the F-S theorem as of practically no force, and as simply the result of a harmless inductive prediction. His interpretation, which appeals to a different, conceptual aspect of Skyrms' framework, emphasizes that adopting a probability-one prediction does not commit an agent to that prediction in ways that undermine their ability to learn new evidence.

In summary, Interpretation (1) reflects a naïve understanding of probabilistic nuance, whereas Interpretation (2) accurately captures the implications of Skyrms' space-extending approach. According to Interpretation (2), the clustering of the probability mass around a single point is a conceptually coherent consequence of Skyrms' space-extension, rather than an inherently problematic one. Indeed, despite the conceptual motivation behind Skyrms' method of space extension - and the fact that any extended space inevitably implies some distribution over future evidence - it remains unclear whether such clustering of probabilities around a single value ought to be considered particularly bad or unacceptable. In the next section, I will shift my focus to Friedman and Shimony's perspective on the alignment of Bayesian conditioning with more general update rules, such as Maximum Entropy.

[19] The general implications of aligning Bayesian conditioning with an update rule will be discussed in Section 5.

[20] Due to space constraints, I cannot explore this approach further here. Briefly: The motivation is that agents ultimately care about their posterior distributions on the six outcomes; thus, it may be natural for an agent to adopt an extended space consisting of probability distributions themselves, directly characterizing the agent's concern. Grove and Halpern (1997) employ and motivate precisely this type of approach in modeling intuitions about the Judy Benjamin problem.

[21] The left figure is generated by sampling 10^9 sequences, each of length 10^8. The right figure is generated by sampling 10^9 distributions from a Dirichlet distribution. In both cases, histograms use a bin size of 0.002.

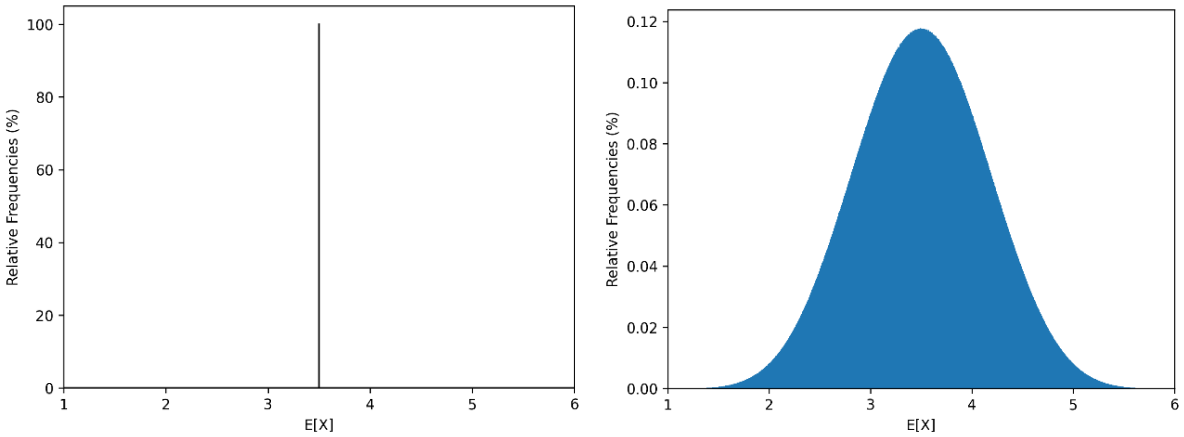


Figure 3: The left panel shows the distribution of elements (sequences) from the probability space $(\Omega_E, \mathscr{F}_E, \mu_E)$, categorized by their expectations. The right panel shows the corresponding distribution of elements (probability distributions) from the alternative probability space $(\Omega'_E, \mathscr{F}'_E, \mu'_E)$. The distributions of elements represent the agent's prior over future evidence $\hat{d}_\varepsilon$.

5 The Distribution of Evidence and Constraints from Alignment

As noted above, Hobson (1972) does not view the F-S theorem as posing a substantial threat to Maximum Entropy. Opponents like Shimony (1973) and Seidenfeld (1986) may nevertheless insist that the theorem reveals a deep flaw inherent in the Maximum Entropy approach, potentially leading to a stalemate in the debate over how best to interpret the F-S theorem. To propel this debate forward in a more constructive manner, I will proceed to explore the general implications of aligning Bayesian conditioning with an update rule, rather than focusing specifically on the alignment with Maximum Entropy. Generally speaking, it is clear that substantial constraints on the prior distribution over future evidence are unavoidable unless the alignment between Bayesian conditioning and the rule is vacuous. Consequently, I will, below, investigate the prospect of achieving a uniform prior on future evidence, arguing that such an approach is largely *ad hoc*.

Admittedly, clustering the mass of probability around a single point results in a degenerate probability model. Whether this degeneracy provides sufficient grounds upon which to reject the approach that leads to the degeneration is a separate question. Mathematically, the clustering of the probability mass, while somewhat extreme, is still, qualitatively, merely a constraint on a probability distribution. Recall that Eq. (1) aligns conditional probabilities with the solution from Maximum Entropy, whereas Eq. (3) enforces this alignment through the law of total expectation. Within the context of the F-S theorem, enforcing this alignment results in a clustering of the mass of probability over $\hat{d}_\varepsilon$ around a single point. However, unless the alignment is trivial (i.e., a uniform distribution $P\left(x_i \mid \hat{d}_\varepsilon\right) = \frac{1}{6}$ for any $i \in \Omega_0$),[22] an alignment under the law of total expectation inevitably places substantial constraints on the probability distribution over $\hat{d}_\varepsilon$. Intuitively, Friedman and Shimony's procedure involves three mutually constraining elements: (1) the prior on the six outcomes, (2) the alignment between Bayesian conditioning and a specific update target (e.g., Maximum Entropy in the context of the F-S theorem), and (3) the (prior) probability distribution on future evidence. Fixing any two of these elements will necessarily impose constraints on the third.

[22] In this case, $\frac{1}{6} = \int_{-\infty}^{+\infty} P\left(x_i \mid \hat{d}_\varepsilon\right) dF\left(\hat{d}_\varepsilon\right)$ holds under any distribution $F\left(\hat{d}_\varepsilon\right)$.

In the context of the F-S theorem, then, both fixing a uniform prior on outcomes and aligning Bayesian conditioning with Maximum Entropy leads to the observed clustered distribution on evidence.

To illustrate how such constraints arise, consider an arbitrary update rule. Suppose that, for $E(X) = \varepsilon$, the update rule distributes probability according to $\varepsilon = p\lfloor\varepsilon\rfloor + (1-\text{p})\lceil\varepsilon\rceil$, yielding $P(\lfloor\varepsilon\rfloor) = \lfloor\varepsilon\rfloor + 1 - \varepsilon, P(\lceil\varepsilon\rceil) = \varepsilon - \lfloor\varepsilon\rfloor$, and 0 otherwise.[23] Enforcement of this arbitrary rule results in constraints including $F\left(\hat{d}_1\right) = 0, F\left(\hat{d}_2\right) \approx 0.4294, F\left(\hat{d}_3\right) \approx 0.6636, F\left(\hat{d}_4\right) \approx 0.7751, F\left(\hat{d}_5\right) \approx 0.8931$ and $F\left(\hat{d}_6\right) = 1$, where F is the cumulative distribution function of $\hat{d}_\varepsilon$.[24] Here, the alignment between Bayesian conditioning and this arbitrary update rule yields constraints that are less restrictive than the constraints observed by Friedman and Shimony (i.e., the clustering of all probability around a single point). Nonetheless, I would argue that, regardless of their apparent lesser force, these constraints are still qualitatively on a par with those specified by Eq. (4) or Eq. (5). As I demonstrated in Section 4, assigning measure zero to values other than 3.5 does not mean that the agent rules out any potential future evidence that could lead their belief elsewhere, nor does it negatively impact their ability to condition within the extended space. Hence, there is little reason to regard a set of constraints that require the mass of probability to cluster around one point as qualitatively inferior to any other possible set of constraints.

Thus, substantial constraints on $F\left(\hat{d}_\varepsilon\right)$ are an inevitable outcome of the type of alignment employed by Friedman and Shimony, unless the alignment target trivially matches the prior on each outcome. The F-S theorem specifically targets the alignment of Bayesian conditioning with Maximum Entropy. Hobson's non-committal stance suggests that the mere presence of constraints on the probability distribution over future evidence does not justify rejecting an approach that involves aligning Bayesian conditioning with some update rules. I would argue further that, as it is difficult to qualitatively differentiate amongst alternative sets of constraints, even if constraints on $F\left(\hat{d}_\varepsilon\right)$ may be generally considered problematic, this is not a problem specific to Skyrms' space-extending approach or the alignment of Bayesian conditioning with Maximum Entropy.

It remains unclear what kind of constraints to the critics like Seidenfeld, Friedman, and Shimony might find acceptable. Perhaps one could reverse-engineer an appropriate extended space from a satisfactory distribution on evidence. To this end, an intriguing possibility might involve adopting a uniform distribution on $\hat{d}_\varepsilon$ that is motivated by the principle of indifference. However, such uniformity would lead to even more questionable consequences. Given the mutual restriction among the three elements (i.e., the prior on the six outcomes, the alignment between Bayesian conditioning and the specific update target, and the probability distribution on future evidence), a uniform $P\left(\hat{d}_\varepsilon\right)$ forces one to either abandon a uniform prior or relinquish a stable alignment. Neither one of these consequences appears more desirable than the clustering of the mass of probability implied by Eq. (4) or Eq. (5). The requirement that one must give up one's uniform prior on the six potential outcomes simply because one encounters a certain kind of evidence seems unreasonable. Such a situation would involve an agent considering (without yet observing) a distribution of certain types of evidence, then deciding not to adopt a uniform prior on outcomes so as to have a uniform prior on that kind of evidence. Furthermore, encountering different kinds of evidence will prompt the agent to engage in different belief revisions. Reverse-engineering an appropriate extended space in this way, then, seems to have no concrete or intelligible basis in reality.

An alternative option involves maintaining uniform priors on both outcomes and evidence, but giving up the notion that a stable alignment is possible. This would be to admit that Maximum Entropy cannot be aligned with Bayesian conditioning. Then, whatever posterior maintained the wanted uniformity would become the correct result of conditioning. Such an approach leads to two problematic implications. First,

[23] To ensure that $P(x_i) = P(x_i \mid E(X) = 3.5)$, an exception can be made for $E(X) = 3.5$.

[24] See the Appendix for a detailed description of the constraints are derived.

alignment with an update target other than Maximum Entropy essentially implies a different method of space-extension (and, more specifically, one that may not refer to sequences of outcomes). This would directly contradict the motivations behind Skyrms' approach, which involves appealing to sequences of outcomes and assigning uniform prior over the sequences. Thus, the potential problems expand beyond merely determining an appropriate distribution on evidence, to questioning the legitimacy of referring to sequences altogether. Unlike Skyrms' method of space extension, such an alternative lacks intuitive justification, making it appear *ad hoc*.

Secondly, and more importantly, consistently maintaining uniform priors on various types of evidence may deprive Bayesian conditioning of systematic structure. Consider the evidence regarding both the expectation (F-S theorem) and variance of the outcomes of the die rolls. Expectation and variance represent distinct types of evidence. Mathematically, it is feasible for an agent to adopt two separate Bayesian conditioning formulas (e.g., $P\left(x_i \mid \hat{d}_\varepsilon\right) = f\left(x_i, \varepsilon\right)$ and $P\left(x_i \mid \hat{v}_\sigma\right) = g\left(x_i, \sigma\right)$) to preserve their uniform priors for expectation and variance, respectively.[25] While adopting two distinct formulas in this way is not inherently problematic, it is unlikely that these two formulas share a common structure. In contrast, Maximum Entropy provides a unified mechanism by which to update multiple types of evidence; this unified mechanism is independently justified by the maximization of uncertainty. The lack of a systematic justification would require additional, possibly *ad hoc*, formulas for each new kind of evidence, undermining the coherence of the Bayesian conditioning approach.

The alignment of Bayesian conditioning with a specific update target necessarily corresponds to a particular extended probability space.[26] As demonstrated in Section 4, the choice of a specific extended space entails a particular prior distribution on future evidence. The analysis that I have presented here further reveals that that alignment imposes substantial constraints on the prior distribution of this evidence. Unless one can qualitatively differentiate between any two extensions, it is unclear how to evaluate the acceptability of these constraints. Thus, the conclusion reached in this section is, necessarily, largely consistent with the conclusion of Section 4: imposing constraints on the distribution of evidence is a universal problem for space extension. Furthermore, the clustering of the mass of probability around a single point is not, in or of itself, a valid basis upon which to criticize Maximum Entropy or Skyrms' approach to space extension. Moreover, the alternative, which involves attempting to simultaneously preserve uniform priors on both outcomes and evidence, does not appear promising, either.

6 Conclusion

The F-S theorem is often interpreted as illustrating how Maximum Entropy fails to properly align with the Bayesian framework. This apparent lack of alignment between the two frameworks has been further leveraged as a critique of Maximum Entropy (or Skyrms' approach to space extension). Nevertheless, when it is considered in isolation, Bayesian conditioning struggles to address learned information that does not correspond to an event in the original outcome space (and therefore necessitates the use of an extended space). Contrary to Seidenfeld's suggestion, the core dilemma here is not that Maximum Entropy fails to serve as an extension of Bayesian conditioning. Rather, the dilemma lies within Bayesianism itself: Bayesian conditioning cannot handle certain types of learning scenarios that Maximum Entropy handles without difficulty.

To handle learned information that does not correspond to an event within the original outcome space,

[25] Needless to say, from the perspective of space-extending, there needs two distinct extended spaces.

[26] An extended probability space refers to a space in which the learned information is an event. In a broader sense, a space does not necessarily have to be constituted of sequences and may have a non-uniform measure over its sample space.

one must extend the space, thereby providing an enriched description of the situation. However, there are an indefinite number of ways of enriching the space, resulting in various possible extensions and posterior answers. Adopting any one specific extension implicitly introduces additional assumptions, which manifest as constraints on one's prior distribution over the evidence. As I have argued above, if the F-S theorem poses difficulties for Maximum Entropy, the reasoning behind those difficulties in fact exposes a universal, unavoidable issue inherent in any approach to space extension. Consequently, critics of Maximum Entropy must either soften their reliance on the F-S theorem, or else reject the entire project of space extension. Abandoning space extension, however, would not only imply that no approach to extending spaces can satisfactorily align Maximum Entropy with Bayesian conditioning, but also acknowledge that Bayesian conditioning is fundamentally incapable of handling information beyond the original outcome space.

References

[1] Jan M. van Campenhout & Thomas Cover (1981): *Maximum entropy and conditional probability*. *IEEE Transactions on Information Theory* 27(4), pp. 483–489, doi:10.1109/TIT.1981.1056374.

[2] Imre Csiszár (1984): *Sanov Property, Generalized I-Projection and a Conditional Limit Theorem*. *The Annals of Probability* 12(3), pp. 768–793, doi:10.1214/aop/1176993187. Available at `http://www.jstor.org/stable/2243326`.

[3] Imre Csiszár & Paul C. Shields (2004): *Information Theory and Statistics: A Tutorial*. Now Publishers, Germany, doi:10.1561/9781933019543.

[4] Amir Dembo & Ofer Zeitouni (1996): *Refinements of the Gibbs conditioning principle*. *Probability Theory and Related Fields* 104, pp. 1–14, doi:10.1007/BF01303799.

[5] Bas van Fraassen (1981): *A Problem for Relative Information Minimizers in Probability Kinematics*. *The British Journal for the Philosophy of Science* 32(4), pp. 375–379, doi:10.1093/bjps/32.4.375.

[6] Kenneth Friedman & Abner Shimony (1971): *Jaynes's maximum entropy prescription and probability theory*. *Journal of Statistical Physics* 3, pp. 381–384, doi:10.1007/BF01008275.

[7] Adam J. Grove & Joseph Y. Halpern (1997): *Probability Update: Conditioning vs. Cross-Entropy*. In: *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 208–214, doi:10.48550/arXiv.1302.1543.

[8] Arthur Hobson (1972): *The interpretation of inductive probabilities*. *Journal of Statistical Physics* 6, pp. 189–193, doi:10.1007/BF01023687.

[9] Edwin T. Jaynes (1957): *Information theory and statistical mechanics*. *Physical Review* 106(4), pp. 620–630, doi:10.1103/PhysRev.106.620.

[10] Per Martin-Löf (1974): *Repetitive structures and the relation between canonical and microcanonical distributions in statistics and statistical mechanics*. In: *Proceedings of Conference on Foundational Questions in Statistical Inference (Aarhus, 1973)*, pp. 271–294.

[11] I.N. Sanov, D. Quade & United States. Air Force. Office of Scientific Research (1958): *On the Probability of Large Deviations of Random Variables*. Air Force Office of Scientific Research Technical Note No, United States Air Force, Office of Scientific Research. Available at `https://books.google.com/books?id=pPnrNz7gDk8C`.

[12] Teddy Seidenfeld (1986): *Entropy and Uncertainty*. *Philosophy of Science* 53(4), pp. 467–491, doi:10.1086/289336. Available at `http://www.jstor.org/stable/187902`.

[13] Abner Shimony (1973): *Comment on the interpretation of inductive probabilities*. *Journal of Statistical Physics* 9, pp. 187–191, doi:10.1007/BF01016847.

[14] Brian Skyrms (1985): *Maximum Entropy Inference as a Special Case of Conditionalization*. *Synthese* 63(1), pp. 55–74, doi:10.1007/BF00485955.

[15] Anubav Vasudevan (2020): *Entropy and Insufficient Reason: A Note on the Judy Benjamin Problem*. *The British Journal for the Philosophy of Science* 71(3), pp. 1113–1141, doi:10.1093/bjps/axy013.

Appendix

Suppose that for $\varepsilon \in [1,6]$ but $\varepsilon \neq 3.5$,

$$P\left(x_i \mid \hat{d}_\varepsilon\right) = \begin{cases} \lfloor \varepsilon \rfloor + 1 - \varepsilon, \text{ when } x_i = \lfloor \varepsilon \rfloor \\ \varepsilon - \lfloor \varepsilon \rfloor, \text{ when } x_i = \lceil \varepsilon \rceil \\ 0, \text{ otherwise} \end{cases}$$

and that when $\varepsilon = 3.5, P\left(x_i \mid \hat{d}_\varepsilon\right) = P\left(x_i\right)$.

According to the law of total expectation, for $x_i \in \{2,3,4,5\}$,

$$\frac{1}{6} = P\left(x_i \mid \hat{d}_\varepsilon\right) = \int_{x_i-1}^{x_i} (\varepsilon - x_i + 1)\, dF\left(\hat{d}_\varepsilon\right) + \int_{x_i}^{x_i+1} (x_i + 1 - \varepsilon)\, dF\left(\hat{d}_\varepsilon\right) \tag{A.1}$$

In addition, by definition, $F\left(\hat{d}_1\right) = 0$ and $F\left(\hat{d}_6\right) = 1$.

Insert x_i of values 1 through 5 into Eq. (A.1). There will be four equations: $F\left(\hat{d}_2\right)$, $F\left(\hat{d}_3\right), F\left(\hat{d}_4\right)$, and $F\left(\hat{d}_5\right)$. As $0 < F\left(\hat{d}_\varepsilon\right) < 1$, resolving these equations results in

$$F\left(\hat{d}_2\right) \approx 0.4294, F\left(\hat{d}_3\right) \approx 0.6636, F\left(\hat{d}_4\right) \approx 0.7751, F\left(\hat{d}_5\right) \approx 0.8931$$

On the Complexity of the Grounded Semantics for Infinite Argumentation Frameworks

Uri Andrews*
Department of Mathematics
University of Wisconsin-Madison
andrews@math.wisc.edu

Luca San Mauro†
Dipartimento di Ricerca e Innovazione Umanistica
University of Bari, Italy
luca.sanmauro@gmail.com

Argumentation frameworks, consisting of arguments and an attack relation representing conflicts, are fundamental for formally studying reasoning under conflicting information. We use methods from mathematical logic, specifically computability and set theory, to analyze the grounded extension, a widely-used model of maximally skeptical reasoning, defined as the least fixed-point of a natural defense operator. Without additional constraints, finding this fixed-point requires transfinite iterations. We identify the exact ordinal number corresponding to the length of this iterative process and determine the complexity of deciding grounded acceptance, showing it to be maximally complex. This shows a marked distinction from the finite case where the grounded extension is polynomial-time computable, thus simpler than other reasoning problems explored in formal argumentation.

1 Introduction

Over the past three decades, formal argumentation has established itself as a prominent research area within Artificial Intelligence, owing to its versatility in addressing various reasoning tasks. These include nonmonotonic reasoning, multi-agent systems, rule-based systems, and the analysis of debates or dialogues. Formal argumentation provides a unifying framework for representing diverse reasoning approaches, ranging from highly skeptical to more permissive forms of inference (for a comprehensive introduction to this area, see the handbook [4]).

At the heart of formal argumentation lies Dung's *abstract argumentation frameworks* (AFs) [15], which are modeled as directed graphs, where nodes correspond to arguments, and directed edges represent the attack relations between them. AFs serve as a common foundational core across various reasoning systems in formal argumentation, with many extensions and refinements, e.g. [9, 13, 16], building upon their structure. Moreover, the simplicity and abstraction of AFs make them a natural testing ground for developing and evaluating ideas before extending them to more sophisticated reasoning paradigms.

Although most of the research has focused on finite AFs, the study of infinite AFs is a natural and increasingly popular topic (see, e.g., [1–3, 7, 12]). A practical reason for studying infinite AFs is that rule-based argumentation formalisms (such as ASPIC [11] and ASPIC$^+$ [21]) are evaluated using Dung-style AFs and can generate infinitely many arguments even starting from a finite set of rules [11, 26]. As a theoretical foundation, infinite AFs enable the modeling of phenomena that finite frameworks cannot capture: for example, Dung [15, §3.1] showed that infinite AFs are needed to model solutions to n-person games. For additional examples of the applications and relevance of infinite AFs, we refer the reader to [3]. In the infinite setting, where conclusions may not always be attained in finite time, focus of computational research into reasoning problems shifts to finding appropriate limiting procedures.

*This research was supported by the NSF grant DMS-2348792.
†San Mauro is a member of INDAM-GNSAGA.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 112–127, doi:10.4204/EPTCS.437.13

© U. Andrews & L. San Mauro
This work is licensed under the
Creative Commons Attribution License.

In this paper, we use methods from mathematical logic, specifically computability theory and set theory, to examine the *grounded extension* in the context of infinite AFs. Introduced by Dung [15], the grounded extension identifies the unavoidable conclusions within an argumentation framework. The grounded extension, especially in the infinite context, plays a role in several applications of argumentation theory. For example, in cooperative games, the set of outcomes of a game forms an infinite AF where the grounded extension is exactly the supercore of the cooperative game (see [23,27]). Our results in this paper apply directly to the problem of determining the supercore of cooperative games.

The grounded extension is formally defined as the smallest fixed-point of the defense function (see Definition 3), but can also be understood as the outcome of an *iterative generative process*: This process begins by accepting arguments that are unanimously approved (i.e., unattacked). Once these are accepted, any argument attacked by an accepted one is invalidated and removed. Repeating this procedure yields a collection of arguments which should be accepted in (almost) any reasoning setting. This provides a natural limiting procedure to compute the grounded extension.

For finite AFs, the behavior of this iterative process is straightforward: If the AF considered contains n arguments, then the process stabilizes (and thus terminates) after at most n steps. It follows that determining membership in the grounded extension is computationally tractable, i.e., solvable in polynomial time. However, for infinite AFs, the situation is markedly different. The iterative generative process may run for infinitely many steps before stabilizing. Dung [15] noted this, as well as noting that for *finitary* AFs—those where each argument is attacked by only finitely many others— the process is guaranteed to terminate in ω many steps (that is, after doing one step for each natural number). Beyond this special case, however, the general behavior of the iterative process remains more complex, and Dung left open the question of how long it might extend.

In fact, it was not immediately clear that grounded extensions exist without assuming the power of full transfinite induction, which is known to be equivalent to the Axiom of Choice. In work exploring the theoretical aspects of the reasoning notions of formal argumentation, Spanring [25] demonstrated that the existence and uniqueness of grounded extensions in infinite AFs follow from Zermelo–Fraenkel set theory (ZF), without invoking the Axiom of Choice.[†] Yet, Spanring's proof is nonconstructive, relying on Hartog's Lemma, which ensures that no set is larger than all ordinals. As a result, the exact length of the iterative process remained undetermined.

We provide a new approach to understanding the grounded extension, rooted in computability theory. This approach allows to determine the exact length of the iterative generative process and to determine the exact computational complexity of the grounded extension. Furthermore, combining this analysis with a set-theoretic forcing argument, we show that not only the Axiom of Choice is unnecessary, but the very weak foundation of Kripke–Platek set theory suffices to define the grounded extension.

We formalize the question of the length of the iterative generative process by defining, for any given argumentation framework $\mathscr{F}$, its *grounding ordinal* (see Definition 6). We show that the grounding ordinal of a countable AF can be an arbitrarily large countable ordinal. In contrast, for computable AFs—which represent a very natural setting: e.g., many infinite AFs generated in applications are computable—we give a much lower, more concrete, and sharp upper bound.

Beyond analyzing the length of the iterative process, we also determine the algorithmic complexity of the grounded extension itself.

Theorem 1. *For computable AFs, the decision problem of determining membership in the grounded extension is* Π^1_1*-complete.*[§]

[†]In contrast, for other semantics widely studied in formal argumentation—such as the preferred and the naive ones—Spanring [24] proved that the existence of extensions in these semantics is equivalent to the axiom of choice over ZF.

[§]This result establishes the complexity of the skeptical acceptance of arguments with respect to the complete semantics for

Π^1_1-completeness (see Definition 17) is the natural analog in the infinite setting of co-NP-completeness. Π^1_1-completeness of the grounded extension shows that in the infinite setting, the grounded extension is as complex as the other computational problems explored for infinite AFs [2]. This result highlights a stark contrast with the finite case, where the grounded extension is solvable in polynomial time, and thus simpler than most other reasoning problems explored in formal argumentation. Moreover, we establish that a single computable AF suffices to achieve this hardness:

Theorem 2. *There is a computable AF $\mathscr{F}$ so that the grounded extension of $\mathscr{F}$ is a Π^1_1-complete set.*

This implies that the Turing degree of a Π^1_1-complete set can be the Turing degree of a grounded extension in a computable argumentation framework. This naturally leads to the question: What is the precise characterization of the collection of Turing degrees of grounded extensions in computable AFs? It is natural to conjecture that every Turing degree of a Π^1_1 set could arise as the degree of the grounded extension of some computable AF; we prove that this is not the case.

The rest of the paper is organized as follows. In Section 2, we review the necessary background on both argumentation theory and computability theory. In Section 3, we give examples of computable AFs with grounding ordinal equal to any $\alpha \leq \omega_1^{\mathrm{CK}}$ (see Definition 10), as well as a computable AF where the grounded extension is Π^1_1-complete. We then shift to the problem of showing that ω_1^{CK} is an upper bound for the grounding ordinal of a computable AF. In Section 4.1, we introduce our main technical tool towards this goal, which is a tree-rank analysis of groundedness or non-groundedness in an argumentation framework. In Section 4.2, we use this analysis to conclude that ω_1^{CK} is an upper bound for the grounding ordinal of any computable AF. In Section 5, we see that we can draw conclusions about noncomputable and even uncountable AFs. Here we establish that Kripke-Platek set theory suffices to define the grounded extension. Finally, in Section 6, we show that there are Π^1_1 Turing degrees which are not the degree of a grounded extension in a computable argumentation framework. Due to lack of space, some technical proofs have been relegated to an appendix.

2 Background

In this section, we provide a concise overview of key concepts in Dung-style argumentation theory, establish the basic terminology of computability theory, and introduce computable argumentation frameworks as a natural way to integrate these two domains. We assume the reader is acquainted with a few fundamental concepts from mathematical logic (such as ordinals and Turing degrees) as presented, e.g., in [22].

2.1 Argumentation theory background

An *abstract argumentation framework* (AF) $\mathscr{F}$ is a pair $(A_{\mathscr{F}}, R_{\mathscr{F}})$ consisting of a set $A_{\mathscr{F}}$ of arguments and an attack relation $R_{\mathscr{F}} \subseteq A_{\mathscr{F}} \times A_{\mathscr{F}}$. If an argument a attacks an argument b, we write $a \rightarrowtail b$ instead of $(a,b) \in R_{\mathscr{F}}$. We call collections of arguments $S \subseteq A_{\mathscr{F}}$ *extensions*. For an extension S, the symbols S^+ and S^- denote, respectively, the arguments that S attacks and the arguments that attack S: $S^+ = \{x : (\exists y \in S)(y \rightarrowtail x)\}$ and $S^- = \{x : (\exists y \in S)(x \rightarrowtail y)\}$.

Definition 3. $S \subseteq A_{\mathscr{F}}$ defends *an argument a, if any argument that attacks a is attacked by some argument in S (i.e., $\{a\}^- \subseteq S^+$). The* defense function *of $\mathscr{F}$ is the following mapping $f_{\mathscr{F}}$ which sends subsets of $A_{\mathscr{F}}$ to subsets of $A_{\mathscr{F}}$:*

$$f_{\mathscr{F}}(S) := \{x : \text{ } x \text{ is defended by } S\}.$$

infinite AFs, providing a proof of a theorem announced but left unproven in [1].

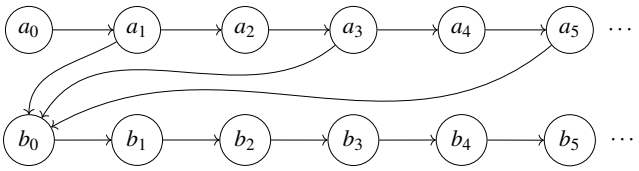


Figure 1: Example from Baumann-Spanring [8] of an AF with grounding ordinal $\omega \cdot 2$.

Dung [15, Definition 20] introduced the grounded extension to capture the collection of arguments which should be accepted by a maximally skeptical reasoner. Dung defined the *grounded extension* $\mathscr{G}$ as the least fixed-point of the defense function and noted that, for any AF, the grounded extension exists and is guaranteed to be conflict-free [15, Theorem 25], i.e., $a \not\rightarrowtail b$ for every $a, b \in \mathscr{G}$. A fundamental property of the grounded extension is that it can be approximated from below, by iterating along the ordinals the generative process sketched in the Introduction (see Definition 5).

Recall that the ordinals, introduced to mathematics by Georg Cantor in 1883, extend the counting numbers into the infinite. We will use the following two properties of the collection of ordinals:

Theorem 4. *The collection of ordinals is linearly ordered, and every non-empty subset of the collection of ordinals has a least element.*

Recall that the ordinal ω is the least infinite ordinal. Recall also that every ordinal α is either the successor of another ordinal β, i.e., $\alpha = \beta + 1$, or is a limit ordinal, i.e., $\alpha = \sup(Y)$, which is the supremum of the ordinals in the set Y. The ordinal ω is the first limit ordinal, and is the limit of the finite ordinals, i.e., the natural numbers.

Definition 5. *For an AF $\mathscr{F}$ and all ordinals α, let $\mathscr{G}_\alpha$ be inductively defined as follows:*

- $\mathscr{G}_0 = \emptyset$;
- *If $\alpha = \beta + 1$, then $\mathscr{G}_\alpha = f_{\mathscr{F}}(\mathscr{G}_\beta)$;*
- *If α is a limit ordinal, then $\mathscr{G}_\alpha = \bigcup_{\beta<\alpha} \mathscr{G}_\beta$.*

It follows from Dung's proof of the existence of the grounded extension that there is an ordinal β so that $\mathscr{G} = \mathscr{G}_\alpha$ for every $\alpha \geq \beta$.

Definition 6. *For an AF $\mathscr{F}$, the least ordinal α so that $\mathscr{G} = \mathscr{G}_\alpha$ is the* grounding ordinal *of $\mathscr{F}$.*

An initial bound on the size of the grounding ordinal for (countable) AFs can be established through a straightforward cardinality argument:

Observation 7. *If $\mathscr{F}$ is a countable AF, then the grounding ordinal of $\mathscr{F}$ is countable.*

Proof. Otherwise we would have that for each countable ordinal α, $\mathscr{G}_{\alpha+1} \supsetneq \mathscr{G}_\alpha$. But there are uncountably many countable ordinals. Thus, $\bigcup_\alpha \mathscr{G}_\alpha \subseteq A_{\mathscr{F}}$ would be uncountable. □

Dung proved that, for *finitary* AFs (i.e., those $\mathscr{F}$ so that the set $\{a\}^-$ is finite, for all $a \in A_{\mathscr{F}}$), the grounding ordinal is $\leq \omega$. Yet not every countable AF has a grounding ordinal $\leq \omega$. Baumann and Spanring [8] provide an example, illustrated in Figure 1, of an AF with a grounding ordinal of $\omega \cdot 2$.

In this paper, we determine exactly how large the grounding ordinal might be for a countable AF. Recall that ω_1 is the least uncountable ordinal. We will prove (Corollary 38) that *an ordinal α is a grounding ordinal of a countable AF if and only if α is a countable ordinal.*

In the next section, we introduce the computable argumentation frameworks and state a contrasting result to this, where we give a much better bound if $\mathscr{F}$ is computable.

2.2 Computability Background

Computable AFs are defined to be the AFs described by some computable function. We denote by $(\Phi_e)_{e\in\mathbb{N}}$ a uniform enumeration of all partial computable functions from $\mathbb{N}\times\mathbb{N}$ to $\{0,1\}$. The number e simply encodes the finite set of instructions of the Turing machine which computes Φ_e.

Definition 8. *A number e is a* computable index for an AF $\mathscr{F} = (A_{\mathscr{F}}, R_{\mathscr{F}})$ *with* $A_{\mathscr{F}} = \{a_n : n \in \mathbb{N}\}$ *if*

$$\Phi_e(n,m) = \begin{cases} 1 & \text{if } a_n \rightarrowtail a_m \\ 0 & \text{otherwise.} \end{cases}$$

An AF $\mathscr{F}$ is computable *if it has a computable index $e \in \mathbb{N}$.*

Computable ordinals are defined in a similar way, as those ordinals which are given by computable functions:

Definition 9. *A number e is a* computable index for a linear order $L_e = \{a_n : n \in \mathbb{N}\}$ *if*

$$\Phi_e(n,m) = \begin{cases} 1 & \text{if } a_n < a_m \\ 0 & \text{otherwise.} \end{cases}$$

If the linear order L_e is isomorphic to the ordinal α, then we say that α is a computable ordinal.

Observe that there are only countably many computable ordinals, and each of them is countable. However, there are uncountably many countable ordinals. Consequently, there exists a least ordinal that is not computable, but which is countable.

Notation 10. *The least non-computable ordinal is commonly denoted by ω_1^{CK}.*

Our main results (Theorems 22 and 35) on the grounding ordinals of computable AFs show that *the grounding ordinals of computable AFs are exactly the ordinals $\leq \omega_1^{CK}$.*

We let $\mathbb{N}^{<\mathbb{N}}$ represent the collection of finite sequences of natural numbers and $\mathbb{N}^{\mathbb{N}}$ represent infinite sequences of natural numbers. A string $\sigma \in \mathbb{N}^{<\mathbb{N}}$ is a *prefix* of a string $\tau \in \mathbb{N}^{<\mathbb{N}}$ or of a sequence $\tau \in \mathbb{N}^{\mathbb{N}}$ if there is some $\rho \in \mathbb{N}^{<\mathbb{N}}$ or $\rho \in \mathbb{N}^{\mathbb{N}}$ so that $\tau = \sigma^\frown\rho$, i.e, τ is the concatenation of σ with ρ.

If σ is a prefix of τ and the length of τ is 1 more than the length of σ, then τ is an *immediate extension* of σ, and we write $\sigma = \tau^-$.

A *tree* is a subset of $\mathbb{N}^{<\mathbb{N}}$ which is closed under prefixes. A path π through a tree T is an element of $\mathbb{N}^{\mathbb{N}}$ so that every prefix σ of π is in T. Note that since our trees may be infinitely branching, even trees which contain arbitrarily long strings may not have paths. Consider for example, the tree containing the empty string and all strings $i^\frown\sigma$ where the length of σ is $\leq i$.

For a tree T which does not have a path, we define the *rank* of T to be an ordinal which captures how difficult it is to see that T has no path.

Definition 11. *For any $\sigma \in T$, if σ is a terminal node in T (i.e., it has no extensions in T), then σ is assigned rank 0. If all immediate extensions of σ are ranked, then σ is ranked and has rank*

$$sup(\{rank(\tau)+1 : \tau \text{ is an immediate extension of } \sigma\}).$$

This process gives a rank to all elements σ so that σ is not a prefix of a path in T. If the root of the tree is assigned a rank α (i.e., there is no path in T), then we say the tree itself has rank α.

The following bounding theorems connect the computable ordinals to the ranks of computable trees:

Theorem 12 (see [22, §16.3, Theorem XIX(a)])**.** *If a computably enumerable tree has no path, then the rank of the tree is an ordinal $< \omega_1^{CK}$.*

Theorem 13 (see [22, §16.3, Theorem XIX(b)])**.** *If α is a computable ordinal, then there exists a computable tree with rank α.*

Next, we shift to the background needed for understanding our complexity result of Π^1_1-completeness of $\mathscr{G}$ for a computable AF.

Definition 14. *The Σ^1_1 sets are formally defined as those subsets of $\mathbb{N}$ that are definable in the language of second-order arithmetic using a single second-order existential quantifier ranging over subsets of $\mathbb{N}$ followed by number quantifiers and the first order functions and relations $(+,\cdot,<,0,1,\in)$; for more details, see [22, §16]. Π^1_1 sets are the complements of Σ^1_1 sets.*

Note that the Σ^1_1 sets are the analog in the infinite setting of the NP problems in the finite setting. Both are described by an exhaustive search over all subsets of the domain. Similarly, the Π^1_1 sets are the analog of the co-NP problems in the finite setting.

Observation 15. *The set of pairs (a,e), where the argument a belongs to the grounded extension of the AF with computable index e, is a Π^1_1 set. This follows from the fact that the grounded extension is the* smallest *fixed-point of the defense function $f_{\mathscr{F}}$. Specifically, $\mathscr{G}$ is defined by the formula $\forall X(f_{\mathscr{F}}(X) = X \to a \in X)$. Note that the defense function is defined in terms of the attack relation, which is computable, and quantifiers over arguments, which are encoded by natural numbers. Therefore, the only second-order quantifier involved is the initial universal quantifier $\forall X$.*

The following theorem states that the set of computable indices for trees which have paths is a maximally complex Σ^1_1 set.

Theorem 16 (Kleene [19])**.** *A set $X \subseteq \mathbb{N}$ is Σ^1_1 if and only if there is a computable sequence of computable trees $(\mathscr{T}^X_n)_{n\in\mathbb{N}}$ so that $n \in X$ iff $\mathscr{T}^X_n$ has a path.*

Definition 17. *A set X is* Π^1_1-complete *if it is Π^1_1 and whenever Y is a Π^1_1 set, there exists a computable function $f : \mathbb{N} \to \mathbb{N}$ so that $n \in Y$ if and only if $f(n) \in X$.*

Finally, all of these notions relativize. For example, for any Turing degree d, we say that an AF is d-computable, if there is a d-computable function Φ giving the attack relation. We say a set is $\Pi^1_1(d)$ if it can be defined in a Π^1_1 way using a predicate for an oracle in d. All the theorems in this section are valid in their relativized form.

3 AFs with maximally large grounding ordinals and Π^1_1-completeness

In this section, we begin with an example to see how we can build computable AFs with large grounding ordinals. Then we use this example to give computable AFs with any possible grounding ordinal, namely each $\alpha \leq \omega_1^{\mathrm{CK}}$.

Example 18. *Let $\mathscr{T}$ be a tree. Let $\mathscr{F}_{\mathscr{T}}$ be defined by having elements $\{a_\sigma : \sigma \in \mathscr{T}\} \cup \{b_\sigma : \sigma \in \mathscr{T}\}$. We let $b_\sigma \rightarrowtail a_{\sigma^-}$ where σ is an immediate extension of σ^- and $a_\sigma \rightarrowtail b_\sigma$. Note that $\mathscr{F}_{\mathscr{T}}$ is a computable AF if $\mathscr{T}$ is a computable tree.*

Theorem 19. *In $\mathscr{F}_{\mathscr{T}}$, $\mathscr{G}_{\beta+1}$ is exactly the set of a_σ so that σ has rank $\leq \beta$.*

We deduce the following theorem by applying Theorem 19 to all computable ordinals:

Theorem 20. *If $\alpha = \beta + 1$ is a computable successor ordinal, then there exists a computable AF $\mathscr{F}_\alpha$ with grounding ordinal α.*

By taking an appropriate union of the AFs from Example 18, we can also build AFs whose grounding ordinal is a computable limit ordinal.

Theorem 21. *If α is any computable ordinal, then there exists a computable AF $\mathscr{F}_\alpha$ with grounding ordinal α.*

We now proceed to construct a computable AF whose grounding ordinal is ω_1^{CK}.

Theorem 22. *There are computable AFs where the grounding ordinal is ω_1^{CK}. There are computable AFs where $\mathscr{G}$ is Π^1_1-complete.*

Proof. We construct the AF $\mathscr{F}$ in the following Lemma:

Lemma 23. *There is a computable AF $\mathscr{F}$ isomorphic to the disjoint union of $\{\mathscr{F}_\mathscr{T} : \mathscr{T}$ a computably enumerable tree$\}$.*

Note that an argument $a \in A_{\mathscr{F}_\mathscr{T}}$ is in $\mathscr{G}_\alpha$ in the disjoint union $\mathscr{F}$ if and only if it is in $\mathscr{G}_\alpha$ in $\mathscr{F}_\mathscr{T}$. Thus for the root $\lambda_\mathscr{T}$ of the tree $\mathscr{T}$, the argument $a_{\lambda_\mathscr{T}}$ is in $\mathscr{G}$ if and only if $\mathscr{T}$ has no path. It follows from Theorem 16 that $\mathscr{G}$ is a Π^1_1-complete set.

Since every computable tree is among the $\mathscr{T}$ included in $\mathscr{F}$, and there are trees with arbitrarily high rank among the computable ordinals by Theorem 13, we cannot have $\mathscr{G} = \mathscr{G}_\alpha$ for any $\alpha < \omega_1^{\mathrm{CK}}$. Thus the grounding ordinal is $\geq \omega_1^{\mathrm{CK}}$. On the other hand, every element of $\mathscr{G}$ is an $a_\sigma \in \mathscr{F}_\mathscr{T}$ where σ is ranked in $\mathscr{T}$ by Theorem 19. Then the set of strings $S = \{\tau : \sigma^\frown\tau \in \mathscr{T}\}$ forms a ranked computably enumerable tree with the same rank as the rank of σ in $\mathscr{T}$. Thus the rank of σ in $\mathscr{T}$ is some $\beta < \omega_1^{\mathrm{CK}}$ by Theorem 12, so $a_\sigma \in \mathscr{G}_{\beta+1} \subseteq \mathscr{G}_{\omega_1^{\mathrm{CK}}}$, which shows that $\mathscr{G} = \mathscr{G}_{\omega_1^{\mathrm{CK}}}$. □

Having established the existence of computable AFs with grounding ordinal α for any $\alpha \leq \omega_1^{\mathrm{CK}}$, we now turn our attention to proving that ω_1^{CK} is the largest possible grounding ordinal for a computable AF.

4 Upper bounds on grounding ordinals

From Observation 15, it follows that the grounded extension of any computable AF $\mathscr{F}$ is a Π^1_1-set. Hence, Theorem 16 ensures that there is a sequence of trees $\mathscr{T}^a$, for $a \in A_\mathscr{F}$, such that $a \in \mathscr{G}$ if and only if $\mathscr{T}^a$ has no path. In order to yield a restriction on the grounding ordinal of an AF, we need to construct these trees in a way that the rank of $\mathscr{T}^a$ corresponds to the least α such that $a \in \mathscr{G}_\alpha$. To do so, we highlight a natural class of extensions in AFs which we call "self-defending" and use these to give a dual understanding of the grounded extension in terms of the largest self-defending extension instead of the least fixed-point of $f_\mathscr{F}$. This construction is detailed in Section 4.1. Once established, we will apply Theorem 12 to conclude that each $a \in \mathscr{G}$ belongs to some $\mathscr{G}_\alpha$ with $\alpha < \omega_1^{\mathrm{CK}}$, thereby showing that $\mathscr{G}_{\omega_1^{\mathrm{CK}}} = \mathscr{G}$.

4.1 Rank analysis for groundedness

In order to show that an element a is in $\mathscr{G}$, we endeavor to show that every b which attacks a is in $\mathscr{G}^+$. That is, we show that $a \in f_\mathscr{F}(\mathscr{G})$, and since $\mathscr{G} = f_\mathscr{F}(\mathscr{G})$, this shows that $a \in \mathscr{G}$. Thus, we first build a tree $T_{\{b\}}$ so that $b \in \mathscr{G}^+$ if and only if $T_{\{b\}}$ has no path. We understand $\mathscr{G}^+$ in terms of the following property of $\mathscr{G}^+$.

Definition 24. *A set $X \subseteq A_{\mathscr{F}}$ is self-defending if whenever $y \in A_{\mathscr{F}}$ attacks some element of X, then there is some $x \in X$ so that $x \rightarrowtail y$. (Note that we are* not *assuming that X is conflict-free.)*

To the best of our knowledge, this result has been previously established, though we do not know an appropriate citation.

Lemma 25. *The largest self-defending extension is the complement of $\mathscr{G}^+$.*

This suggests a strategy for showing, in a Σ^1_1 way, that an element is in $A_{\mathscr{F}} \setminus \mathscr{G}^+$: Show that it is contained in a self-defending extension. For any finite set S, we produce a tree so that the paths through S encode self-defending extensions containing S.

Recall that we fix an enumeration of the arguments of $\mathscr{F}$ as $A_{\mathscr{F}} = \{a_n : n \in \mathbb{N}\}$.

Definition 26. *Let S be any finite subset of $A_{\mathscr{F}}$.*

For $\sigma \in \mathbb{N}^{<\mathbb{N}}$, define $mran_S(\sigma) = S \cup \{a_{\sigma(m)-1} : \sigma(m) \geq 1\}$.

For $\pi \in \mathbb{N}^{\mathbb{N}}$, define $mran(\pi) = \bigcup_{\sigma \preceq \pi} mran(\sigma)$.

We use a computable function $x, y \mapsto \langle x, y \rangle$ which is a bijection from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$.

Definition 27. *For S any finite subset of $A_{\mathscr{F}}$, we define a tree T_S as follows:*

- *The empty string is in T_S.*
- *Let $\sigma \in T_S$ with $|\sigma| = \langle n, m \rangle$.*
  - *Case 1: a_n attacks some element of $mran_S(\sigma)$. Then $\sigma^\frown i \in T_S$ if and only if $i > 0$ and $a_{i-1} \rightarrowtail a_n$;*
  - *Case 2: a_n does not attack any element of $mran_S(\sigma)$. Then $\sigma^\frown i \in T_S$ if and only if $i = 0$.*

Informally, each path π describes the set $\mathrm{mran}(\pi)$. At each level of the tree, we consider an element a_n and for any σ at that level of the tree, the immediate extension of σ must justify how the set $\mathrm{mran}(\pi)$ is self-defending in the face of a possible attack by a_n. The extension $\sigma^\frown 0$ represents the justification that there is no attack from a_n to any element of $\mathrm{mran}(\sigma)$, and the extension $\sigma^\frown (i+1)$ represents the justification that $a_i \in \mathrm{mran}(\pi)$ defends against an attack from a_n.

Lemma 28. *S is contained in a self-defending extension if and only if T_S has a path.*

Lemma 29. *$S \cap \mathscr{G}^+ = \emptyset$ if and only if T_S has a path.*

Proof. This follows directly from Lemmas 25 and 28. □

The following Lemma is the property of these trees which bridges between tree-rank and grounding ordinals.

Lemma 30. *For any $\rho \in T_S$, if the rank of ρ is $\leq \alpha$, then there is a member of $\mathscr{G}_{\alpha+1}$ which attacks some element of $mran_S(\rho)$.*

Proof. We prove this by induction on α. For $\alpha = 0$, let ρ be given and let $|\rho| = \langle n, m \rangle$. Then $\rho \in T_S$ having rank 0 means it is a terminal node, so a_n attacks some member of $\mathrm{mran}_S(\rho)$ and no element attacks a_n, i.e., $a_n \in \mathscr{G}_1$. Let $\alpha > 0$ be fixed and fix $\rho \in T_S$ with rank $\leq \alpha$. Let $|\rho| = \langle n, m \rangle$. If a_n does not attack any element of $\mathrm{mran}_S(\rho)$, then the tree above ρ has rank 1 more than the rank of the tree above $\rho^\frown 0$, i.e., $\alpha = \mathrm{rank}(\rho^\frown 0) + 1$. By inductive hypothesis applied to $T_S, \rho^\frown 0, \mathrm{rank}(\rho^\frown 0)$, some element of $\mathscr{G}_\alpha$ attacks some element of $\mathrm{mran}_S(\rho^\frown 0) = \mathrm{mran}_S(\rho)$.

Next, suppose that a_n does attack an element of $\mathrm{mran}_S(\rho)$. If $a_n \in \mathscr{G}_{\alpha+1}$, we are done, so we suppose $a_n \notin \mathscr{G}_{\alpha+1}$. Then there must be some i so that $a_i \rightarrowtail a_n$ and $a_i \notin \mathscr{G}^+_\alpha$. By inductive hypothesis, some member of $\mathrm{mran}_S(\rho^\frown (i+1)) = \mathrm{mran}_S(\rho) \cup \{a_i\}$ is attacked by a member of $\mathscr{G}_\alpha$. Since this cannot be a_i, we have that a member of $\mathscr{G}_\alpha \subseteq \mathscr{G}_{\alpha+1}$ attacks a member of $\mathrm{mran}_S(\rho)$. □

Applying this Lemma to the empty string, we get the following corollary:

Corollary 31. *If the rank of $T_{\{b\}}$ is less than α, then b is attacked by an element of $\mathscr{G}_{\alpha+1}$.*

We can thus use the trees $T_{\{b\}}$ for each b attacking a to analyze whether or not a has an attacker which is not in $\mathscr{G}^+$. Putting all these trees together, we define a tree $\mathscr{T}^a$ as follows:

Definition 32. *For each $a \in A_{\mathscr{F}}$, we define the tree $\mathscr{T}^a$ to be the tree which puts the empty string $\lambda \in \mathscr{T}^a$ and $i^\frown\sigma \in \mathscr{T}^a$ if and only if $a_i \rightarrowtail a$ and $\sigma \in T_{\{a_i\}}$.*

Corollary 33. *$\mathscr{T}^a$ has a path if and only if $a \notin \mathscr{G}$.*

Proof. $\mathscr{T}^a$ has a path if and only if there is some a_i with $a_i \rightarrowtail a$ and so $\mathscr{T}_{\{a_i\}}$ has a path if and only if there is some $a_i \rightarrowtail a$ so $a_i \notin \mathscr{G}^+$, if and only if a is not defended by $\mathscr{G}$ if and only if $a \notin \mathscr{G}$. □

Finally, we connect the rank of the tree $\mathscr{T}^a$ to the ordinal α so that $a \in \mathscr{G}_\alpha$.

Corollary 34. *If $\mathscr{T}^a$ has rank $\leq \alpha$, then $a \in \mathscr{G}_{\alpha+1}$.*

Proof. If $\mathscr{T}^a$ has rank $\leq \alpha$ then every a_i which attacks a has $\text{rank}(T_{\{a_i\}}) < \alpha$. Thus, each element of $\mathscr{A}_{\mathscr{F}}$ which attacks a is itself attacked by a member of $\mathscr{G}_\alpha$, and therefore $a \in \mathscr{G}_{\alpha+1}$. □

4.2 Concluding Upper Bounds

Theorem 35. *Let $\mathscr{F}$ be a computable AF. Then the grounding ordinal is $\leq \omega_1^{CK}$. That is, the grounded extension equals $\mathscr{G}_{\omega_1^{CK}}$, i.e., $\mathscr{G} = \mathscr{G}_{\omega_1^{CK}}$.*

Proof. Suppose that $a \in \mathscr{G}$. Then the computable tree $\mathscr{T}^a$ has no path. By Theorem 12, the rank α of $\mathscr{T}^a$ is $< \omega_1^{\text{CK}}$. It follows from Corollary 34 that $a \in \mathscr{G}_{\alpha+1}$, thus $a \in \mathscr{G}_{\omega_1^{\text{CK}}}$. □

5 Conclusions beyond computable AFs

Though our analysis was focused on computable AFs, everything relativizes to any Turing degree:

Theorem 36. *Let d be a Turing degree. There are d-computable AFs where the grounding ordinal is $\omega_1^{CK}(d)$. There are d-computable AFs where $\mathscr{G}$ is $\Pi^1_1(d)$-complete.*

Proof. We use the same construction as in Theorem 22 relativized to d. We take a d-computable sequence of all d-computable trees $\mathscr{T}_i$ and construct a d-computable AF $\mathscr{F}$ so that $\mathscr{G}$ is $\Pi^1_1(d)$-complete. It follows from the relativized version of Theorem 13 that the grounding ordinal of $\mathscr{F}$ is $\geq \omega_1^{\text{CK}}(d)$. It follows from the relativized version of Theorem 35 that the grounding ordinal of $\mathscr{F}$ is $\leq \omega_1^{\text{CK}}(d)$. □

Theorem 37. *For any countable ordinal α, there is a countable AF $\mathscr{F}$ such that the grounding ordinal for $\mathscr{F}$ is α.*

Proof. For any countable ordinal α, there is an ordering on $\mathbb{N}$ which is isomorphic to α. Let d be the Turing degree of this ordering. Then α is d-computable. By Theorem 21 relativized to d, there is a d-computable AF $\mathscr{F}$ so that the grounding ordinal is α. □

The following Corollary is immediate from Observation 7 and Theorem 37.

Corollary 38. *An ordinal α is a grounding ordinal of a countable AF if and only if α is countable.*

Since our analysis is constructive and we have strong bounds on the levels in the analysis, we are also able to extend the work begun by Spanring [25] and show that not only is the Axiom of Choice unnecessary to prove the existence and uniqueness of the grounded extension, but the very weak base theory of Kripke–Platek set theory (KP) suffices (see [6] for a reference on KP). ¶

Theorem 39. *There exists a formula $\psi(x,y)$ in the language of set theory so that, for any AF $\mathscr{F}$ in a model of KP, $\psi(x,\mathscr{F})$ defines the grounded extension of $\mathscr{F}$. Further, KP proves that for every AF $\mathscr{F}$, $\psi(x,\mathscr{F})$ defines a fixed point of $f_{\mathscr{F}}$ in $\mathscr{F}$. Further, KP proves for each $\rho(x,y)$ that for every Y, if $\rho(x,Y)$ defines a fixed point of $f_{\mathscr{F}}$ in $\mathscr{F}$, then $\forall x(\psi(x,\mathscr{F}) \to \rho(x,Y))$; that is, among all classes, ψ is the least fixed-point of $f_{\mathscr{F}}$ in $\mathscr{F}$.*

6 The Turing degrees of grounded extensions

We finally turn to the question of determining the exact complexity of the membership problem for the grounded extension of a given computable AF $\mathscr{F}$. Since $\mathscr{G}$ is a Π^1_1 set, and Theorem 22 demonstrated that $\mathscr{G}$ can be maximally complex among the Π^1_1-sets, it is natural to conjecture that the Turing degree of *any* Π^1_1 set could correspond to the complexity of $\mathscr{G}$ for some computable $\mathscr{F}$. It is a common phenomenon in computability theory that once the complexity of natural class $\mathscr{C}$ of sets is identified, this gives a characterization of the Turing degrees of members of $\mathscr{C}$. For example, the Turing degrees of word problems of finitely presented groups are exactly the Turing degrees of semicompuable sets [10, 14, 17].

However, we show that this is not the case for grounded extensions. We show this by identifying a new notion of a set being Π^1_1 *uniformly self-witnessing by paths* which we show holds for grounded extensions, showing that they have more structural uniformity than a general Π^1_1 set. This notion is of independent interest in computability theory.

Definition 40. *A set X is* Π^1_1 uniformly self-witnessing by paths *if there is a computable sequence of computable trees $(\mathscr{T}_i)_{i\in\mathbb{N}}$ and a Turing functional $\Psi(A,n)$ so that $\mathscr{T}_i$ has no path if and only if $i \in X$. Further, whenever $i \notin X$, then $\Psi(X',i)$ is a path in $\mathscr{T}_i$.*

Lemma 41. *Let $\mathscr{F}$ be a computable argumentation framework. Then the grounded extension is Π^1_1 uniformly self-witnessing by paths.*

The proof of the following Lemma requires several ideas from computability, including relativized hyperimmune-freeness and Friedberg jump-inversion.

Lemma 42. *There are Turing degrees of Π^1_1 sets which do not contain a set which is Π^1_1 uniformly self-witnessing by paths.*

These two Lemmas yield the following theorem.

Theorem 43. *There are Π^1_1-sets X so that the Turing degree of X is not the Turing degree of the grounded extension of any computable argumentation framework.*

A natural direction for future research is to characterize exactly which Turing degrees contain grounded extensions of computable AFs. Is it the degrees of sets which are Π^1_1 uniformly self-witnessing by paths?

¶We thank Noah Schweber for providing the proof of Theorem 39.

References

[1] Uri Andrews & Luca San Mauro (2024): *On computational problems for infinite argumentation frameworks: Hardness of finding acceptable extensions*. In: *Proceedings of the 10th Workshop on Formal and Cognitive Reasoning (FCR-2024)*, CEUR Workshop Proceedings. Available at `https://ceur-ws.org/Vol-3763/paper4.pdf`.

[2] Uri Andrews & Luca San Mauro (2024): *On computational problems for infinite argumentation frameworks: The complexity of finding acceptable extensions*. In: *Proceedings of the 22nd International Workshop on Nonmonotonic Reasoning (NMR-2024)*, CEUR Workshop Proceedings.

[3] Pietro Baroni, Federico Cerutti, Paul E Dunne & Massimiliano Giacomin (2013): *Automata for infinite argumentation structures*. *Artificial Intelligence* 203, pp. 104–150, doi:10.1016/j.artint.2013.05.002.

[4] Pietro Baroni, Dov Gabbay, Massimilino Giacomin & Leendert van der Torre, editors (2018): *Handbook of Formal Argumentation*. College Publications, London, England.

[5] Kenneth Jon Barwise (1967): *Infinitary Logic and Admissible Sets*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

[6] Kenneth Jon Barwise (1975): *Admissible Sets and Structures: An Approach to Definability Theory*. Springer-Verlag, doi:10.1007/978-3-662-11035-5. ISBN: 3540074511.

[7] Ringo Baumann & Christof Spanring (2015): *Infinite Argumentation Frameworks: On the Existence and Uniqueness of Extensions*. In: *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation: Essays Dedicated to Gerhard Brewka on the Occasion of his 60th Birthday*, Springer, pp. 281–295, doi:10.1007/978-3-319-14726-0_19.

[8] Ringo Baumann & Christof Spanring (2017): *A Study of Unrestricted Abstract Argumentation Frameworks*. 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. 807–813, doi:10.24963/IJCAI.2017/112.

[9] Dorothea Baumeister, Daniel Neugebauer, Jörg Rothe & Hilmar Schadrack (2018): *Verification in incomplete argumentation frameworks*. *Artificial Intelligence* 264, pp. 1–26, doi:10.1016/j.artint.2018.08.001.

[10] William W. Boone (1966): *Word Problems and Recursively Enumerable Degrees of Unsolvability. A Sequel on Finitely Presented Groups*. *Annals of Mathematics* 84(1), pp. 49–84, doi:10.2307/1970530. Available at `http://www.jstor.org/stable/1970530`.

[11] Martin Caminada & Leila Amgoud (2007): *On the evaluation of argumentation formalisms*. *Artificial Intelligence* 171(5-6), pp. 286–310, doi:10.1016/j.artint.2007.02.003.

[12] Martin Caminada & Nir Oren (2014): *Grounded semantics and infinitary argumentation frameworks*. In: *Proceedings of the 26th Benelux Conference on Artificial Intelligence*, BNAIC, pp. 25–32.

[13] Claudette Cayrol & Marie-Christine Lagasquie-Schiex (2013): *Bipolarity in argumentation graphs: Towards a better understanding*. *International Journal of Approximate Reasoning* 54(7), pp. 876–899, doi:10.1016/j.ijar.2013.03.001.

[14] C. R. J. Clapham (1964): *Finitely Presented Groups with Word Problems of Arbitrary Degrees of Insolubility*. *Proceedings of the London Mathematical Society* s3-14(4), pp. 633–676, doi:10.1112/plms/s3-14.4.633. arXiv:https://academic.oup.com/plms/article-pdf/s3-14/4/633/4257975/s3-14-4-633.pdf.

[15] 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.

[16] Bettina Fazzinga, Sergio Flesca & Francesco Parisi (2015): *On the complexity of probabilistic abstract argumentation frameworks*. *ACM Transactions on Computational Logic (TOCL)* 16(3), pp. 1–39, doi:10.1145/2749463.

[17] A. A. Fridman (1962): *Degrees of unsolvability of the word problem for finitely presented groups*. *Dokl. Akad. Nauk SSSR* 147, pp. 805–808.

[18] Richard Friedberg (1957): *A criterion for completeness of degrees of unsolvability*. The Journal of Symbolic Logic 22(2), pp. 159–160, doi:10.2307/2964177.

[19] Stephen C Kleene (1955): *Arithmetical predicates and function quantifiers*. Transactions of the American Mathematical Society 79(2), pp. 312–340, doi:10.1090/S0002-9947-1955-0070594-4.

[20] Webb Miller & D. A. Martin (1968): *The degrees of hyperimmune sets*. Z. Math. Logik Grundlagen Math. 14, pp. 159–166, doi:10.1002/malq.19680140704.

[21] 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.

[22] Hartley Rogers Jr (1987): *Theory of recursive functions and effective computability*. MIT press.

[23] Alvin E. Roth (1976): *Subsolutions and the Supercore of Cooperative Games*. Mathematics of Operations Research 1(1), pp. 43–49, doi:10.1287/moor.1.1.43. arXiv:https://doi.org/10.1287/moor.1.1.43.

[24] Christof Spanring (2014): *Axiom of choice, maximal independent sets, argumentation and dialogue games*. In: 2014 Imperial College Computing Student Workshop, Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

[25] Christof Spanring (2017): *Set- and graph-theoretic investigations in abstract argumentation*. Ph.D. thesis, University of Liverpool, UK.

[26] Hannes Strass (2013): *Approximating operators and semantics for abstract dialectical frameworks*. Artificial Intelligence 205, pp. 39–70, doi:10.1016/j.artint.2013.09.004.

[27] Anthony P. Young, David Kohan Marzagão & Josh Murphy (2020): *Continuum Argumentation Frameworks from Cooperative Game Theory*. In Henry Prakken, Stefano Bistarelli, Francesco Santini & Carlo Taticchi, editors: Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020, Frontiers in Artificial Intelligence and Applications 326, IOS Press, pp. 403–410, doi:10.3233/FAIA200524.

A Additional details for Section 3

Theorem 19. *In $\mathscr{F}_{\mathscr{T}}$, $\mathscr{G}_{\beta+1}$ is exactly the set of a_σ so that σ has rank $\leq \beta$.*

Proof. We check by an induction on ordinals β that $\mathscr{G}_{\beta+1}$ is exactly the set of a_σ so that σ has rank $\leq \beta$. For $\beta = 0$, we note that $\mathscr{G}_1$ is the unattacked elements of $\mathscr{F}_{\mathscr{T}}$, i.e., the a_σ where σ is a terminal node on $\mathscr{T}$, meaning σ has no immediate extensions in $\mathscr{T}$. These are exactly the a_σ where σ has rank ≤ 0.

For successor $\beta = \gamma+1$, $a_\sigma \in \mathscr{G}_{\beta+1} = \mathscr{G}_{\gamma+2}$ if and only if $a_\tau \in \mathscr{G}_{\gamma+1}$ for every immediate extension τ of σ. This, by inductive hypothesis is exactly if every τ an immediate successor of σ has rank $\leq \gamma$, which is equivalent to the rank of σ being $\leq \beta = \gamma+1$.

Finally, if β is a limit, then $a_\sigma \in \mathscr{G}_{\beta+1}$ if and only if every a_τ for τ an immediate extension of σ is in $\mathscr{G}_\beta$. Since β is a limit, this means every a_τ for τ an immediate extension of σ is in some $\mathscr{G}_{\delta_\tau+1}$ for some $\delta_\tau < \beta$. But this is equivalent by the inductive hypothesis to the rank of τ being $\leq \delta_\tau < \beta$ for every τ an immediate extension of σ. But this is equivalent to the rank of σ being $\leq \beta$. □

Theorem 20. *If $\alpha = \beta + 1$ is a computable successor ordinal, then there exists a computable AF $\mathscr{F}_\alpha$ with grounding ordinal α.*

Proof. By Theorem 13, we can take a tree $\mathscr{T}$ with rank β. Then the grounded extension in $\mathscr{F}_{\mathscr{T}}$ in Example 18 is the set of all a_σ since every $\sigma \in \mathscr{T}$ is ranked. Further $\mathscr{G} = \mathscr{G}_\alpha$ since every element of $\mathscr{T}$ has rank $\leq \beta$ and α is least such. □

Theorem 21. *If α is any computable ordinal, then there exists a computable AF $\mathscr{F}_\alpha$ with grounding ordinal α.*

Proof. If α is a successor ordinal, then the result is established in Theorem 20, so we suppose that α is a limit ordinal. Let e be a computable index for a linear order L_e isomorphic with α. Then by considering the sequence of sets $S_a := \{x \in L_e : x < a_n\}$ for each $n \in \mathbb{N}$, we get a uniformly computable sequence of indices $(i_n)_{n\in\omega}$ for computable ordinals β_n so that $\alpha = \sup(\{\beta_n : n \in \mathbb{N}\})$. Using Theorem 13, we can get a computable sequence $\mathscr{T}_n$ of computable trees so that the rank of $\mathscr{T}_n$ is β_n. Applying Example 18, we get a uniformly computable sequence $\mathscr{F}_{\mathscr{T}_n}$ of AFs with grounding ordinal $\beta_n + 1$. Finally, we let $\mathscr{F}$ be the disjoint union of the AFs $\mathscr{F}_{\mathscr{T}_n}$. Then the grounding ordinal of $\mathscr{F}$ is the supremum of $\{\beta_n + 1 : n \in \mathbb{N}\}$, which is α since each β_n is $< \alpha$ and α is a limit ordinal. □

Lemma 23. *There is a computable AF $\mathscr{F}$ isomorphic to the disjoint union of $\{\mathscr{F}_{\mathscr{T}} : \mathscr{T}$ a computably enumerable tree$\}$.*

Proof. We begin by indexing the computably enumerable (c.e.) presentations of all trees. For each index e, define the tree $\mathscr{T}_e$ such that $\sigma \in \mathscr{T}_e$ if and only if $\varphi_e(\sigma)\downarrow = 1$ for every $\tau \preceq \sigma$. Note that every computably enumerable tree can be represented as some $\mathscr{T}_e$, and each $\mathscr{T}_e$ is indeed a c.e. tree.

We construct an argumentation framework $\mathscr{H} = (A_{\mathscr{H}}, R_{\mathscr{H}})$ as follows $A_{\mathscr{H}} = \{(x,i) : x \in \mathscr{F}_{\mathscr{T}_i}\}$, where $\mathscr{F}_{\mathscr{T}_i}$ denotes the argumentation framework derived from tree $\mathscr{T}_i$ as in Example 18. The attack relation is defined by $(x,i) \to (y,j)$ if and only if $i = j$ and $x \to y$ in $\mathscr{F}_{\mathscr{T}_i}$.

This framework $\mathscr{H}$ is almost computable: its domain is a computably enumerable set, though not necessarily a computable one. To obtain a fully computable argumentation framework, observe that for any infinite c.e. set X, there exists a computable bijection $\beta : \omega \to X$. We use this to define a new framework $\mathscr{F} = (A_{\mathscr{F}}, R_{\mathscr{F}})$, where $A_{\mathscr{F}} = \{a_i : i \in \mathbb{N}\}$, and $a_x \to a_y$ in $\mathscr{F}$ if and only if $\beta(x) \to \beta(y)$ in $\mathscr{H}$. □

B Additional details for Section 4

Lemma 25. *The largest self-defending extension is the complement of $\mathscr{G}^+$.*

Proof. We first show that $\mathscr{G}^+ \cap X = \emptyset$ whenever X is self-defending. Suppose otherwise that X is self-defending and some element of $\mathscr{G}$ attacks an element of X. Let α be least so that some $a \in \mathscr{G}_\alpha$ attacks an element of X. Then since X is self-defending, some element $x \in X$ attacks a. But since $a \in \mathscr{G}_\alpha$, some element of $\mathscr{G}_\beta$ for $\beta < \alpha$ must attack x contradicting the minimality of α. It follows that any self-defending extension is contained in the complement of $\mathscr{G}^+$.

Now we show that the complement of $\mathscr{G}^+$ is self-defending. If a attacks $x \in A_{\mathscr{F}} \setminus \mathscr{G}^+$, then $a \notin \mathscr{G}$. Thus $a \notin f_{\mathscr{F}}(\mathscr{G})$. Thus there is some element $b \rightarrowtail a$ so that b is not in $\mathscr{G}^+$, so $b \in A_{\mathscr{F}} \setminus \mathscr{G}^+$ showing that $A_{\mathscr{F}} \setminus \mathscr{G}^+$ is self-defending. □

Lemma 28. *S is contained in a self-defending extension if and only if T_S has a path.*

Proof. Suppose that T_S has a path π. Let $X = \mathrm{mran}_S(\pi)$. We show that X is a self-defending extension. If a_n attacks an element of $\mathrm{mran}_S(\pi)$, then for some sufficiently large k, $\pi(\langle n,k\rangle) > 0$ and $a_{\pi(\langle n,k\rangle)-1} \in \mathrm{mran}_S(\pi)$ attacks a_n. Thus X is self-defending.

Suppose that S is contained in a self-defending extension X. Then we can define a path π in T_S by extending finite strings σ according to the following rule: Whenever $|\sigma| = \langle n,m\rangle$ and a_n attacks some element of $\mathrm{mran}_S(\sigma)$, extend σ by $i+1$ for some i with $a_i \in X$ and $a_i \rightarrowtail a_n$. Since X is self-defending, we can continue as such infinitely giving a path π in T_S. □

C Additional details for Section 5

We note that the proof of the following theorem uses the machinery of set-theoretic forcing. Loosely speaking, we understand truth in a given model A of KP-set theory by constructing a related larger model $A[G]$ with a convenient property, here that a given AF $\mathscr{F}$ is countable in $A[G]$, and then we use the relationship between A and $A[G]$ to transfer the definition of the grounded extension of $\mathscr{F}$ from $A[G]$ to A.

Theorem 39. *There exists a formula $\psi(x,y)$ in the language of set theory so that, for any AF $\mathscr{F}$ in a model of KP, $\psi(x,\mathscr{F})$ defines the grounded extension of $\mathscr{F}$. Further, KP proves that for every AF $\mathscr{F}$, $\psi(x,\mathscr{F})$ defines a fixed point of $f_{\mathscr{F}}$ in $\mathscr{F}$. Further, KP proves for each $\rho(x,y)$ that for every Y, if $\rho(x,Y)$ defines a fixed point of $f_{\mathscr{F}}$ in $\mathscr{F}$, then $\forall x(\psi(x,\mathscr{F}) \to \rho(x,Y))$; that is, among all classes, ψ is the least fixed-point of $f_{\mathscr{F}}$ in $\mathscr{F}$.*

Proof. We first show that there is a formula $\psi(x,y)$ so that whenever $\mathscr{F}$ is an AF in a model A of KP, then $\psi(x,\mathscr{F})$ defines the grounded extension of $\mathscr{F}$.

We note that our construction of $\mathscr{T}^a$ above requires listing off the elements of $\mathscr{F}$. Thus we will first give a formula ψ_c so that $\psi_c(x,\mathscr{F})$ defines the grounded extension for any countable $\mathscr{F}$ in a model A of KP. The formula $\psi_c(x,\mathscr{F})$ says that there exists an ordinal α which is a rank for the tree $\mathscr{T}^x$ as constructed above. Since the construction of $\mathscr{T}^x$ above is computable, this is uniformly defined in the model A. If A is a model of KP containing $\mathscr{F}$ and a is in the grounded extension of $\mathscr{F}$, then there is an $\mathscr{F}$-computable ordinal α so that α is the rank of $\mathscr{T}^a$. Since $A \models$ KP, this ordinal α is in A, so the formula $\psi_c(a,\mathscr{F})$ will hold. On the other hand, if a is not in the grounded extension of $\mathscr{F}$, then $\mathscr{T}^a$ has a path, so no ordinal can be a rank for $\mathscr{T}^a$. Thus $\psi_c(x,\mathscr{F})$ defines the grounded extension $\mathscr{F}$ in A whenever $\mathscr{F}$ is countable and A models KP.

We now use a forcing argument to extend this to all AFs $\mathscr{F}$ in any model A of KP: Suppose A is a model of KP with an infinite set and let $\mathscr{F}$ be an AF in A. Then the forcing partial order $Col(\omega, A_{\mathscr{F}})$ is an element of A, and it is known (see Barwise [5, Final Lemma in Appendix B]) that set forcing preserves admissibility, i.e., the forced model $A[G]$ also models KP, and satisfies the definability theorem in the following sense:

1. If G is a filter through $Col(\omega,X)$ meeting every definable-over-A dense subset of $Col(\omega,X)$, then $A[G]$ is a model of KP and $\mathscr{F}$ is countable in $A[G]$. Call such G "strongly generic over A."
2. Moreover, for each fixed formula φ the relation $p \Vdash \varphi(\nu)$, where p ranges over conditions and ν ranges over A-names, is definable in A. (Here we read "$p \Vdash \theta$" as "For every strongly generic G extending p, we have $A[G] \models \theta$.")

We can now reason as follows. Since $\psi_c(x,\mathscr{F})$ defines $\mathscr{G}$ whenever A models KP and $\mathscr{F}$ is a countable AF in A, *for every generic G*, $\psi_c(x,\mathscr{F})$ defines exactly $\mathscr{G}$ in $A[G]$. Thus, for every $a \in \mathscr{G}$, $\emptyset \Vdash \psi_c(a,\mathscr{F})$. Similarly, for every $a \notin \mathscr{G}$, $\emptyset \Vdash \neg\psi_c(a,\mathscr{F})$. Let $\psi(a,\mathscr{F})$ be the formula whose existence is guaranteed in (2) which defines the condition $\emptyset \Vdash \psi_c(a,\mathscr{F})$. Now $\psi(x,\mathscr{F})$ defines $\mathscr{G}$ for every A modeling KP and every AF $\mathscr{F}$ in A.

Now we shift to showing that KP proves that $\psi(x,\mathscr{F})$ defines a fixed-point of $f_{\mathscr{F}}$ and that for every other definable fixed-point $\rho(x,Y)$, KP proves that $\rho(x,Y)$ contains the set defined by $\psi(x,\mathscr{F})$. This argument is an application of the completeness theorem for first-order logic. Since we have shown that $\psi(x,\mathscr{F})$ defines the grounded extension $\mathscr{G}$ in any model of KP, the completeness theorem for first-order logic implies that every first order consequence of this fact is provable in KP. In particular, KP proves that $\psi(x,\mathscr{F})$ defines a fixed point of $f_{\mathscr{F}}$. Similarly, whenever $\rho(x,Y)$ defines a fixed-point of $f_{\mathscr{F}}$ in a model

of KP, it necessarily contains $\mathscr{G}$ since $\mathscr{G}$ is the least fixed-point of $f_{\mathscr{F}}$, and $\mathscr{G}$ is defined by $\psi(x,\mathscr{F})$. Thus, every model of KP satisfies the statement that whenever $\rho(x,Y)$ defines a fixed-point of $f_{\mathscr{F}}$, then $\rho(x,Y)$ contains $\psi(x,\mathscr{F})$, and the completeness theorem again shows that KP proves this first-order fact. □

D Additional details for Section 6

Lemma 41. *Let $\mathscr{F}$ be a computable argumentation framework. Then the grounded extension is Π^1_1 uniformly self-witnessing by paths.*

Proof. For each a, we have the computable tree $\mathscr{T}^a$ so that a is in the grounded extension if and only if $\mathscr{T}^a$ has no path. Note that if a is not in the grounded extension, then $\mathscr{T}^a$ has a path computed by $\mathscr{G}^+$. Namely, the path extends i where i is least so that $a_i \notin \mathscr{G}^+$. As in Lemma 28, we can compute a path through $T_{\{a_i\}}$ by successively either extending by 0 when in case 2 of Definition 27 or by $i+1$ where i is least so that $a_i \rightarrowtail a_n$ and $a_i \notin \mathscr{G}^+$. This is uniformly computable from $\mathscr{G}^+$, which is computable from $\mathscr{G}'$. □

We finally include the proof for Lemma 42. To do so, we introduce two additional concepts from computability theory: The hyperimmune-free degrees (relativized) and Friedberg jump inversion.

Definition 44. *For a fixed Turing degree d, a Turing degree c is said to be d-hyperimmune-free if whenever $f : \mathbb{N} \to \mathbb{N}$ is a function which is computable from c, there exists a d-computable function $g : \mathbb{N} \to \mathbb{N}$ so that $\forall x(f(x) \leq g(x))$.*

The notion of d-hyperimmune-freeness captures the notion that, though c may compute non-d-computable functions, none of these have faster growth-rates than d-computable functions. The following theorem is proven via a standard computability-theoretic construction for hyperimmune-free sets. We include this proof, which requires only straightforward changes from the original proof by Miller and Martin [20] of the existence of hyperimmune-free degrees.

Notation 45. *The symbol $\mathbf{0}$ represents the Turing degree that contains the computable sets.*

For a Turing degree d, the symbol d' represents the result of applying the Halting jump operator to the Turing degree d.

Generally, $d^{(n)}$ is the result of applying the Halting jump operator n times to the Turing degree d.

Theorem 46. *For any Turing degrees $c \geq d'$, there exists a d-hyperimmune-free degree a so $d <_T a <_T c'$, and $a \not\leq_T c$.*

Proof. We employ a forcing argument with d-computable function-trees.

Definition 47. *A function-tree is a function $T : 2^{<\mathbb{N}} \to 2^{<\mathbb{N}}$ so that $\sigma \preceq \rho$ if and only if $T(\sigma) \preceq T(\rho)$.*

We say that σ is on the tree T if there is some α so that $\sigma = T(\alpha)$. We say that $\widehat{T}$ refines T if every α which is on $\widehat{T}$ is also on T.

We build a sequence of d-computable function-trees T_i for $i \in \mathbb{N}$ so that T_{i+1} always refines T_i. Though each tree is d-computable, the sequence will be c'-computable.

Then the sequence $T_i(\lambda)$ where λ is the empty string satisfies $T_i(\lambda) \preceq T_{i+1}(\lambda)$. Let Y be the unique set so that $j \in Y$ if and only if $T_i(\lambda)(j) = 1$ for some or equivalently all sufficiently large i. This Y will be our hyperimmune-free set. Since the sequence of trees is c'-computable, it follows that Y is c'-computable.

Fix D to be any set in the degree d. We begin with T_0 being the d-computable function-tree $T_0(\sigma) = \sigma(0)D(0)\sigma(1)D(1)\dots\sigma(|\sigma|-1)D(|\sigma|-1)$. Since the set Y we construct is a path through T_0, we have already ensured that $Y \geq_T d$. At even stages we extend to ensure that $T_{2n}(\lambda)$ is incompatible with φ_n^c. That is, one of $T_{2n-1}(0)$ or $T_{2n-1}(1)$ is incompatible with φ_n^c. Using c' we can determine which. Suppose $T_{2n-1}(0)$ is incompatible with φ_n^c. Then we define $T_{2n}(\sigma) = T_{2n-1}(0^\frown\sigma)$. These stages collectively ensure that Y is not computable from c.

At odd stages, we ensure that Y is d-hyperimmune-free by forcing that either the nth computable function φ_n is total on T_{2n+1} or that φ_n is partial on T_{2n+1}. In the former case, this will ensure that φ_n^Y is d-computably-dominated. In the latter, this will ensure that φ_n^Y is partial.

To do so, we define a function-tree $\widehat{T}$ as follows: We let $\widehat{T}(\lambda) = T_{2n}(\lambda)$. Then for each σ, we search if there is some α so that $\widehat{T}(\sigma) \preceq T_{2n}(\alpha)$ so that $\varphi_n^\alpha(|\sigma|)\downarrow$. If found, then we let $\widehat{T}(\sigma^\frown i) = T_{2n}(\alpha^\frown i)$ for each $i \in \{0,1\}$. Using d'', which is $\leq_T c'$, we can check if $\widehat{T}$ defines a total function-tree. If so, we let $T_{2n+1} = \widehat{T}$. If not, then there is some α and m so that no $\beta \succeq \alpha$ has $\varphi_n^\beta(m)\downarrow$. In this case we let $T_{2n+1}(\sigma) = T_{2n}(\alpha^\frown\sigma)$ and we have ensured that $\varphi_n^Y(m)\uparrow$.

Finally, we need to check that Y is hyperimmune-free. Fix $e \in \mathbb{N}$. We must see that if φ_e^Y is a total function, then it is dominated by a d-computable function. If at step $2e+1$ we did not make $T_{2e+1} = \widehat{T}$, then we ensured that $\varphi_e^Y(m)$ does not converge, so φ_e^Y is not a total function. On the other hand, if we made $T_{2e+1} = \widehat{T}$, then we ensured that $\varphi_e^{\widehat{T}(\sigma)}(m)$ converges for each σ of length $m+1$. Thus $\varphi_e^Y(m) \leq \max\{\varphi_e^{\widehat{T}(\sigma)}(m) : |\sigma| = m+1\}$. This gives a d-computable function (since $\widehat{T}$ is d-computable) bounding φ_e^Y. □

Theorem 48 (Friedberg [18])**.** *Let d be any Turing degree above $\mathbf{0}'$. Then there exists another Turing degree c so that $d = c'$.*

Lemma 42. *There are Turing degrees of Π^1_1 sets which do not contain a set which is Π^1_1 uniformly self-witnessing by paths.*

Proof. Let $Y \geq_T \mathbf{0}'$ be a $\mathbf{0}^{(6)}$-computable set which is not $\mathbf{0}^{(5)}$-computable and is $\mathbf{0}'$-hyperimmune free; we apply Theorem 46 with $d = \mathbf{0}'$ and $c = \mathbf{0}^{(5)}$ to get such a set Y. Let e be the degree of a jump-inverse of Y as in Theorem 48. Since e is computable from $\mathbf{0}^{(6)}$, every set in e is Π^1_1.

Let X be any set in e. We claim that X is not Π^1_1 uniformly self-witnessing by paths. Suppose towards a contradiction that there were a computable sequence of trees $\mathscr{T}_i$ witnessing that X was Π^1_1 uniformly self-witnessing by paths. Then $i \in X$ if and only if Y computes a path through $\mathscr{T}_i$. But note that since Y is $\mathbf{0}'$-hyperimmune free, Y computing a path through $\mathscr{T}_i$ implies that there is a $\mathbf{0}'$-computable total function g so that $\mathscr{T}_i \cap \{\sigma \in \mathbb{N}^{<\mathbb{N}} : \forall n < |\sigma|\,(\sigma(n) < g(n))\}$ has a path. Thus, $i \in X$ if and only if $\exists g \leq_T \mathbf{0}'$ total so $\mathscr{T}_i^g := \mathscr{T}_i \cap \{\sigma \in \mathbb{N}^{<\mathbb{N}} : \forall n < |\sigma|\sigma(n) < g(n)\}$ has a path. We can now apply König's Lemma, which states that any finitely branching tree which has arbitrarily long strings must also have a path. Since $\mathscr{T}_i^g$ is finitely branching, it having a path is equivalent to it containing arbitrarily long strings. The tree T_i^g is uniformly computed from g, thus from $\mathbf{0}'$. Thus $\mathbf{0}^{(2)}$ can check the condition that the tree has arbitrarily long strings, and $\mathbf{0}^{(3)}$ can check that g is total. Finally, $\mathbf{0}^{(4)}$ can check the existence of a g satisfying the full condition. Thus X is computable in $\mathbf{0}^{(4)}$ showing that Y is computable in $\mathbf{0}^{(5)}$, which is a contradiction. □

Conditionals Based on Selection Functions, Modal Operators and Probabilities

Tommaso Flaminio
IIIA - CSIC
Campus de la UAB
Barcelona, Spain
tommaso@iiia.csic.es

Lluis Godo
IIIA - CSIC
Campus de la UAB
Barcelona, Spain
godo@iiia.csic.es

Giuliano Rosella
Department of Philosophy
University of Turin
Turin, Italy
giuliano.rosella@unito.it

1 Introduction

Methods for probability updating, of which Bayesian conditionalization is the most well-known and widely used, are modeling tools that aim to represent the process of modifying an initial epistemic state, typically represented by a prior probability function P, which is adjusted in light of new information. Notably, updating methods and conditional sentences seem to intuitively share a deep connection, as is evident in the case of conditionalization. Starting with a probability P, the question of what the probability of an event B is, given that A holds (i.e., $P(B \mid A)$), appears to involve conditional reasoning of the form "if A then B". Indeed, Adams [1] argues that the assertability of (indicative) conditionals, represented as "if A then B", aligns with the corresponding conditional probability $P(B \mid A)$. However, this intuitive connection between conditionals and updating methods has been challenged by Lewis's celebrated triviality result [20]. This result demonstrates that conditional probability $P(B \mid A)$ cannot generally be equated with the probability of the corresponding conditional, $P(A \rhd B)$, without leading to trivializing constraints on the initial probability functions.

Building upon these foundational works, research on the relationship between updating methods and conditionals has been and continues to be prolific. This research has primarily focused on understanding whether and how updating methods can be semantically represented in terms of conditional connectives, or conversely, what kind of updating methods are associated with specific semantic conditional operators (see, for instance, [22, 9, 15, 10, 7, 23]).

The present work contributes to this line of research and aims at shedding new light on the relationship between updating methods and conditional connectives. Departing from previous literature that often focused on a specific type of conditional or a particular updating method, our goal is to prove general results concerning the connection between conditionals and their probabilities. This will allow us to characterize the probabilities of certain conditional connectives and to understand what class of updating procedures can be represented using specific conditional connectives. Broadly, we adopt a general perspective that encompasses a large class of conditionals and a wide range of updating methods, enabling us to prove some general results concerning their interrelation.

Let us begin by providing some background. As observed in [14] (see also [9]), conditionalization updating can be situated within the broader context of *imaging* updating methods. Thge imaging method was initially introduced by Lewis [20] as a way to circumvent his triviality result and it hs been extended and generalized by several authors, notably by Gärdenfors [12]. While conditionalization is an updating method that requires only an algebra of events and a (positive) probability function, imaging methods update an a prior probability by employing a *closeness relation* among possible worlds (the atoms of the algebra), often interpreted as a measure of similarity. This closeness relation imposes a certain structure

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 128–139, doi:10.4204/EPTCS.437.14

© T. Flaminio, L. Godo & G. Rosella
This work is licensed under the
Creative Commons Attribution License.

on the underlying set of possible worlds that is employed within the context of imaging updating methods to determine which possible worlds the probability mass should be redistributed to, typically the closest (most similar) ones to a given one.

Such a closeness relation can also be used to specify the truth conditions of a wide range of conditional connectives, such as variably strict conditionals [19] and preferential conditionals [3, 21, 11]. The informal idea is that a conditional $a \rhd b$ is true at a world w if and only if b is true in the closest a-worlds to w, i.e., those closest worlds that make a true. The structure of closeness has been represented through various model-theoretic frameworks, among which ordered models and sphere models are particularly noteworthy. Sphere models were prominently used by Lewis [19] as the intended semantic device for specifying the truth conditions of variably strict conditionals (including counterfactual conditionals), while ordered models have been used, for example, to specify the truth conditions of preferential conditionals. Consequently, this closeness relation of possible worlds has been employed both in specifying imaging-like updating methods and in establishing a semantics for a large class of conditional connectives. Hence, intuitively, one might expect a deep relationship between imaging-like updating methods and conditionals based on a closeness relation, given that both rely on the same structure.

Our investigation will specifically focus on these two dimensions: imaging-like updating methods and conditionals, both having a semantics grounded on a closeness relation. Given our aim for generality, we will slightly deviate from the standard model-theoretic approach to conditionals and updating methods that relies on sphere models or ordered models to represent closeness relations, and we will utilize models based on *selection functions*. Notably, certain classes of ordered models and sphere models have equivalent representations in terms of selection functions (see, e.g., [18] and also [13]). One of the main advantages of the selection function is its technical utility: it is a very general tool for representing a general structure over worlds. However, one of its main shortcomings is conceptual: while orders and spheres offer an intuitive representation of a certain closeness relation, the relation induced by a selection function can sometimes be obscure and abstract.

In the present contribution we aim to provide a further generalization of imaging methods by relying on this more general setting of selection functions. Simultaneously, we aim to present results that conceptually enhance the understanding of certain updating procedures and selection function models. Inspired by this generality, we will primarily work within an algebraic setting, which also generalizes possible worlds models and imaging updating methods. Our basic framework will consist of a finite Boolean algebra $\mathbf{A}$, whose domain is denoted by the corresponding non-bold italic letter A, and whose atoms, forming the set $\mathrm{at}(\mathbf{A})$, will be thought as possible worlds. The finite algebra $\mathbf{A}$ will be equipped with a probability function P, whence $\mathrm{at}(\mathbf{A})$ will be endowed with a probability distribution that we still denote by P. The structure on A that we need to establish our general updating methods is represented by two basic tools (more technical details will be provided in the next sections):

- a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$ that, for every atom $\alpha \in \mathrm{at}(\mathbf{A})$ and every $a \in A$, specifies an element of A that correspond to the set of worlds closest to α according to a;
- a function $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$ that will be used to determine how to distribute the probability mass of each $\alpha \in \mathrm{at}(\mathbf{A})$ to its closest worlds $f(a,\alpha)$.

While the next section will introduce the basic notions and results that constitute the ground of our investigation, Section 3 will connect selection function-based conditionals with modalities and probabilities, while Section 4 will more directly deal with general updating methods. We will end the present paper with some conclusions that will be put forward in Section 5.

2 Selection Functions and Conditionals

Let us begin by establishing the fundamental components that define our framework. Boolean algebras will be presented in their usual language $(\wedge, \vee, \neg, \bot, \top)$ of type $(2,2,1,0,0)$ and where the implication operator $\rightarrow$ will be defined as $a \rightarrow b = \neg a \vee b$. All Boolean algebras that we consider in this paper are assumed to be finite; a generic Boolean algebra will be denoted by $\mathbf{A}$ and the set of its atoms is given by $\mathrm{at}(\mathbf{A}) = \{\alpha_1, \ldots, \alpha_n\}$. Let $\wp(X)$ stand for the powerset of a set X. Thus, $\mathbf{A}$ and $\wp(\mathrm{at}(\mathbf{A}))$ are isomorphic algebras whence we shall henceforth identify elements of A with subsets of $\mathrm{at}(\mathbf{A})$ and use a typical set-theoretical notation without danger of confusion. For instance, for $a \in A$, we will write $|a|$ to denote the cardinality of the subset of $\mathrm{at}(\mathbf{A})$ below a itself. Moreover, for $a, b \in A$ we write $a \subseteq b$ to state that a lies below b in the lattice order of $\mathbf{A}$ while, when $\alpha \in \mathrm{at}(\mathbf{A})$ and $a \in A$ we will simply write $\alpha \in a$ to denote that α is an atom among those that lie below a.

A selection function for $\mathbf{A}$ is a mapping $f : A \times \mathrm{at}(\mathbf{A}) \rightarrow A$ that associate to each pair of an atom and an element of A another element of A that we will identified with a set of atoms of $\mathbf{A}$ as observed above. Initially, we do not impose any further properties on this function. Intuitively, the selection function specifies the set of closest worlds (atoms) to a given world, relative to a certain element. For instance, $f(a, \alpha)$ represents the set of worlds closest to α, as determined by a. For every selection function $f : A \times \mathrm{at}(\mathbf{A}) \rightarrow A$, let us define an *selection function-based conditional* operator (or simply a *conditional* operator)

$$\rhd_f \colon A \times A \rightarrow A$$

as follows: for all $a, b \in A$,

$$a \rhd_f b = \{\alpha \in \mathrm{at}(\mathbf{A}) \mid f(a, \alpha) \subseteq b\}. \tag{1}$$

In other words, for all $\alpha \in \mathrm{at}(\mathbf{A})$,

$$\alpha \in a \rhd_f b \text{ iff } f(a, \alpha) \subseteq b.$$

As we have already anticipated, this algebraic setting has a very intuitive possible-worlds semantic counterpart: an atom α belonging to $a \rhd_f b$ can be regarded as the possible world α making $a \rhd_f b$ true. Consequently, given an intuitive interpretation of the selection function in terms of a closeness relation, the above condition is telling us that $a \rhd_f b$ is true at α if and only if all the closest a-worlds make b true.

This closeness relation, specified by f, is at this stage very general while it can be instantiated in different ways to represent, for instance, an order of similarity [19] or normality among worlds [3]. The logical properties of $\rhd_f$ will be induced by the constraints imposed on f, i.e., the type of closeness relation we aim to represent. Indeed, having imposed no specific properties on the selection function f does not inherently explain why the operator $\rhd_f$ defined as above should be considered a *conditional* operator. For example, it might be desirable that $a \rhd_f a$ is always true, or that $\rhd_f$ satisfies the modus ponens inference rule, and so on. All these properties can be enforced by imposing relevant constraints on the closeness relation involved in evaluating the conditional, namely the selection function. Below, we list several relevant constraints, including some that are discussed in [18]: consider a finite Boolean algebra $\mathbf{A}$ and a selection function $f : A \times \mathrm{at}(\mathbf{A}) \rightarrow A$, we can define the following properties for all $\alpha \in \mathrm{at}(\mathbf{A})$:

- $f(\bot, \alpha) = \bot$ (emptyness)
- for $a \in A \setminus \{\bot\}$, $f(a, \alpha) \neq \bot$ (normality)
- for $a \in A$, $f(a, \alpha) \subseteq a$ (identity)
- if $\alpha \in a$, then $\alpha \in f(a, \alpha)$ (centering-1)

- if $\alpha \in a$, then $f(a,\alpha) \subseteq \alpha$ (centering-2)
- if $\alpha \in a$, $f(a,\alpha) = \alpha$ (centering)
- $|f(a,\alpha)| \leq 1$ (uniqueness)
- if $f(a,\alpha) \subseteq b$ and $f(b,\alpha) \subseteq a$, then $f(a,\alpha) = f(b,\alpha)$ (well-order)
- $f(a \vee b,\alpha) \subseteq a$ or $f(a \vee b,\alpha) \subseteq b$ or $f(a \vee b,\alpha) = f(a,\alpha) \vee f(b,\alpha)$ (nesting)

The connection to the corresponding conditional connective is straightforward, and this allows to prove the following result, which readily follows from the conditions imposed on the selection function and equation (1) and from well-known results proved in [18]:

Fact 1. *Consider a finite Boolean algebra* $\mathbf{A}$ *and a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$*; the following hold for all* $a,b,c \in A$*, and all* $d \in A \setminus \{\bot\}$*:*

$\bot \rhd_f a = \top$	$\Leftrightarrow$	*f satisfies emptyness*
$d \rhd_f b \subseteq \neg(d \rhd_f \neg b)$	$\Leftrightarrow$	*f satisfies normality*
$a \rhd_f a = \top$	$\Leftrightarrow$	*f satisfies identity*
$a \rhd_f b \subseteq \neg a \vee b$	$\Leftrightarrow$	*f satisfies centering-1*
$a \wedge b \subseteq a \rhd_f b$	$\Leftrightarrow$	*f satisfies centering-2*
$a \wedge b \subseteq a \rhd_f b \subseteq \neg a \vee b$	$\Leftrightarrow$	*f satisfies centering*
$(a \rhd_f \neg b) \vee (a \rhd_f b) = \top$	$\Leftrightarrow$	*f satisfies uniqueness*
$(a \rhd_f b) \wedge (b \rhd_f a) \subseteq (a \rhd_f c) \leftrightarrow (b \rhd_f c)$	$\Leftrightarrow$	*f satisfies well-order*
$((a \vee b) \rhd_f a) \vee ((a \vee b) \rhd_f b) \vee (((a \vee b) \rhd_f c) = ((a \rhd_f c) \wedge (b \rhd_f c)))$	$\Leftrightarrow$	*f satisfies nesting*

For convenience, we may refer to $\rhd_f$ as to a *counterfactual conditional* when f satisfies the following properties: (i) identity, (ii) well-ordering, (iii) nesting, and (iv) centering. We call $\rhd_f$ a *variably strict conditional* when f satisfies (i) identity, (ii) well-ordering, and (iii) nesting. Finally, we use the term *Stalnaker conditional* for $\rhd_f$ when f satisfies (i) identity, (ii) well-ordering, (iii) nesting, (iv) centering, and (v) uniqueness. This terminology is justified by well-established results in the literature implying that that when $\rhd_f$ is a counterfactual conditional, it satisfies precisely the logical principles of Lewis's counterfactuals, namely it obeys the conditional logic C1 (see [18]). When $\rhd_f$ is a variably strict conditional, it obeys the conditional logic C0 (see [18]), which is the basic variably strict conditional logic (see [19]). And when $\rhd_f$ is a Stalnaker conditional, it obeys the logic C2 (see [18]), which corresponds to Stalnaker's logic of conditionals [25].

3 Modalities and Probabilities

In this section, we will establish some key results concerning the relationship between normal modal operators and selection functions. These results will subsequently be employed as a tool to begin exploring the connection between selection function-based conditionals and probability.

Consider a finite Boolean algebra $\mathbf{A}$ and a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$; let the binary relation $R^f_a \subseteq \mathrm{at}(\mathbf{A}) \times \mathrm{at}(\mathbf{A})$ be defined as follows: for all $\alpha \in A$, $R^f_a(\alpha) = f(a,\alpha)$. In other words,

$$\alpha R^f_a \beta \text{ iff } \beta \in f(a,\alpha).$$

For each such R^f_a, let $\Box^f_a : A \to A$ be defined as standard normal modal operator: for all $b \in A$

$$\Box^f_a(b) = \{\alpha \in \mathrm{at}(\mathbf{A}) \mid R^f_a(\alpha) \subseteq b\} = \{\alpha \in \mathrm{at}(\mathbf{A}) \mid f(a,\alpha) \subseteq b\}.$$

Notice that $\Box_a^f$ is, indeed, a normal modal operator in the sense of [2, Definition 1.42], that is to say, $\Box_a^f \top = \top$ and $\Box_a^f(b \to c) \subseteq (\Box_a^f b \to \Box_a^f c)$. Moreover, since $R_a^f(\alpha) = f(a,\alpha)$, comparing the latter with (1) we can immediately show the following result:

Fact 2. *For a finite Boolean algebra* $\mathbf{A}$ *and a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$*; for all* $a, b \in A$*, it holds that*

$$\alpha \in a \rhd_f b \text{ iff } \alpha \in \Box_a^f(b).$$

In other words, for every $a \in A$, $a \rhd_f (\cdot)$ and $\Box_a^f(\cdot)$ are the same unary operator on A. This implies that selection function-based conditionals can also be equivalently defined in terms of a family of indexed normal modal operators, $\{\Box_a^f \mid a \in A\}$, each of which behaves as a conditional with a fixed antecedent. Consequently, the structure $(\mathbf{A}, \{\Box_a^f\}_{a\in A})$ forms a *Boolean algebra with operators* in the sense of [2]. Following standard practice in modal logic, we define $\Diamond_a^f x$ as $\neg\Box_a^f\neg x$ and for all $b \in A$

$$\Diamond_a^f(b) = \{\beta \in \mathrm{at}(\mathbf{A}) \mid R_a^f(\beta) \wedge b \neq \bot\}.$$

Treating a conditional (binary) operator as a family of indexed normal modal (unary) operators provides a valuable framework for gaining new insights into the properties of conditional operators. This approach allows us to leverage the well-established behavior of normal modal operators and to demonstrate relevant results that connect the properties of a selection function f (and consequently, the corresponding conditional $\rhd_f$) with the properties of the induced indexed modal operators. The following result contributes to this direction:

Fact 3. *For a finite Boolean algebra* $\mathbf{A}$ *and selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$ *the following holds for all* $a \in A$*:*

$$|f(a,\alpha)| = 1 \Leftrightarrow (\textit{for all } x \in A,\ \Box_a^f x = \Diamond_a^f x).$$

Proof. The identity $\Box x = \Diamond x$ holds in a Boolean algebra with a modal operator $(\mathbf{A}, \Box)$ iff its associated Kriple frame $(\mathrm{at}(\mathbf{A}), R)$ satisfies that for all $\alpha \in \mathrm{at}(\mathbf{A})$, there exists a unique $\beta \in \mathrm{at}(\mathbf{A})$ such that $\alpha R \beta$, and hence R is a function from $\mathrm{at}(\mathbf{A})$ to itself, [2]. Thus in particular, $\Box_a^f x = \Diamond_a^f x$ for all $x \in A$ if and only if R_a^f is a function and hence, if and only if, for all α, there is a unique β such that $\alpha R_a^f \beta$, that is to say, there exists a unique $\beta \in f(a,\alpha)$. Therefore, if and only if $|f(a,\alpha)| = 1$. □

We now have the basic ingredients to begin investigating the relationship between conditionals and probability using the introduced general framework. For our investigation we will henceforth always assume that probability functions are positive, *i.e.*, they assign 0 only to the impossible event. Given the finite algebraic environment we are adopting, this choice comes at no cost. So, let $P : \mathbf{A} \to [0,1]$ be a positive probability on a finite Boolean algebra $\mathbf{A}$ and consider a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$. Then, given the above assumptions and definitions, we define

$$P(a \rhd_f b) = \sum_{\alpha \in a \rhd_f b} P(\alpha).$$

That is, the probability of a conditional $a \rhd_f b$ is given by the sum of the probabilities of the atomic states (worlds) where the conditional is satisfied. In other words, $P(a \rhd_f b)$ represents the probability that the conditional $a \rhd_f b$ is true. Directly from Fact 2, we obtain the following.

Fact 4. *For a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$*, and probability* $P : \mathbf{A} \to [0,1]$ *the following holds for all* $a, b \in A$*:*

$$P(a \rhd_f b) = P(\Box_a^f(b)).$$

As a direct consequence of the above Fact 4 and by the main results of [17], the map defined as

$$P(a \rhd_f (\cdot)) = P(\Box_a^f(\cdot)) \tag{2}$$

is a *belief function* in the sense of Dempster-Shafer theory [5]. Clearly, this observation does not preclude the possibility that $P(a \rhd_f (\cdot))$ is a probability function for some $a \in A$.

Let us now examine in more detail this connection between conditionals and belief functions. First, recall that belief functions differ from probability functions in their treatment of disjunction: probabilities are additive, satisfying $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$, whereas belief functions, denoted Bel, are superadditive, that is to say, $Bel(a \vee b) \geq Bel(a) + Bel(b) - Bel(a \wedge b)$.

Besides the axiomatic definition of belief functions, that the interest reader can find for instance in [16] and on which we will not rely for the rest of the present paper, it is convenient to recall how belief functions can be described in terms of their *mass distribution*. Specifically, every belief function is induced by a mass distribution m over a Boolean algebra $\mathbf{A}$, i.e., a function $m : A \to [0,1]$ such that $\sum_{a \in A} m(a) = 1$. From this mass distribution, the belief function is then defined as follows:

$$Bel(a) = \sum_{b \subseteq a} m(a). \tag{3}$$

Observe that, while probability functions on $\mathbf{A}$ are fully determined by distributions on $\mathrm{at}(\mathbf{A})$, belief functions on $\mathbf{A}$ are fully described by mass functions on A itself, and not only on its atoms. This easy observation marks the key difference between the two uncertain measures.

Given this definition, we can more vividly represent the belief function induced by a conditional $\rhd_f$ according to equation 2 by appealing to a generalization of the *imaging-like updating method* for belief functions introduced in [6]. Specifically, let us consider a finite Boolean algebra $\mathbf{A}$, a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$, and a probability $P : A \to [0,1]$. For a given element $a \in A$ we can define a mass distribution $m_a : A \to [0,1]$ as follows: $m_a(b) = \sum_{\alpha : f(a,\alpha) = b} P(\alpha)$. Then the corresponding *imaged* belief function is:

$$Bel_a(b) = \sum_{c \subseteq b} m_a(c) = \sum_{c \subseteq b} \left(\sum_{\alpha : f(a,\alpha) = c} P(\alpha) \right) = \sum_{\alpha : f(a,\alpha) \subseteq b} P(\alpha). \tag{4}$$

As a consequence, we obtain the following:

Fact 5. *Consider a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$*, and a probability* $P : \mathbf{A} \to [0,1]$*, the following holds for all* $a, b \in A$*:*

$$P(a \rhd_f (\cdot)) = P(\Box_a^f(\cdot)) = Bel_a(\cdot)$$

$$P(a \rhd_f b) = P(\Box_a^f(b)) = Bel_a(b)$$

Specifically, the probability of a selection function conditional $a \rhd_f b$ is equal to the corresponding belief function of the consequent imaged on the antecedent, i.e. $Bel_a(b)$.

This general result implies that the probability of a broad class of conditionals can be characterized in terms of an updated belief function. For instance, variably strict conditionals, which encompass both Stalnaker conditionals and Lewis counterfactuals (see [19, 18]), are known to be characterizable as selection function-based conditionals. Similarly, preferential conditionals [21, 11, 3] also fall under this category. Indeed, the semantics of preferential conditionals is typically defined using preorders, such that a preferential conditional $a \rhd_f b$ is true at a world α if and only if the minimal a-worlds with respect to the preorder associated with α are also b-worlds. This semantics can be reformulated such that the

set of minimal a-worlds, with respect to the preorder associated with α, constitutes a selection function $f(a,\alpha)$. Hence, the aforementioned result also holds for preferential conditionals. This result generalizes Lewis's [20] work on Stalnaker conditionals. Lewis demonstrated that the probability of a Stalnaker conditional $a \rhd_f b$ corresponds to the probability of the consequent b imaged on the antecedent a. Essentially, the probability of a Stalnaker conditional aligns with a specific updated probability distribution. Given that belief functions generalize probabilities, our finding indicates that the probability of selection function-based conditionals generally follows an updated belief function, and in specific instances, for instance in the case of Stalnaker conditionals, this updated belief function can indeed reduce to a probability, thereby implying Lewis's original result (further details are in subsequent sections).

Therefore, it is legitimate to ask under what conditions the belief function induced by a conditional is also a probability function. The next result provides an answer to this question, bringing together some of the results seen so far and characterizing the conditions under which the map defined in (2) is a probability function.

Proposition 1. *Consider a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$*, and a positive probability* $P : A \to [0,1]$*, for all* $a \in A$*, the following conditions are equivalent:*

(i) $P(a \rhd_f (\cdot)) : A \to [0,1]$ *is a probability function,*

(ii) $|f(a,\alpha)| = 1$*,*

(iii) R^f_a *is a function,*

(iv) for all $x \in A$*,* $\Box^f_a x = \Diamond^f_a x$*.*

Proof. The equivalence between (ii), (iii), and (iv) follows from Fact 3 and the definition of R^f_a. We now prove that (iv) implies (i). Indeed, if $\Box^f_a x = \Diamond^f_a x$ holds for all $x \in A$, we have that $\Box^f_a$ distributes over disjunction, namely $\Box^f_a(c \vee b) = \Box^f_a c \vee \Box^f_a b$, and so, by Fact 5, we would have that $Bel_a(c \vee b) = Bel_a(b) + Bel(b) - Bel_a(c \wedge b)$. Namely, the belief function Bel_a is additive and hence a probability; namely $P(a \rhd_f (\cdot)) : A \to [0,1]$ is a probability function. We now prove that (i) implies (ii). Let us reason by contraposition and assume that f doesn't satisfy uniqueness, namely that there is an atom α and an element $a \in A$ such that $|f(a,\alpha)| > 1$ or $f(a,\alpha) = \bot$. If the former holds, let $b = f(a,\alpha)$, then, by definition of Bel_a, we would have that $Bel_a(b) > 0$, since $m_a(b) > 0$. Thus, take $\beta \in b$ so that $b = \beta \vee (b \wedge \neg\beta)$, and then one has $Bel_a(b) > Bel_a(\beta) + Bel_a(b \wedge \neg\beta)$, whence Bel_a is not additive. Finally, if $f(a,\alpha) = \bot$, then $Bel_a(\bot) > 0$. Therefore, Bel_a is not a probability either. So, by Fact 5, $P(a \rhd_f (\cdot))$ is not a probability. □

The above results are general findings concerning conditionals and probability, and they can be interpreted as having limitative implications. On one hand, they demonstrate that the probability of a selection function-based conditional aligns with imaged belief functions. On the other hand, they establish that the belief function induced by a conditional is a probability function if and only if the selection function underlying the conditional satisfies the uniqueness property. Given these general results, we can now proceed to investigate whether more specific relationships exist between updating methods for probability and selection function-based conditionals.

4 Dealing with General Updates

As briefly mentioned in the introduction, Lewis [20] introduced *imaging* as an alternative to conditionalization for updating a probability measure $P : \mathbf{A} \to [0,1]$. Roughly, this procedure changes the initial

probability P by acting as if a certain event a holds. This update is accomplished by leveraging a similarity structure among possible worlds (or, equivalently, a selection function): the updated a-imaged probability P_a is obtained by shifting the probability mass of each atom α to its closest a-world. Crucially, this updating procedure relies on a uniqueness assumption: for each $\alpha \in \mathrm{at}(\mathbf{A})$, there is only one closest a-world to which α transfers its probability mass $P(\alpha)$. Gärdenfors [12] further generalized Lewis's imaging by relaxing this uniqueness assumption, allowing for multiple closest worlds. Building on this prior work and drawing inspiration from Güenther [14], we now introduce a general framework for defining updating methods for probabilities that generalizes both Lewis and Gärdenfors's imaging.

First, given a finite Boolean algebra $\mathbf{A}$ and a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$, we define a *distribution function for* f as a map $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$ satisfying the following constraints:

- $\lambda(a,\alpha)(\beta) = 0$ if and only if $\beta \notin f(a,\alpha)$
- $\sum_{\beta \in f(a,\alpha)} \lambda(a,\alpha)(\beta) = 1$

Observe that the second item above states that, for all $a \in A$ and $\alpha \in \mathrm{at}(\mathbf{A})$, $\lambda(a,\alpha)$ is a probability distribution on $f(a,\alpha)$.

Now, consider a finite Boolean algebra $\mathbf{A}$, a positive probability $P : \mathbf{A} \to [0,1]$, a selection function $f : A \times \mathrm{at}(\mathbf{A}) \to A$, and distribution function $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$ for f. Given an element $a \in A \setminus \{\bot\}$, we can define a probability distribution P_a^λ over $\mathrm{at}(\mathbf{A})$ as follows: for all $\beta \in \mathrm{at}(\mathbf{A})$

$$P_a^\lambda(\beta) = \sum_{\alpha : \beta \in f(a,\alpha)} \lambda(a,\alpha)(\beta) \cdot P(\alpha) \tag{5}$$

Specifically, the probability distribution $P_a^\lambda(\cdot)$ is obtained by redistributing the original probability mass of each atom α among all atoms in $f(a,\alpha)$, resulting in a new probability distribution. Hence, the corresponding updated probability function can be expressed as:

$$P_a^\lambda(b) = \sum_{\beta \in b} P_a^\lambda(\beta) = \sum_{\beta \in b} \left(\sum_{\alpha : \beta \in f(a,\alpha)} \lambda(a,\alpha)(\beta) \cdot P(\alpha) \right). \tag{6}$$

$P_a^\lambda(\cdot)$ represents an updated probability distribution where the shift of probability mass is performed according to the selection function f. The constraints imposed on the distribution function λ ensure that $P_a^\lambda(\cdot)$ is indeed a probability distribution. Moreover, all the probability mass is transferred to the atoms selected by f. Indeed, the following result can be proven from equation 6 and the definition of the distribution function λ:

Fact 6. *Consider a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$ *and a distribution mass* $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$*. For* $a \in A \setminus \{\bot\}$*,* $P_a^\lambda(\cdot)$ *is a probability function*

According to the general updating method for probabilities defined above, which utilizes the distribution function λ, several different updating procedures can be represented, including conditionalization [14], imaging [20], and general imaging [12]. Furthermore, since λ allows for considerable flexibility in the redistribution of probability mass, numerous other updating procedures, beyond these standard and common ones, can also be defined within this framework.

Having established this foundation, an important question remains: what is the relationship between selection function-based conditionals and λ-updating methods for probability? In light of the preceding results, we can demonstrate that these updating methods can be characterized as a probability induced

by a conditional only in specific cases, thus proving that selection function-based conditionals cannot capture the entire scope of λ-updating procedures.

First, let us prove a general result concerning the relationship between the probability of conditionals and updating methods, namely that the probability of a conditional $a \rhd_f b$ (with f satisfying normality) will always be less than or equal than the corresponding λ-updated probability of b given a:

Fact 7. *Consider a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$ *satisfying normality, and a distribution mass* $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$*. For* $a \in A \setminus \{\bot\}$*, and for all* $b \in A$*,*

$$P(a \rhd_f b) \leq P_a^{\lambda}(b)$$

Proof. (Sketch) We show that for any $\alpha \in a \rhd_f b$, the initial probability mass of $P(\alpha)$ contributes to the calculation of $P_a^{\lambda}(b)$. Specifically, let us consider $\alpha \in \mathrm{at}(\mathbf{A})$ such that $\alpha \in a \rhd_f b$. This implies that $f(a,\alpha) \subseteq b$. According to the definition of $P_a^{\lambda}(\cdot)$, the probability mass $P(\alpha)$ is redistributed among the elements in $f(a,\alpha)$. However, since $f(a,\alpha) \subseteq b$, we know that for every $\beta \in f(a,\alpha)$, $\beta \in b$. This implies that the mass $P(\alpha)$ from every $\alpha \in a \rhd_f b$ is redistributed among atoms that are below b. Consequently, this mass will be included in the summation that yields $P_a^{\lambda}(b)$. Thus, we have $\sum_{\alpha \in a \rhd_f b} P(\alpha) \leq P_a^{\lambda}(b)$, which is equivalent to $P(a \rhd_f b) \leq P_a^{\lambda}(b)$. □

We now have all the necessary components to show the conditions under which the probability of a conditional corresponds to an updated probability.

Theorem 1. *Consider a finite Boolean algebra* $\mathbf{A}$*, a selection function* $f : A \times \mathrm{at}(\mathbf{A}) \to A$ *satisfying normality, a distribution function* $\lambda : A \times \mathrm{at}(\mathbf{A}) \to [0,1]^{|\mathrm{at}(\mathbf{A})|}$*, and a positive probability* $P : A \to [0,1]$*. The following holds:*

$$(\textit{for all } a \in A \setminus \{\bot\} \textit{ and } b \in A,\ P(a \rhd_f b) = P_a^{\lambda}(b)) \Leftrightarrow \textit{ } f \textit{ satisfies uniqueness}$$

Proof. (Sketch) ($\Leftarrow$). This direction is straightforward. ($\Rightarrow$) By contraposition, assume that f doesn't satisfy uniqueness, namely there are α and a such that $|f(a,\alpha)| > 1$, by normality. Then, by Proposition 1, we have that $P(a \rhd_f (\cdot))$ is not a probability function, hence, a fortiori, it cannot be $P_a^{\lambda}(\cdot)$ which is a probability by Fact 6; therefore there must be some $a \in A \setminus \{\bot\}$ and $b \in A$ such that $P(a \rhd_f b) \neq P_a^{\lambda}(b)$. □

In conclusion of this section, let us examine the previous result in more detail. First of all, let us observe that if f satisfies the uniqueness constraint (along with normality), then for all $a \in A \setminus \{\bot\}$, $P(a \rhd_f (\cdot))$ and $P_a^{\lambda}(\cdot)$ coincide. This is because, under uniqueness, $\lambda(a,\alpha)$, being a probability distribution on $f(a,\alpha) = \{\beta\}$, it can only take values of 0 or 1; specifically, $\lambda(a,\alpha)(\beta) = 1$ and 0 otherwise. This means that each world (or atom) α transfers its entire probability mass to its unique closest a-world, as determined by f. Hence, due to uniqueness, the mass m_a, reduces to a probability distribution.

Conversely, when f does not satisfy uniqueness, it generally holds that for some $a, b \in A$, with $a \neq \bot$, $P(a \rhd_f b) < P_a^{\lambda}(b)$ as the following example shows.

Example 1. *Let us consider the algebra* $\mathbf{A}$ *of three atoms, i.e.* $\mathrm{at}(\mathbf{A}) = \{\alpha_1, \alpha_2, \alpha_3\}$*. Furthermore, conforming to our convention, let us represent the elements of A as subsets of* $\mathrm{at}(\mathbf{A})$*.*

Let us consider the selection function f *such that* $f(\{\alpha_2,\alpha_3\},\alpha_1) = \{\alpha_2,\alpha_3\}$*,* $f(\{\alpha_2,\alpha_3\},\alpha_2) = \{\alpha_2\}$ *and* $f(\{\alpha_2,\alpha_3\},\alpha_3) = \{\alpha_3\}$*. Notice that* f *does satisfy normality, yet not the uniqueness property because* $|f(\{\alpha_2,\alpha_3\},\alpha_1)| = 2$*.*

Let P be a positive probability on $\mathbf{A}$ *and* λ *be a distribution function. Recall from what we stated at the very beginning of this section that for all* $a \in A$ *and* $\alpha \in \mathrm{at}\mathbf{A}$*,* $\lambda(a,\alpha)$ *is a probability distribution on* $f(a,\alpha)$*. Thus, assuming* λ *to be positive, we have that*

- $\lambda(\{\alpha_2,\alpha_3\},\alpha_1)$ *distributes postively on* $f(\{\alpha_2,\alpha_3\},\alpha_1)=\{\alpha_2,\alpha_3\}$ *and therefore* $\lambda(\{\alpha_2,\alpha_3\},\alpha_1)(\alpha_2)>0$ *and* $\lambda(\{\alpha_2,\alpha_3\},\alpha_1)(\alpha_3)>0$*;*
- $\lambda(\{\alpha_2,\alpha_3\},\alpha_2)$ *can only assign* 1 *to* α_2 *being* α_2 *the unique element in its domain.*

Then, by definition:

- $P(\{\alpha_2,\alpha_3\}\rhd_f\alpha_2)=P(f(\{\alpha_2,\alpha_3\},\alpha_2))=P(\alpha_2)$*;*
- $P^{\lambda}_{\{\alpha_2,\alpha_3\}}(\alpha_2)=\lambda(\{\alpha_2,\alpha_3\},\alpha_1)(\alpha_2)\cdot P(\alpha_1)+\lambda(\{\alpha_2,\alpha_3\},\alpha_2)(\alpha_2)\cdot P(\alpha_2)=\lambda(\{\alpha_2,\alpha_3\},\alpha_1)(\alpha_2)\cdot P(\alpha_1)+1\cdot P(\alpha_2)$.

Therefore, since P is positive and $\lambda(\{\alpha_2,\alpha_3\},\alpha_1)(\alpha_2)>0$*,*

$$P(\{\alpha_2,\alpha_3\}\rhd_f\alpha_2)<P^{\lambda}_{\{\alpha_2,\alpha_3\}}(\alpha_2).$$

5 Conclusion

The present work aimed to demonstrate that the probability of selection function conditionals can be characterized in terms of updated belief functions within the framework of Dempster-Shafer theory. Conversely, classical Bayesian probabilistic updates (defined in terms of λ) cannot be characterized as the probability of a selection function conditional, except in a highly restricted case, specifically when the set of closest worlds involved in defining both the update and the conditional is a singleton. These results are quite general, as selection function conditionals represent a broad class of conditionals, including variably strict conditionals and preferential conditionals.

The results presented here can be further generalized to gain new insights into the logic of conditionals. In particular, the finding that the probability of selection function-based conditionals can be represented as a belief function suggests that this type of conditional might be representable in terms of a standard normal modal operator from modal logic. Indeed, our results in Section 3 can be further generalized to explore whether selection function-based conditionals can be represented as $a\rhd_f b\equiv\Box(a\to b)$, where $\to$ can be another binary connective. Previous work in this direction by [22] has shown that Lewis counterfactuals $a\rhd_f b$ can be represented as $\Box(b\mid a)$ within the framework of Boolean algebras of conditionals [8], where $(\cdot\mid\cdot)$ is a conditional that satisfies suitable algebraic counterparts of the laws of conditional probability. Given that Lewis counterfactuals are a specific type of selection function conditional, our findings in this work suggest that the representation result in [22] can also be extended to other kinds of selection function conditionals by imposing appropriate constraints on $\Box$.

On a more conceptual level, we have shown that a broad class of updating methods, including conditionalization, cannot be interpreted as the probability that a certain conditional connective holds. A question remains open regarding whether there exists a specific operator whose probability can represent these kinds of updating methods. Our results here indicate that the possible candidates are limited.

Acknowledgments

The authors are grateful to the anonymous referees for their careful reading and valuable suggestions to clarify several parts of the present paper. Flaminio and Godo acknowledge the Spanish projects SHORE (PID2022-141529NB-C21) and LINEXSYS (PID2022-139835NB-C21) respectively, both funded by MCIU/AEI/10.13039/501100011033. Flaminio and Godo also acknowledge partial support by the H2020-MSCA-RISE-2020 project MOSAIC (Grant Agreement number 101007627). Giuliano Rosella acknowledges financial support from the Italian Ministry of University and Research (MUR) through the PRIN 2022 grant n. 2022ARRY9N funded by the European Union (Next Generation EU).

References

[1] Ernest W. Adams (1975): *The Logic of Conditionals*. Reidel, Dordrecht, doi:10.1007/978-94-015-7622-2.

[2] Patrick Blackburn, Marteen de Rijke, Yde Venema (2002): *Modal Logic*. Cambridge University Press, Cambridge, doi:10.1017/CBO9781107050884.

[3] John P. Burgess (1981): *Quick Completeness Proofs for Some Logics of Conditionals*. Notre Dame Journal of Formal Logic 22(1), pp. 76–84, doi:10.1305/ndjfl/1093883341.

[4] Brian F. Chellas (1975): *Basic conditional logic*. Journal of Philosophical Logic 4(2), pp. 133–153, doi:10.1007/bf00693270.

[5] Arthur P. Dempster (1967): *Upper and lower probabilities induced by a multivalued mapping*. Ann. Math. Stat 38(2), pp. 325–339, doi:10.1214/aoms/1177698950.

[6] Didier Dubois & Henri Prade (1994): *A survey of belief revision and updating rules in various uncertainty models*. International Journal of Intelligent Systems 9(1), pp. 61–100, doi:10.1002/int.4550090105.

[7] Paul Égré, Lorenzo Rossi & Jan Sprenger (forthcoming): *Certain and Uncertain Inference with Indicative Conditionals*. Australasian Journal of Philosophy.

[8] Tommaso Flaminio, Lluis Godo & Hykel Hosni (2020): *Boolean algebras of conditionals, probability and logic*. Artificial Intelligence 286, p. 103347, doi:10.1016/j.artint.2020.103347.

[9] Tommaso Flaminio, Giuliano Rosella & Lluis Subirana (2025): *From conditionalization to imaging through the triviality glass*. Submitted Manuscript.

[10] Bas C. van Fraassen (1976): *Probabilities of Conditionals*. In: Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Springer Netherlands, pp. 261–308, doi:10.1007/978-94-010-1853-1_10.

[11] Marianna Girlando, Sara Negri & Nicola Olivetti (2021): *Uniform labelled calculi for preferential conditional logics based on neighbourhood semantics*. Journal of Logic and Computation 31(3), pp. 947–997, doi:10.1093/logcom/exab019.

[12] Peter Gärdenfors (1982): *Imaging and Conditionalization*. Journal of Philosophy 79, pp. 747–760, doi:10.2307/2026039.

[13] Gösta Grahne (1998): *Updates and Counterfactuals*. Journal of Logic and Computation 8, pp. 87–117, doi:10.1093/logcom/8.1.87.

[14] Mario Günther (2022): *Probabilities of Conditionals and Imaged Probabilities*. Draft manuscript.

[15] Alan Hájek (1989): *Probabilities of conditionals - Revisited*. Journal of Philosophical Logic 18(4), doi:10.1007/bf00262944.

[16] Joseph Y. Halpern (2003): *Reasoning about Uncertainty*. The MIT press, Cambridge, Massachusetts London, England, doi:10.7551/mitpress/10951.001.0001.

[17] David Harmanec, George J. Klir & Germano Resconi (1994): *On modal logic interpretation of Dempster-Shafer theory of evidence*. International Journal of Intelligent Systems 9(10), pp. 941–951, doi:10.1002/int.4550091003.

[18] David Lewis (1971): *Completeness and decidability of three logics of counterfactual conditionals 1*. Theoria 37(1), pp. 74–85, doi:10.1111/j.1755-2567.1971.tb00061.x.

[19] David Lewis (1973): *Counterfactuals*. Blackwell, Oxford, doi:10.2307/2273738.

[20] David Lewis (1976): *Probabilities of Conditionals and Conditional Probabilities*. Philosophical Review 85(3), pp. 297–315, doi:10.2307/2184045.

[21] Sara Negri & Nicola Olivetti (2015): *A Sequent Calculus for Preferential Conditional Logic Based on Neighbourhood Semantics*, pp. 115–134. Springer International Publishing, doi:10.1007/978-3-319-24312-2_9.

[22] Giuliano Rosella, Tommaso Flaminio & Stefano Bonzio (2023): *Counterfactuals as Modal Conditionals, and Their Probability*. Artificial Intelligence 323(C), p. 103970, doi:10.1016/j.artint.2023.103970.

[23] Paolo Santorio (forthcoming): *Probabilities of Counterfactuals are Counterfactual Probabilities*. Journal of Philosophy.

[24] Glenn Shafer (1976): *A Mathematical Theory of Evidence*. Princeton University Press, doi:10.2307/1268172.

[25] Robert Stalnaker (1968): *A Theory of Conditionals*. In Nicholas Rescher, editor: Studies in Logical Theory, American Philosophical Quarterly Monograph Series 2, Blackwell, Oxford, pp. 98–112, doi:10.1007/978-94-009-9117-0_2.

[26] Robert Stalnaker (1970): *Probability and Conditionals*. Philosophy of Science 37, pp. 64–80, doi:10.1086/288280.

Group Knowledge of Hypothetical Values

Alexandru Baltag
Institute for Logic, Language and Computation
University of Amsterdam
Amsterdam, the Netherlands
thealexandrubaltag@gmail.com

Sonja Smets
Institute for Logic, Language and Computation
University of Amsterdam
Amsterdam, the Netherlands
s.j.l.smets@uva.nl

In recent years, epistemic logics have been extended with operators $K_a x$ for knowledge of (the value of) a variable x (by an agent a). We study dynamic versions of these logics, enriched with modalities for semi-public data-exchange events (e.g., public announcements, data-sharing within a subgroup, or changing the value of a variable). To obtain a complete axiomatization of data-exchange events, in the presence of equality $x = y$ and $K_a x$, one needs to extend the logic further: first, with an operator for distributed knowledge $K_A x$ of the value (by a *group* of agents A); next, with a conditional version of this: distributed knowledge $K_A^{\varphi} x$ (of the value by a group) given some hypothetical condition (expressed by some proposition φ); then, with definite descriptions x_A^{φ}, denoting the 'hypothetical' value of x according to A's (distributed) knowledge given condition φ. In order to deal with common knowledge in the presence of semi-public data exchanges, we also need to add a *novel conditional version* of the recent concept of *common distributed knowledge*. We investigate the resulting logic, giving examples and presenting a complete axiomatization and a decidability proof.

1 Introduction

Information may come in both propositional and non-propositional (e.g. numerical or graphic) form. Such *data* can be encoded as *values of local variables* (e.g. a cryptographic key or a password), that can be stored in certain *locations* or 'sources' (e.g., databases, websites, etc.). Each source can be associated with an *agent*, either because it actually is the knowledge base of a (natural or artificial) agent, or because we think of the source itself as an abstract 'agent' (possessing exactly the information that is stored at it). More complex data may have an *extended location*, being *distributed* among a number of agents.[1]

In this setting, we will consider a number of *data-exchange actions* or 'events'. The agents may *jointly receive some propositional information*, in the form of a *public announcement*. An agent may *change the values of her local variables*, according to some substitution function. Or she may *gain access to another site*, in which case she can be assumed to instantly 'read' (and copy in her own knowledge base) all the information stored at that source. An agent may in fact gain simultaneous access to multiple such sources, and at the same time different agents may gain access to various sources. In this way, an agent may store data that are not 'owned' by her (e.g. values of some other agents' local variables)[2]. Groups of agents may *share or exchange data* in a symmetric or asymmetric manner. Finally, we can also consider events that combine all these features. In general, such data-exchange events can be *public* (when it is common knowledge that the data stored at a given location has become visible to everybody, cf. the WikiLeaks case), or *semi-public* (when the data are accessible only to some agents, but this fact is common knowledge), or *semi-private* (when some databases are publicly accessed, others are privately accessed, and these facts are known to some but not to all of the agents), or even *fully private* (e.g. a *secret hacking* of a database by an unsuspected intruder).

[1] One can recover the value of such complex variables only by accessing several sources.

[2] But we assume that each agent can change only the values of her own local variables.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 140–159, doi:10.4204/EPTCS.437.15

© A. Baltag, S. Smets
This work is licensed under the
Creative Commons Attribution License.

However, in this extended abstract, we follow the approach in [9], by *restricting ourselves to 'semi-public' data-exchanges*. The reason is that the syntax, the axioms and the proofs are simpler in this case, allowing us to focus at the new sources of complexity posed by the introduction of explicit variables and of the new operator of (conditional, distributed) knowledge of the value of a variable. The generalization to arbitrary events is left to the full journal version of this paper.

To reason about these scenarios, we introduce a logic with data variables x and equality of values $x = y$ (as well as other predicates and functions of variables), endowed with a Kripke-based semantics that can model both the traditional account of 'knowing that' ($K_a\varphi$ for *agent a knows that* φ) and an account of 'knowing what' ($K_a x$ for *a knows the value of* x). While work on the former ('propositional') knowledge $K_a\varphi$ is well known in standard epistemic logic, work on the later form $K_a x$ traces back to Plaza [21] and subsequent investigations by [16, 23, 24, 25, 26, 27, 28, 2] on formalizing 'knowledge de re'. While Plaza axiomatized the static logic of 'knowing what', the extension with public announcement operators $[!\varphi]\psi$ was only later axiomatized by Wang and Fan [25] via the introduction of a *conditional or 'hypothetical' version of 'knowing what'*: $K_a^\varphi x$ means that a knows the value of x given the information that proposition φ was the case. Note that this type of *conditional knowledge may not imply the knowledge of the 'real' value of* x (in the actual world), but only the fact that the agent *can uniquely determine a hypothetical value* of x (that would apply only to worlds in which φ was the case).

Part of the novelty of our setting is that we have a *way to explicitly refer to such hypothetical values*, via the introduction of a new term x_a^φ, denoting the *(unique) value that x would have according to agent a if* φ *were true*. (Here, we take x_a^φ to be undefined if such a unique value doesn't exist, either because the agent a knows that φ is false, or because the hypothesis φ is not enough for a to determine a unique such value of x.) These complex terms denoting hypothetical values seem to be required by the presence of equality $x = y$ and other predicates: e.g. we need a way to express the fact that the current value of x (at a world s at which φ is false) is the same as the hypothetical value of y (at worlds indistinguishable from s in which φ were true); and this can be expressed by writing that $x = y_a^\varphi$ holds at s.

The presence of data-exchange events forces us to further extend the static base of our logic. As noted in [9], when an agent a gains access to another agent b's knowledge base, her new state of knowledge matches the *distributed knowledge* of the group $\{a,b\}$. So on the side of propositional knowledge, we need operators $K_A\varphi$ for *distributed knowledge within group* A.[3] A similar move is also necessary on the side of "knowing what": one needs operators $K_A x$ that express that *the group A has distributed knowledge of the value of x*. And, because of the simultaneous presence of public announcements, we need in fact a *conditional version* of these operators: $K_A^\varphi x$ stands for group A having a *conditional distributed knowledge value of x, given the hypothesis* φ. Furthermore, due again to the presence of equality and other predicates, we are lead to generalize the terms for hypothetical-values to groups: x_A^φ will denote the *(unique) value that x would have according to (the distributed knowledge of) group A if* φ *were true*.[4]

We similarly follow [9] into generalizing common knowledge $C_A\varphi$ to super-groups $\mathfrak{A}$: for a family $\mathfrak{A}$ of groups of agents, $C_{\mathfrak{A}}\varphi$ denotes *common distributed knowledge* of φ, equivalent to the infinite conjunction of all finite iterations of $K_A K_B \ldots \varphi$ for all $A, B, \ldots \in \mathfrak{A}$: every group $A \in \mathfrak{A}$ has distributed knowledge that φ, and also has distributed knowledge that every other group $B \in \mathfrak{A}$ has distributed knowledge that φ, etc. Once again, this move is necessitated by the presence of (semi-public) data-exchange events, but

[3] The notation $D_A\varphi$ is typically used for distributed knowledge, but we prefer here $K_A\varphi$ because we think of this as capturing a natural notion of (virtual) group knowledge. Furthermore, this will allow us to use the notation $C_{\mathfrak{A}}\varphi$ for "common distributed knowledge", playing on the analogy between the relationship of individual knowledge K_a with common knowledge C_A on the one hand, and the relationship of distributed knowledge K_A with common distributed knowledge $C_{\mathfrak{A}}$ on the other hand.

[4] In addition to such terms for hypothetical values x_A^φ, we also have other complex terms of more traditional kinds, formed by applying functions $F(x_1, \ldots, x_n)$, or defined by cases.

the notion of common distributed knowledge has independent motivation and applications [9]. In fact, a further extension is needed here, in presence of public announcements: the novel notion of *conditional common distributed knowledge* $C^{\theta}_{\mathfrak{A}}\varphi$ appears for the first time in our current submission.[5]

In this paper we study the resulting logic of semi-public data-exchange events in the presence of (conditional) group knowledge of (hypothetical) values of variables, (conditional) common distributed knowledge and the other ingredients mentioned above. The main results are a complete axiomatization and a proof of decidability for this Dynamic Logic of (group) Knowledge of (hypothetical) values *DLKV*.

2 Syntax and Semantics

A *vocabulary* $\mathscr{V} = (V, \mathscr{A}, ag, Pred, Funct, ar)$ consists of the following:

- a (finite or infinite) set $\mathscr{A}$ of *agents* (or 'locations', or databases) $a, b, \ldots$;
- a (finite or infinite) set C of *constants*, that include *two logical constants* $0, 1$, and *one 'improper' constant* $\uparrow$ (denoting the '*undefined*' value).
- a (finite or infinite) set V of *(basic) local variables* $v^1_a, v^2_a, \ldots$, each being 'located' at some location $a \in \mathscr{A}$ (or 'owned' by some agent a) where its correct value is always stored; we indicate the location/owner via the subscript in v_a, but we sometimes drop the subscript, writing just v (or $v^1, v^2, \ldots$), when we are not interested in the variable's location;
- a set $Pred$ of *predicate symbols* $P, Q, \ldots$, including the *equality predicate* $(=)$, together with an *arity* $ar : Pred \to \mathbb{N}$, mapping each symbol $P \in Pred$ to a number $ar(P) \in \mathbb{N}$, where $ar(=) = 2$
- and a set $Funct$ of *function symbols* $F, G, \ldots$, also endowed with an *arity map* $ar : Funct \to \mathbb{N}$, associating to each function symbol $F \in Funct$ a natural number $ar(F) \in \mathbb{N}$.

Groups and Super-groups Given the vocabulary $\mathscr{V}$, a *group* $A \subseteq \mathscr{A}$ is a finite, non-empty set of agents in $\mathscr{A}$. We use letters $A, B, \ldots$ to denote groups. A *supergroup* $\mathfrak{A} \subseteq \mathscr{P}(\mathscr{A})$ is a finite, non-empty set of groups. We use letters $\mathfrak{A}, \mathfrak{B}, \ldots$ to denote supergroups.

Syntax of *DLKV*. The *Dynamic Logic of (group) Knowledge of (hypothetical) Values* $DLKV(\mathscr{V})$ (over the vocabulary $\mathscr{V}$) has a syntax, consisting of: a set $Var := Var(\mathscr{V})$ of compound *variables* (or 'terms') x (over the given vocabulary $\mathscr{V}$); a set $Fml; = Fml(\mathscr{V})$ of *formulas* φ (over $\mathscr{V}$); and a set $\mathscr{E}$ of ('semi-public') *events* e; these sets are simultaneously defined by mutual recursion:

$$\begin{array}{lcl} x & ::= & c \mid v \mid x|_{\varphi} y \mid F(\overline{x}) \mid x^{\varphi}_{A} \mid e(x) \\ \varphi & ::= & P\overline{x} \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_A\varphi \mid C^{\varphi}_{\mathfrak{A}}\varphi \mid [e]\varphi \\ e & ::= & !\Phi/\sigma \end{array}$$

where: $c \in C$ is any constant symbol; $v \in V$ is any basic variable; $x \in Var$ is any variable (term), and $\overline{x} = (x_1, \ldots, x_n)$ are n-tuples of variables in Var; $P \in Pred$ is any predicate symbol of arity n; F is any n-ary function symbols[6]; $a \in \mathscr{A}$ is any agent, $A \subseteq \mathscr{A}$ is any group of agents, and $\mathfrak{A} \subseteq \mathscr{P}(\mathscr{A})$ is any supergroup; $\Phi \subseteq Fml$ is a finite set of (already formed) formulas; and finally $\sigma : V \cup \mathscr{A} \to Var \cup \mathscr{P}(\mathscr{A})$ is a 'substitution' map, assigning to each basic variable $v \in V$ some term $\sigma(v) \in Var$ and assigning to each agent $a \in \mathscr{A}$ some group of agents $\sigma(a) \subseteq \mathscr{A}$. These components are subject to the following two constraints, for all $a \in \mathscr{A}$ and all basic local variables v_a of agent a:

[5] We stopped here, due to the limited size of this extended abstract. But the generalization to arbitrary events would require a further, wilder extension of the notion of common (distributed) knowledge, namely the introduction of polyadic modalities as in [10], denoting "conditional levels of group knowledge" in the spirit of [19].

[6] In particular, recall that the predicate symbols include the equality predicate $=$, and the constants $0, 1, \uparrow$. As usual, we will write $x = y$ instead of $= xy$.

1. $a \in \sigma(a)$ (*each agent continuously accesses her own database*);
2. $K_a\sigma(v_a) \in \Phi$ (*agents know what values they reassign to their own local variables*).

The static fragment *LKV*. The *Logic of (group) Knowledge of (hypothetical) Variables LKV*$(\mathscr{V})$ is the 'static fragment' of *DLKV*, obtained by removing the dynamic operators $[e]\varphi$ and $e(x)$ from (all the three components of) the recursive definition of the language *DLKV*.

Precondition, Access and Postconditions of an event Furthermore, for any event $e = !\Phi/\sigma$, any basic variable $v \in V$ and any agent $a \in \mathscr{A}$, we use the notations

$$pre_e := \bigwedge \Phi, \quad e(a) := \sigma(a), \quad post_e(v) := \sigma(v)$$

The first of these is called the *precondition* of event e (intuitively characterizing the conditions of possibility of this event); the second is called the *access map* of event e (intuitively giving for each agent $a \in \mathscr{A}$ the list $e(a) \subseteq \mathscr{A}$ of agents or locations whose information is being accessed by agent a during this event); while the third is the *postcondition map* of event e, assigning to each variable $v \in V$ the term $\sigma(v) \in Var$ (and intuitively capturing the variable-changes induced by the event).

Informal meaning of our language Our variables are meant to express 'data' (which can be numerical, propositional, graphical, etc), and will take various values in various states (or else they can be undefined, in which case they will be assigned the 'improper' value $\uparrow$, for 'undefined'). The basic variables are either constants c (taking the same value in all states, and keeping their value unchanged) or local variables v_a, each 'owned' by some agent a. From these, we recursively form more complex variables by: *applying function* symbols $F(x_1,\ldots,x_n)$; *defining new terms by cases* ($x|_\varphi y$ is a variable taking the same value as x if φ is true, and otherwise taking the same value as y); forming *definite descriptions* x_A^φ to denote the hypothetical value that x would take according to group A if φ were true; and *applying the (substitution involved in) event e to any variable x:* $e(x)$ takes the value that x will acquire after event e

At the propositional side, we have: atoms $P\overline{x}$ describing relations between the values of a tuple $\overline{x} = (x_1,\ldots,x_n)$ of variables; Boolean operators; group (distributed) knowledge $K_A\varphi$; conditional common distributed knowledge $C^\theta_{\mathfrak{A}}\varphi$ (pf *varphi*, among the groups in a supergroup $\mathfrak{A}$, conditional on some hypothesis θ); and dynamic modalities $[e]\varphi$, saying as usually that φ will hold after event e.

Finally, the expressions $e := !\Phi/\sigma$ for semi-public data-exchange events are of a particular simple kind. Our events have no epistemic structure (-no epistemic relations as in in standard Dynamic Epistemic Logic), since the structure of a 'semi-public' event is fully transparent: *what is happening* (-the presuppositions of this action, who gains access to whose database and which variables are changed and how) *is common knowledge*; although *the content of this happening* (-the actual data that are being accessed, the current values of variables, as well as the next values acquired after the event) *may remain private*. So event expressions are simple pairs of a set Φ of precondition formulas and a substitution σ. The set Φ comprises all presuppositions or 'preconditions' of event e: the propositions that must be true in order for the event to happen. The precondition formula pre_e collects them into one proposition, capturing in a sense the propositional information carried by event e. The substitution σ captures the event's *effect*: for any agent a, $e(a) := \sigma(a)$ is the list of all the sources accessed by a during this event; while for any local variable v, $post_e(v) := \sigma(v)$ denotes the value that v acquires after the event.

Extending Access to Groups and Supergroups For any event $e \in \mathscr{E}$, we can extend its access map from single agents to groups A and supergroups $\mathfrak{A}$, by putting:

$$e(A) := \bigcup_{a\in A} e(a), \qquad e[\mathfrak{A}] := \{e(A) : A \in \mathfrak{A}\}$$

More Abbreviations. We define as usual the other Boolean connectives. We also use the abbreviations, for variables $x \in Var$ and tuples $\overline{x} = (x_1,\ldots,x_n) \in Var^n$:

$$\top := (\uparrow{=}\uparrow), \quad \bot := \neg\top, \quad ?_\varphi := 1|_\varphi 0, \quad x{\uparrow} := (x{=}\uparrow), \quad x{\downarrow} := \neg x{\uparrow}, \quad \overline{x}{\downarrow} := \bigwedge_{i=1}^{n} x_i{\downarrow},$$

$$K_A^\theta\varphi := K_A(\theta\to\varphi), \quad K_A^\theta x := K_A^\theta(x{=}x_A^\theta), \quad K_A^\theta\overline{x} := \bigwedge_{i=1}^{n} K_A^\theta x_i, \quad K_A x := K_A^\top x, \quad K_A\overline{x} := K_A^\top\overline{x},$$

$$\langle K_A\rangle\varphi := \neg K_A\neg\varphi, \quad \langle K_A^\theta\rangle\varphi := \neg K_A^\theta\neg\varphi, \quad \langle C_{\mathfrak{A}}^\theta\rangle\varphi := \neg C_{\mathfrak{A}}^\theta\neg\varphi, \quad \langle e\rangle\varphi := \neg[e]\neg\varphi,$$

$$C_A^\theta\varphi := C_{\{\{a\}:a\in A\}}^\theta\varphi, \quad C_A\varphi := C_A^\top\varphi, \quad C_{\mathfrak{A}}x := C_{\mathfrak{A}}(\bigwedge_{A\in\mathfrak{A}} K_A x), \quad C_A x := C_{\{\{a\}:a\in A\}}x = C_A(\bigwedge_{a\in A} K_a x).$$

So *'knowledge what' is definable in our logic, in all its forms* (individual, common, distributed and common distributed knowledge of the value of a variable), via the above abbreviations defining $K_a x$, $K_A x$, $C_A x$ and $C_{\mathfrak{A}} x$. Also, note that, for the empty tuple $\lambda = ()$, we have $K_A^\theta\lambda = \bigwedge\emptyset = \top$ by definition; while the Boolean variable $?_\varphi$ takes value 1 if φ is true, and value 0 otherwise.

Abbreviations for events Whenever some component of an event is 'trivial', we may skip it. E.g. if the set of preconditions $\Phi = \emptyset$ in $e = !\Phi/\sigma$ (so $pre_e = \top$), we can write $e = !\sigma$; if in addition the access map is trivial, i.e. we have $\sigma(a) = \{a\}$ for all agents $a \in \mathscr{A}$, then we can just write the (non-trivial) post-conditions (only for those variables v for which $\sigma(v)$ is not the same term as v): e.g. writing $!(v := x)$ for a semi-public substitution event which only affects variable x; and dually, if both $\Phi = \emptyset$ and the post-conditions are trivial (leaving all variables the same), we can write just the (non-trivial part of the) access map, e.g. writing $!(a : b)$ for a semi-public event in which the only thing happening is that agent a gains access to agent b's knowledge base (so $\sigma(a) = \{a,b\}$ and $\sigma(i) = \{i\}$ for all other agents $i \neq a$).

Extended location For any *static expression* (i.e., term or formula $\alpha \in Var \cup Fml$ in the static language LKV), the set $\mathscr{A}(\alpha)$ of *agents involved in* α can be thought of as the "*extended location*" of that expression: its value (or truth value) is distributed among the agents in $\mathscr{A}(\alpha)$. Formally, we put:

$$\mathscr{A}(v_a) = \{a\}, \quad \mathscr{A}(c) = \emptyset, \quad \mathscr{A}(\varphi\wedge\psi) = \mathscr{A}(\varphi)\cup\mathscr{A}(\psi), \quad \mathscr{A}(\neg\varphi) = \mathscr{A}(\varphi),$$

$$\mathscr{A}(x_A^\varphi) = \mathscr{A}(K_A\varphi) = A, \quad \mathscr{A}(F(\overline{x})) = \mathscr{A}(P\overline{x}) = \bigcup_{i=1,n} \mathscr{A}(x_i) \quad \text{for } \overline{x} = (x_i)_{1\le i\le n}.$$

Semantics: epistemic data models We assume given a *first-order model* $\mathbf{D} = (D, I)$, consisting of: a *domain of values (objects)* D, that includes an object $\uparrow_{\mathbf{D}} \in D$ (the *'undefined' value*); and an *interpretation function* I, mapping each constant symbol $c \in C$ into some value $c_{\mathbf{D}} := I(c) \in D$, each functional symbol f of arity n into a function $I(f) : D^n \to D$, and each relational symbol P of arity n into a set $I(P) \subseteq D^n$ of n-tuples of values in D, with *the equality symbol $=$ being interpreted as the identity relation $I(=) = \{(d,d) : d \in D\}$ on the set D.*

An *epistemic data model over* $\mathbf{D}$ is a tuple $\mathbf{M} = (S, \sim, \bullet(\bullet))$, where: S is a set of *states* (or 'possible worlds'), typically denoted by $s, w, \ldots$; $\sim : \mathscr{A} \to \mathscr{P}(S\times S)$ maps each agent $a \in \mathscr{A}$ to some *equivalence relation* $\sim_a \subseteq W \times W$, called *agent a's 'indistinguishability' relation*; and $\bullet(\bullet) : S \times V \to D$ maps states s and basic variables $v \in V$ into *values* $s(v) \in D$; where we require that *agents know their own data*, i.e.:

$$s \sim_a w \text{ implies } s(v_a) = w(v_a).$$

Group indistinguishability For every group $A \subseteq \mathscr{A}$, we introduce *group indistinguishability relations* $\sim_A := \bigcap_{a\in A} \sim_a$ by *taking intersections*, i.e.: $s \sim_A w$ iff $s \sim_a w$ for all $a \in A$.

Satisfaction, Term-Interpretation, Update We simultaneously define three semantic notions: (1) a *satisfaction relation* $s \models_{\mathbf{M}} \varphi$ between states $s \in S$ (in a data model $\mathbf{M}$) and formulas $\varphi \in Fml$; (2) an *extended value function* mapping $(state, variable)$-pairs $(s,x) \in S \times Var$ into values $s(x)_{\mathbf{M}} \in D$; and (3) an *update function* mapping data models $\mathbf{M} = (S, \sim, \bullet(\bullet))$ and events e into *updated models* $e(\mathbf{M}) =$

$(S^e, \sim^e, \bullet(\bullet)^e)$ (over the same first-order model $\mathbf{D}$). Whenever the model is understood, we skip the subscript $\mathbf{M}$, writing simply $s \models \varphi$ and $s(x)$. The definition is by mutual recursion:

$$
\begin{array}{lll}
s \models_{\mathbf{M}} Px_1 \dots x_n & \text{iff} & (s(x_1)_{\mathbf{M}}, \dots, s(x_n)_{\mathbf{M}}) \in I(P) \\
s \models_{\mathbf{M}} \neg\varphi & \text{iff} & s \not\models_{\mathbf{M}} \varphi \\
s \models_{\mathbf{M}} \varphi \wedge \psi & \text{iff} & w \models_{\mathbf{M}} \varphi \text{ and } w \models_{\mathbf{M}} \psi \\
s \models_{\mathbf{M}} K_A\varphi & \text{iff} & w \models_{\mathbf{M}} \varphi \text{ for all } w \sim_A s \\
s \models_{\mathbf{M}} C^{\theta}_{\mathfrak{A}}\varphi & \text{iff} & w \models_{\mathbf{M}} \varphi \text{ for every chain } s = s_0 \sim_{A^1} s_1 \sim_{A^2} \cdots \sim_{A^n} s_n = w \\
 & & \text{(of any finite length } n \geq 0) \text{ with } A_k \in \mathfrak{A} \text{ and } s_k \models_{\mathbf{M}} \theta \text{ for all } k \geq 1 \\
s \models_{\mathbf{M}} [e]\varphi & \text{iff} & s \models_{\mathbf{M}} pre_e \text{ implies } s \models_{e(\mathbf{M})} \varphi
\end{array}
$$

$$
\begin{array}{lll}
s(c)_{\mathbf{M}} & = & c_{\mathbf{D}} \text{ as given by the interpretation } I(c) \\
s(v)_{\mathbf{M}} & = & s(v) \text{ as given in the model } \mathbf{M} \\
s(x|_{\varphi}y)_{\mathbf{M}} & = & s(x)_{\mathbf{M}} \quad \text{if } s \models_{\mathbf{M}} \varphi, \text{ and} \\
s(x|_{\varphi}y)_{\mathbf{M}} & = & s(y)_{\mathbf{M}} \quad \text{if } s \not\models_{\mathbf{M}} \varphi \\
s(F(x_1, \dots, x_n))_{\mathbf{M}} & = & (I(F))(s(x_1)_{\mathbf{M}}, \dots, s(x_n)_{\mathbf{M}}) \\
s(x^{\varphi}_A)_{\mathbf{M}} & = & \text{the unique } d \in D \text{ s.t. } w(x)_{\mathbf{M}} = d \text{ for all } w \sim_A s \text{ with } w \models_{\mathbf{M}} \varphi, \\
 & & \text{if such a unique } d \text{ exists, and} \\
s(x^{\varphi}_A)_{\mathbf{M}} & = & \uparrow_{\mathbf{D}} \quad \text{otherwise.} \\
s(e(x))_{\mathbf{M}} & = & s(x)_{e(\mathbf{M})} \quad \text{if } s \models_{\mathbf{M}} pre_e, \text{ and} \\
s(e(x))_{\mathbf{M}} & = & \uparrow_{\mathbf{D}} \quad \text{otherwise.}
\end{array}
$$

$$
\begin{array}{lll}
\mathbf{M} = (S, \sim, \bullet(\bullet)) & \mapsto & e(\mathbf{M}) = (S^e, \sim^e, \bullet(\bullet)_{e(\mathbf{M})}), \quad \text{where} \\
S^e & = & \|pre_e\|_{\mathbf{M}} = \{s \in S : s \models_{\mathbf{M}} pre_e\} \\
s \sim^e_a w & \text{iff} & s \sim_{e(a)} w \quad (\text{for } s, w \in S^e) \\
s(v)_{e(\mathbf{M})} & = & s(post_e(v))_{\mathbf{M}} \quad (\text{for } s \in S^e).
\end{array}
$$

Proposition 1. *Let s, w be states s.t. $s \sim_A w$ for some group $A \subseteq \mathscr{A}$, let φ be some formula s.t. $\mathscr{A}(\varphi) \subseteq A$, and let $x \in Var$ be a term s.t. $\mathscr{A}(x) \subseteq A$. Then we have:*

1. *$s \models \varphi$ iff $w \models \varphi$;*
2. *$s(x) = w(x)$.*

3 An Example: The Numbers' Game Puzzle

There are two brothers, Alex and Bob, their sister Daniela, as well as an evil intruder Eve. Each of Alex, Bob and Daniela is given a sheet of paper on which there is a secret natural number $n_a, n_b, n_d \in \{0, 1, 2, \dots\}$ (Alex receives n_a, etc), so that each of them can see only his/her number. It is common knowledge that Daniela's number is the absolute difference of the other two numbers ($n_d = |n_a - n_b|$).

The following scenario ensues (in which all announcements are public and it is common knowledge that nobody lies):

Daniela shows to Alex her sheet of paper (containing the number n_d). It is common knowledge that this is happening, but nobody else can see the number.

Then Daniela asks Alex "Do you know Bob's number?", and Alex answers "I don't know".

Next, Daniela also shows to Bob her sheet of paper (with her number n_d). Again, it is common knowledge that this is happening, but nobody else can see the number.

Then Daniela asks Bob "Do you know Alex's number?" and Bob answers "I don't know".

Finally, Alex says "*Now* I know Bob's number".

Exercise:

1. Show that, at the end of the above scenario, it is common knowledge among all agents (including Eve) that $0 < n_d \leq n_a < 2 \cdot n_d \leq n_b$.
2. Show that, at the end of the above scenario, all three numbers are common knowledge among Alice, Bob and Daniela, but it is also common knowledge that Eve doesn't know any of the numbers.
3. Suppose that, after the above scenario, Daniela announces that her number is $n_d = 1$. Show that, after that, it is common knowledge (among *all* agents) that the numbers are $n_a = 1$, $n_b = 2$, $n_c = 1$.

Modelling the Numbers' Game Example. We take $D = \mathbb{N}$ (the set of natural numbers) and work with four agents (a,b,d,e) and three basic local variables n_a, n_b, n_d, each in possession of the corresponding agent. We use addition $x+y$ as our only function symbol, and equality $x = y$, inequality $x \leq y$ and strict inequality $x < y$ as our predicate symbols (with their standard interpretations on $\mathbb{N}$. The set of states in the initial model $\mathbf{S}^0$ is $S^0 := \{s = (s_a, s_b, s_d) \in \mathbb{N} \times \mathbb{N} \times \mathbb{N} : s_a = s_b + s_d \text{ or } s_b = s_a + s_d\}$. The epistemic equivalence relations $\sim_i \subseteq S^0 \times S^0$ are given by putting $s \sim_i w$ iff $s_i = w_i$, for each $i \in \{a,b,d\}$, while $\sim_e = S^0 \times S^0$ is the universal relation on S^0. The value map is the obvious one: $s(n_i) := s_i$, for $i \in \{a,b,d\}$.

Our scenario starts with the semi-public data-sharing event $!(a : d)$ (in which Daniela shows her sheet to Alex): this changes our model to an updated model $\mathbf{S}^1$, obtained by resetting Alex's relation to the intersection $\sim'_a := \sim_a \cap \sim_d$ of his old relation with Daniela's (and leaving everything else the same). The next step is the announcement $!(\neg K_a n_b)$ (in which Alex publicly announces that he doesn't know Bob's number): this deletes all states having $s_d = 0$ or $s_a < s_d$ (since these states satisfy $K_a n_b$ in $\mathbf{S}^1$), leaving everything else the same. So in the resulting model $\mathbf{S}^2$, we are left only states $s = (s_a, s_b, s_d)$ with $0 < s_d \leq s_a$ (and $|s_a - s_b| = s_d$). In the next round, we have another semi-public data-sharing event $!(b : d)$, which resets Bob' relation to the intersection $\sim'_b := \sim_b \cap \sim_d$, and leaving everything else the same in the resulting model $\mathbf{S}^3$. After this, the public announcement $!(\neg K_b n_a)$ eliminates all states having $0 \leq s_b < 2 \cdot s_d$ (since these states satisfy $K_b n_a$ in $\mathbf{S}^3$). So in the updated model $\mathbf{S}^4$ we are left only with states with $0 < s_d \leq s_a$ and $2 \cdot s_d \leq s_b$. Finally, the public announcement $!(K_a n_b)$ eliminates all states with $2 \cdot s_d \leq s_a$ (since those satisfy $\neg K_a n_b$ in $\mathbf{S}^4$). Hence, the final model $\mathbf{S}^5$ consists only of states satisfying $0 < s_d \leq s_a < 2 \cdot s_d \leq s_b$ (and $|s_a - s_b| = s_d$). It is obvious that the formula $C_{\{a,b,d,e\}}(0 < n_d \leq n_a < 2 \cdot n_d \leq n_b)$ is true in all the states of $\mathbf{S}^5$, which takes care of item 1 in our Exercise.

For item 2, it is easy to see that the formula $C_{\{a,b,e\}}\overline{n}$ holds in all states of $\mathbf{S}^5$, where $\overline{n} = (n_a, n_b, n_e)$ is the tuple consisting of our three local variables. Finally, to deal with item 3, we perform on $\mathbf{S}^5$ another public announcement $!(n_d = 1)$, which deletes all states except for $s = (1,2,1)$.

4 Axiomatizations and Main Results

The proof system **DLKV** for the full dynamic logic *DLKV* is in Table 1, where we used the abbreviations $K_A^\theta \varphi := K_A(\theta \to \varphi)$, $K_a \varphi := K_{\{a\}} \varphi$, $K_A^\varphi x := K_A^\varphi (x = x_A^\varphi)$, and $K_A x := K_A^\top x$. The proof system **LKV** is obtained by taking only the axioms and rules in groups (I)-(V) (restricted to the static fragment *LKV*).

Theorem 1. *The proof system* **LKV** *is sound and complete for the static fragment LKV, and its logic LKV is decidable.*

Using this result, we can establish completeness and decidability for the full dynamic logic *DLKV*:

Theorem 2. *The proof system* **DLKV** *in Table 1 is sound and complete for the dynamic logic DLKV. Moreover, the logic DLKV is provably co-expressive with the static logic LKV, and thus it is decidable.*

Soundness of **LKV** and **DLKV** are easy verifications, so we skip the details. The *proofs of completenes and decidability* are intricate, and they form the content of the next two sections.

(I) Axioms and rules of classical propositional logic	
(II) Axioms for equality:	
(Reflexivity)	$x = x$
(Indiscernability)	$\overline{x} = \overline{y} \to (P\overline{xz} \leftrightarrow P\overline{yz})$
(Functionality)	$\overline{x} = \overline{y} \to F(\overline{x}) = F(\overline{y})$
(Definition by Cases)	$(\varphi \to x\|_\varphi y = x) \wedge (\neg\varphi \to x\|_\varphi y = y)$
(III) Axioms and rules for (distributed) knowledge:	
(Necessitation)	From φ, infer $K_A\varphi$
(Distribution)	$K_A(\varphi \to \psi) \to (K_A\varphi \to K_A\psi)$
(Veracity)	$K_A\varphi \to \varphi$
(Positive Introspection)	$K_A\varphi \to K_AK_A\varphi$
(Negative Introspection)	$\neg K_A\varphi \to K_A\neg K_A\varphi$
(Group-Monotonicity)	$K_A\varphi \to K_B\varphi$, whenever $A \subseteq B$
(IV) Axioms and rules for conditional common distributed knowledge:	
(C-Necessitation)	From φ, infer $C^\theta_{\mathfrak{A}}\varphi$
(C-Distribution)	$C^\theta_{\mathfrak{A}}(\varphi \to \psi) \to (C^\theta_{\mathfrak{A}}\varphi \to C^\theta_{\mathfrak{A}}\psi)$
(Fixed Point)	$C^\theta_{\mathfrak{A}}\varphi \to \left(\varphi \wedge \bigwedge_{A\in\mathfrak{A}} K^\theta_A C^\theta_{\mathfrak{A}}\varphi\right)$
(Induction)	$C^\theta_{\mathfrak{A}}(\varphi \to \bigwedge_{A\in\mathfrak{A}} K^\theta_A\varphi) \to (\varphi \to C^\theta_{\mathfrak{A}}\varphi)$
(V) Axioms for conditional knowledge of values	
(Non-vacuous Knowledge of Values)	$x^\varphi_A{\downarrow} \to \left(K^\varphi_A x \wedge \langle K_A\rangle\varphi\right)$
(Knowledge of Constants & Local Variables)	$K_a c \wedge K_a(v_a)$
(Knowledge of Hypothetical Values)	$K_A x^\varphi_A$
(Knowledge of Predicates)	$K_A\overline{x} \to (P\overline{x} \to K_AP\overline{x})$
(Knowledge of Functions)	$K^\varphi_A\overline{x} \to K^\varphi_A F(\overline{x})$
(Known Equality)	$K^\theta_A(x = y) \to (K^\theta_A x \to K^\theta_A y)$
(Anti-Monotonicity)	$K_A(\varphi \to \theta) \to (K^\theta_A x \to K^\varphi_A x)$
(VI) Dynamic reduction axioms for propositional formulas	
($[e]$-Necessitation)	From φ, infer $[e]\varphi$
($[e]$-Distribution)	$[e](\varphi \to \psi) \to ([e]\varphi \to [e]\psi)$
(Atomic Change)	$[e]Px_1 \dots x_n \leftrightarrow (pre_e \to Pe(x_1) \dots e(x_n))$
(Partial Functionality)	$[e]\neg\varphi \leftrightarrow (pre_e \to \neg[e]\varphi)$
(Knowledge Update)	$[e]K_A\varphi \leftrightarrow \left(pre_e \to K_{e(A)}[e]\varphi\right)$
($C_{\mathfrak{A}}$-Update)	$[e]C^\theta_{\mathfrak{A}}\varphi \leftrightarrow \left(pre_e \to C^{\langle e\rangle\theta}_{e[\mathfrak{A}]}[e]\varphi\right)$
(VII) Dynamic reduction axioms for terms	
(Survival of Value)	$e(x){\downarrow} \to pre_e$
(Preservation of Constants)	$pre_e \to e(c) = c$
(Change of Basic Values)	$pre_e \to e(v) = post_e(v)$
(Change of Disjunctive Terms)	$pre_e \to e(x\|_\varphi y) = e(x)\|_{\langle e\rangle\varphi} e(y)$
(Change of Functional Terms)	$pre_e \to e(F(x_1,\dots,x_n)) = F(e(x_1),\dots,e(x_n))$
(Change of Hypothetical Values)	$pre_e \to e(x^\varphi_A) = e(x)^{\langle e\rangle\varphi}_{e(A)}$

Table 1: The Proof System **DLKV**. The Proof System **LKV** consists only of the axioms and rules in groups (I)-(V) (restricted to the static fragment *LKV*).

To illustrate the power of our axiomatic system, we note a number of interesting theorems.

Proposition 2. *The following are provable in the system* **LKV***:*

- *(Strong Introspection)* $\varphi \to K_A\varphi$, *whenever* $\mathscr{A}(\varphi) \subseteq A$
- *(Term Introspection)* $K_A x$, *whenever* $\mathscr{A}(x) \subseteq A$
- *(Explicit Value Introspection)* $\langle K_A\rangle\varphi \to x_A^\varphi = x$, *whenever* $\mathscr{A}(x) \subseteq A$
- *(Conditional Knowledge)* $K_A^\varphi\psi \leftrightarrow K_A(\varphi \to \psi)$
- *(Non-Vacuous Knowledge)* $\langle K_A\rangle\varphi \to \left(K_A^\varphi x \leftrightarrow x_A^\varphi{\downarrow}\right)$
- *(True Conditional Value)* $(x_A^\varphi{\downarrow} \wedge \varphi) \to x_A^\varphi = x)$
- $K_A(\varphi \to x = y) \to x_A^\varphi = y_A^\varphi$
- $(K_A(\varphi \to x = y) \wedge \varphi \wedge x{\downarrow}) \to x_A^\varphi{\downarrow}$, *whenever* $\mathscr{A}(y) \subseteq A$

5 Completeness and Decidability for the Static Logic *LKV*

Throughout this section, *we fix a formula* $\varphi_0 \in Fml$. We'll prove Theorem 1 by the method of *quasi-models*. To obtain an appropriate analogue of Fischer-Ladner closure, we need an auxiliary notion:

Restricted Vocabulary and Restricted Terms For any finite set of formulas $\Sigma \subseteq Fml$, the Σ-*restricted set of terms* Var_Σ and the Σ-*restricted vocabulary* $\mathscr{V}_\Sigma :== (V_\Sigma, \mathscr{A}_\Sigma, ag_\Sigma, Pred_\Sigma, Funct_\Sigma, ar_\Sigma)$ are formed as follows: Var_Σ is *the set of all subterms (of any term) occurring in formulas of* Σ; $V_\Sigma := V \cap Var_\Sigma$ is *the set of all basic variables occurring (as subterms of any term) in formulas of* Σ; $\mathscr{A}_\Sigma$ is the set of agents occurring (inside terms or modalities in formulas) in Σ or as locations $ag(v)$ of basic variables $v \in V_\Sigma$; Pre_Σ is the set of predicates (occurring in formulas) in Σ; $Funct_\Sigma$ is the set of function symbols occurring in (terms of) Var_Σ; ag_Σ is the restriction of ag to V_Σ; and ar_Σ is the restriction of ar to $Pred_\Sigma \cup Funct_\Sigma$.

Clearly, if Σ is finite, then Var_Σ and all the sets in $\mathscr{V}_\Sigma$ are finite. But note that the Σ-restricted set of terms Var_Σ is only *a finite subset* of the (typically infinite) set $Var(\mathscr{V}_\Sigma)$ of terms of the language $LKV(\mathscr{V}_\Sigma)$.

Closure Given our fixed formula φ_0, *the closure of* φ_0 is a finite set $\Sigma = \Sigma(\varphi_0) \subseteq Fml$, defined as the smallest set of formulas satisfying the following properties: $\varphi_0 \in \Sigma$; Σ is closed under subformulas and single negations $\sim\varphi$; if $(K_A\varphi) \in \Sigma$ and $B \subseteq \mathscr{A}_\Sigma$, then $(K_B\varphi) \in \Sigma$; if $K_A(\varphi \to \psi) \in \Sigma$, then $(K_A\varphi) \in \Sigma$ (and so also $(K_B\varphi) \in \Sigma$; if $(x|_\varphi y) \in Var_\Sigma$, then $\varphi \in \Sigma$; if $(x_A^\varphi), y, z \in Var_\Sigma$, then $K_A(\varphi \to y = z) \in \Sigma$ (and so also $(K_A^\varphi x), \langle K_A\rangle(\varphi \wedge y \neq z), \langle K_A\rangle\varphi, \varphi \in \Sigma$); if $\overline{x} = (x_1, \ldots, x_n)$ has all $x_1, \ldots, x_n \in Var_\Sigma$ and $P \in Pred_\Sigma$, then $P\overline{x}, (\overline{x}{\downarrow}) \in \Sigma$ (-and so in particular, if $x, y \in Var_\Sigma$ then $(x = y), x{\downarrow}, x{\uparrow} \in \Sigma$ and $\uparrow \in Var_\Sigma$).

Types Let Σ be the closure of φ_0. A Σ-*type* is a subset of Σ with the following properties:

1. for every $\varphi \in \Sigma$: $(\sim\varphi) \in \Delta$ iff $\varphi \notin \Delta$;
2. for every $(\varphi \wedge \psi) \in \Sigma$: $(\varphi \wedge \psi) \in \Delta$ iff $\varphi \in \Delta$ and $\psi \in \Delta$;
3. for every $x \in Var_\Sigma$: $(x = x) \in \Delta$;
4. for every $P\overline{xz} \in \Sigma$: if $(\overline{x} = \overline{y}), P\overline{xz} \in \Delta$ then $P\overline{yz} \in \Delta$;[7]
5. for $F(\overline{x}), F(\overline{y}) \in Var_\Sigma$: if $(\overline{x} = \overline{y}) \in \Delta$, then $(F(\overline{x}) = F(\overline{y})) \in \Delta$;
6. for every $(x|_\varphi y) \in \Sigma$: $\varphi \in \Delta$ implies $(x|_\varphi y) \in \Delta$; and $\varphi \notin \Delta$ implies $(x|_\varphi y = y) \in \Delta$;
7. if $(K_A\varphi) \in \Delta$ then $\varphi \in \Delta$;

[7] This, together with clause 3, has the consequence that Δ-*equality is an equivalence relation on variables in* Var_Σ: $(x = y) \in \Delta$ implies $(y = x) \in \Delta$; and $(x = x'), (x' = x'') \in \Delta$ imply $(x = x'') \in \Delta$.

8. if $(x^{\varphi}_{A}\downarrow) \in \Delta$, then $K^{\varphi}_{A}x, \langle K_A\rangle\varphi \in \Delta$;
9. if $\varphi, (x^{\varphi}_{A}\downarrow) \in \Delta$, then $(x = x^{\varphi}_{A}) \in \Delta$;
10. if $K_A(\varphi \to x = y) \in \Delta$ then $(x^{\varphi}_{X} = y^{\varphi}_{A}) \in \Delta$;
11. $(c^{\varphi}_{A} = c) \in \Delta$ or $(c^{\varphi}_{A} = \uparrow) \in \Delta$, for all $c \in C \cap Var_{\Sigma}$;
12. for every $x^{\varphi}_{A} \in Var_{\Sigma}$ and every $y \in Var_{\Sigma}$ s.t. $\mathscr{A}(y) \subseteq A$: if $(x^{\varphi}_{A}\uparrow), \varphi, x\downarrow \in \Delta$, then $\langle K_A\rangle(\varphi \wedge x \neq y) \in \Delta$;
13. if $(C^{\theta}_{\mathfrak{A}}\varphi) \in \Delta$ and $A \in \mathfrak{A}$, then $\varphi, (K^{\theta}_{A}C^{\theta}_{\mathfrak{A}}\varphi) \in \Delta$.

Accessibility relations on types For types Δ, Δ' and group $A \subseteq \mathscr{A}_{\Sigma}$, we put:

$$\Delta \sim_A \Delta' \text{ iff } \varphi \in \Delta \Leftrightarrow \varphi \in \Delta' \text{ for all } \varphi \in \Sigma \text{ s.t. } \mathscr{A}(\varphi) \subseteq A.$$

Proposition 3. *The relations $\sim_A$ are equivalence relations on types, satisfying Group Monotonicity:*

$$\Delta \sim_A \Delta' \text{ and } B \subseteq A \text{ imply } \Delta \sim_B \Delta'.$$

Proof. This follows directly from the definition of relations $\sim_A$ on types. □

Proposition 4. *If $\Delta \sim_A \Delta'$ and $(K_A\varphi) \in \Delta$, then $\varphi \in \Delta'$.*

Proof. Since $\mathscr{A}(K_A\varphi) = A$ and $\Delta \sim_A \Delta'$, we use the definition of $\sim_A$ on types and the fact that $(K_A\varphi) \in \Delta$ to infer that $(K_A\varphi) \in \Delta'$. This together with condition 7 on types, gives us the desired conclusion. □

Hat notation. For every type Δ, we put $\widehat{\Delta} := \bigwedge\Delta$ for the *conjunction of all formulas in* Δ.

Proposition 5. *Given types Δ and Λ, if $\widehat{\Delta} \wedge \langle K_A\rangle\widehat{\Lambda}$ is consistent, then $\Delta \sim_A \Lambda$.*

Proof. Let Δ and Λ be types as above, and suppose towards a contradiction that we have $\Delta \not\sim_A \Lambda$. Then there must exist $\varphi \in \Sigma$ s.t. $\mathscr{A}(\varphi) \subseteq A$ and $\varphi \in \Delta$, but $(\sim\varphi) \in \Lambda$. From this together with the assumption that $\widehat{\Delta} \wedge \langle K_A\rangle\widehat{\Lambda}$ is consistent, we infer that $\varphi \wedge \langle K_A\rangle\neg\varphi$ is consistent, and thus $\varphi \wedge \neg K_A\varphi$ is consistent.

But this contradicts the fact that $\vdash \varphi \to K_A\varphi$ is an **LDK**-theorem for formulas φ with $\mathscr{A}(\varphi) \subseteq A$. □

Quasi-Models A *quasi-model for* φ_0 is a set S of types over $\Sigma = \Sigma(\varphi_0)$, with the following properties:

(*) $\varphi_0 \in \Delta_0$ for some type $\Delta_0 \in S$;

(**) if $\langle K_A\rangle\varphi \in \Delta \in S$, then there is some $\Delta' \in S$ with $\Delta \sim_A \Delta'$ and $\varphi \in \Delta'$;

(***) if $\langle C^{\theta}_{\mathfrak{A}}\rangle\varphi \in \Delta \in S$, then there is some finite chain $\Delta = \Delta_0 \sim_{A^1} \Delta_1 \sim_{A^2} \cdots \sim_{A^n} \Delta_n$, with $0 \leq n \leq |S| - 1$, $A^1, \ldots, A^n \in \mathfrak{A}$, $\theta \in \Delta_1, \ldots, \Delta_n \in S$ and $\varphi \in \Delta_n$.

The Σ-canonical quasi-model A Σ-*theory* is any maximally consistent subset of Σ. We denote by S_{Σ} the *set of all Σ-theories*. We will *show that S_{Σ} is a (finite) "canonical" quasi-model* for Σ.

Lemma 1. *Every Σ-theory is a Σ-type.*

Proof. This is easy to check, using the axioms of **LKV** and Proposition 2. □

Lemma 2. *For every $\Delta \in S_{\Sigma}$, if $\langle K_A\rangle\varphi \in \Delta$ then there is some $\Delta' \in S_{\Sigma}$ with $\Delta \sim_A \Delta'$ and $\varphi \in \Delta'$.*

Proof. Put $\Delta_A := \{\theta : \theta \in \Delta \text{ s.t. } \mathscr{A}(\theta) \subseteq A\} \cup \{\sim\theta : \theta \in (\Sigma - \Delta) \text{ s.t. } \mathscr{A}(\theta) \subseteq A\}$.

Claim: $\Delta_A \cup \{\varphi\}$ is consistent wrt the system **LKV**.

Proof of Claim: Suppose not. Then we have $\vdash \widehat{\Delta_A} \to \sim\varphi$. Applying Necessitation and Distribution, we obtain $\vdash K_A\widehat{D_A} \to K_A \sim \varphi$. On the other hand, by inspecting the structure of Δ_A, it is easy to see that $\mathscr{A}(\Delta_A) \subseteq A$, so by Strong Introspection (Proposition 29) we have $\vdash \widehat{\Delta_A} \to K_A\widehat{\Delta_A}$. Putting these together, we get $\vdash \widehat{\Delta_A} \to K_A \sim \varphi$. Since $\Delta_A \subseteq \Delta$ and Δ is closed under Σ-consequences, we have $(K_A \sim \varphi) \in \Delta$. But this contradicts the assumption that $\langle K_A\rangle\varphi \in \Delta$ (given the consistency of Δ).

Using our Claim and the standard Lindenbaum Lemma, we get that $\Delta_A \cup \{\varphi\}$ has a Σ-maximally consistent extension $\Delta' \in S'_{\Sigma}$. So we have $\varphi \in \Delta'$ and $\Delta_A \subseteq \Delta'$, which implies that $\Delta \sim_A \Delta'$. □

Lemma 3. *For every Σ-theory $\Delta \in S_\Sigma$, if $\langle C^\theta_{\mathfrak{A}}\rangle\varphi \in \Delta$, then there is some finite chain of Σ-theories $\Delta = \Delta_0 \sim_{A^1} \Delta_1 \sim_{A^2} \cdots \sim_{A^n} \Delta_n$, with $0 \leq n \leq |S_\Sigma| - 1$, $A^1, \ldots, A^n \in \mathfrak{A}$, $\theta \in \Delta_1, \ldots, \Delta_n \in S_\Sigma$ and $\varphi \in \Delta_n$.*

Proof. Put $\Delta^\Delta_{\mathfrak{A}} := \{\Gamma \in S_\Sigma : \Delta = \Gamma_0 \sim_{A^1} \Gamma_1 \cdots \sim_{A^n} \Gamma_n = \Gamma \text{ s.t.} n \geq 0, \theta \in \Gamma_k \in S_\Sigma \text{ and } A^k \in \mathfrak{A} \text{ for all } k \geq 1\}$. We need to prove that there exists $\Gamma \in \Delta^\Delta_{\mathfrak{A}}$ s.t. $\varphi \in \Gamma$.[8] Suppose not. Then $(\sim \varphi) \in \Gamma$ for all $\Gamma \in \Delta^\Delta_{\mathfrak{A}}$. Take the formula $\eta := \bigvee\{\widehat{\Gamma} : \Gamma \in \Delta^\Delta_{\mathfrak{A}}\}$. So we have $\vdash \eta \rightarrow \sim\varphi$, thus also $\vdash C^\theta_{\mathfrak{A}}\eta \rightarrow C^\theta_{\mathfrak{A}} \sim \varphi$.

Claim: $\vdash \eta \rightarrow C^\theta_{\mathfrak{A}}\eta$.

Proof of Claim: Using the Induction axiom and $C^\theta_{\mathfrak{A}}$-Necessitation, we see that is enough to show that $\vdash \eta \rightarrow K^\theta_A\eta$ for all $A \in \mathfrak{A}$. Suppose not: then we have that $\eta \wedge \langle K^\theta_A\rangle\neg\eta$ is consistent, for some $A \in \mathfrak{A}$. Unfolding the structure of η and using the (easily verifiable) fact that $\vdash \bigvee\{\widehat{\Gamma} : \Gamma \in S_\Sigma\}$ is a theorem, we see that there must exist Σ-theories $\Gamma \in \Delta^\theta_{\mathfrak{A}}$ and $\Gamma' \in (S_\Sigma - \Delta^\theta_{\mathfrak{A}})$ s.t. $\widehat{\Gamma} \wedge \langle K^\theta_A\rangle\widehat{\Gamma'}$ is consistent, i.e. $\widehat{\Gamma} \wedge \langle K_A\langle(\theta \wedge \widehat{\Gamma'})$ is consistent. This means that on the one hand $\theta \wedge \widehat{\Gamma'}$ is consistent, hence $\theta \in \Gamma'$ (since $\theta \in \Sigma$ and Γ' is a maximally consistent subset of Σ); and on the other hand, $\widehat{\Gamma} \wedge \langle K_A\rangle\widehat{\Gamma'}$ is consistent, hence by Proposition 5 (and the fact that Γ and Γ' are types by Lemma 1) we have $\Gamma \sim_A \Gamma'$. But given the way we defined the set $\Delta^\theta_{\mathfrak{A}}$, we see that from $\Gamma \in \Delta^\theta_{\mathfrak{A}}$, $A \in \mathfrak{A}$, $\Gamma \sim_A \Gamma'$ and $\theta \in \Gamma'$ we can infer that $\Gamma' \in \Delta^\theta_{\mathfrak{A}}$, which contradicts the fact that $\Gamma' \in (S_\Sigma - \Delta^\theta_{\mathfrak{A}})$.

Using now our Claim and the fact that $\vdash C^\theta_{\mathfrak{A}}\eta \rightarrow C^\theta_{\mathfrak{A}} \sim \varphi$, we obtain that $\vdash \eta \rightarrow C^\theta_{\mathfrak{A}} \sim \varphi$. But we also have $\vdash \widehat{\Delta} \rightarrow \eta$ (by the structure of η and the fact that $\Delta \in \Delta^\theta_{\mathfrak{A}}$ by definition), hence we infer $\vdash \widehat{\Delta} \rightarrow C^\theta_{\mathfrak{A}} \sim \varphi$. But this contradicts the assumption that $\langle C^\theta_{\mathfrak{A}}\rangle\varphi \in \Delta$ (given the consistency of Δ). □

Proposition 6. *If $\varphi_0 \in \Sigma$ is consistent, then there exists a quasi-model for φ_0.*

Proof. Take the set S_Σ of all Σ-theories. By the Lindenbaum Lemma, there exists a maximally consistent subset $\Delta_0 \in S_\Sigma$, such that $\varphi_0 \in \Delta_0$. By Lemmas 1, 2 and 3, S_Σ is a quasi-model for φ_0. □

Putting this together with the *soundness* of the proof system **LKV**, we obtain the following:

Corollary. *If φ_0 is satisfiable then there exists a quasi-model for φ_0.*

The hard part is to prove the *converse* of this:

Proposition 7. *If S is a quasi-model for φ_0, then φ_0 is satisfiable.*

Note that this would immediately give us the completeness and decidability of **LKV**:

Proof of Theorem 1: By Proposition 7, Proposition 6 and the soundness of the system **LKV**, the following three notions are equivalent: (a) φ_0 is consistent; (b) there exists a quasi-model for φ_0; (c) φ_0 is satisfiable. *Completeness* immediately follows. For decidability: to decide whether φ_0 is satisfiable, take the closure $\Sigma = \Sigma(\varphi_0)$ of φ_0, and check whether φ belongs to any type in a quasi-model over Σ.[9]

The rest of this section is dedicated to the proof of Proposition 7.

Unravelling: the tree of histories Let us fix a quasi-model S, a formula φ_0 and a type $\Delta_0 \in S$ with $\varphi_0 \in \Delta_0$. We will construct a model for φ, essentially based on an unravelling of S around Δ_0. A *history* is a finite sequence $h = (\Delta_0, A^1, \Delta_1, \ldots, A^n, \Delta_n)$ of any length $n \geq 0$, where $\Delta_1, \ldots, \Delta_n \in S$ are types and $A^1, \ldots, A^n \subseteq \mathscr{A}_\Sigma$ are groups, such that we have $\Delta_{i-1} \sim_{A^i} \Delta_i$ for all $i = 1, n$. We denote by $last(h) := \Delta_n$ the last state in history h, and by $\rightarrow_A$ the natural *one-step relation* on histories, given by putting: $h \rightarrow_A h'$

[8] This shows that there exists a finite $\mathfrak{A}$-chain of Σ-theories $\Delta = \Gamma_0, \Gamma_1, \ldots, \Gamma_n = \Gamma$, with $n \geq 0$, $\varphi \in \Gamma$ and with $\theta \in \Gamma_k$ for all $j \geq 1$. The fact that we can take $n \leq |S| - 1$ follows from the fact that there are no more than $|S_\Sigma|$ distinct such theories, so if the chain is longer we can just shorten it by cutting the repeating theories (together with the subchains connecting them).

[9] Note that there are only finitely many types over Σ, so only finitely many quasi-models, and in fact we can calculate an upper bound on their number, that is a computable function of the complexity of φ_0. Equally important, checking that a finite subset of Σ is a type, and that a finite set of types is a quasi-model, are both computable tasks.

iff $h' = (h, A, \Delta')$ (with $last(h) \sim_A \Delta' = last(h')$). The one-step relations structure H into a *tree rooted at* Δ_0 (with the immediate successor relation given by $h \to h'$ iff $h \to_A h'$ for some group A). In particular, we have *the tree property* : *every two nodes h, h' of the tree are connected by a unique non-redundant path* $h = h_0 \leftarrow_{A^1} h_1 \leftarrow_{A^2} \ldots \leftarrow_{A^i} h_i \to_{A^{i+1}} \ldots \to_{A^n} h_n = h'$ (-in which neighboring nodes on the path are immediate successors of one another, in one order or another, and no nodes are repeated).

Epistemic relations on histories To make this tree into a model for our restricted vocabulary $\mathcal{V}_\Sigma$, we define our *single-agent indistinguishability relations* $\sim_a \subseteq H \times H$ on histories, by putting

$$\sim_a := \left(\bigcup_{A \ni a} \to_A \cup \bigcup_{A \ni a} \leftarrow_A \right)^*,$$

where $\leftarrow_A$ is the converse of $\to_A$, the unions range over groups $A \subseteq \mathcal{A}_\Sigma$ s.t. $a \in A$, and R^* is the reflexive-transitive closure of R. Since our goal is to build a standard model, the *group indistinguishability relations* $\sim_A \subseteq H \times H$ are taken to be simply the *intersections* of all the individual relations:

$$\sim_A := \bigcap_{a \in A} \sim_a$$

It is useful to give a more concrete characterizations of the knowledge relations $\sim_a$ and $\sim_A$ on histories.

Lemma 4. *The following are* equivalent*, for all $a \in \mathcal{A}_\Sigma$ and histories $h, h' \in H$:*

1. $h \sim_a h'$;
2. *$a \in A^i$, for all groups A^i that appear as transition labels on the non-redundant path from h to h'.*

Proof. This is obvious, by the definition of $\sim_a$ on H, and the uniqueness of the non-redundant path. □

Lemma 5. *The following are* equivalent*, for $A \subseteq \mathcal{A}_\Sigma$ and histories $h, h' \in H$:*

1. $h \sim_A h'$;
2. *$A \subseteq A^i$, for all groups A^i that appear as transition labels on the non-redundant path from h to h'.*

Proof. This follows from Lemma 4, by the uniqueness of the non-redundant path from h to h'. □

Next, we show that the map *last* preserves $\sim_A$-relations:

Lemma 6. *If $h \sim_A h'$, then $last(h) \sim_A last(h')$.*

Proof. The proof is by *induction on the length N of the non-redundant path* from h to h':

Base case: $h = h'$. The conclusion follows trivially (given that $\sim_A$ are equivalence relations).

Inductive case: Suppose the non-redundant path from h and h' has length $N+1$, and let us look at the last transition on this path. Given Lemma 5, this transition can be either of the form $h_N \to_{A^N} h_{N+1} = h'$, or of the form $h_N \leftarrow_{A^N} h_{N+1} = h'$, with $A^N \supseteq A$. By definition of $\to_A$ on histories, we have either $h' = (h_N, A^N, last(h'))$ or $h_N = (h', A^N, last(h_N))$, with $last(h_N) \sim_{A^N} last(h')$ in both cases. By Monotonicity (Proposition 3) and the fact that $A \subseteq A^N$, we obtain $last(h_N) \sim_A last(h')$. On the other hand, we also have $last(h) \sim_A last(h_N)$ (-since the non-redundant path from h to h_N has length N, so by the induction hypothesis the pair (h, h_N) satisfies the conclusion of our Lemma, with h' replaced by h_N). Putting these two together (and using the transitivity of $\sim_A$), we conclude that $last(h) \sim_A last(h')$, as desired. □

The Value Domain: a Quotient Construction The *set of values D* of our model will be given by a quotient of the Cartesian product $H \times Var_\Sigma$. We *will define an equivalence relation* $\approx$ on pairs $(history, variable)$ in $H \times Var_\Sigma$ (telling us *when two such pairs represent the same value*), then we'll define our canonical set of objects D to be *the quotient of $H \times Var_\Sigma$ with respect to* $\approx$.

For this, we first introduce another *equivalence relation* $\sim$ on $H \times Var_\Sigma$, representing "*identity of objects at a given node*". This is given by putting:

$$(h, x) \sim (h', x') \text{ iff } h = h' \text{ and } (x = x') \in last(h).$$

Proposition 8. *$\sim$ is an equivalence relation.*

Proof. This follows immediately from the fact that Δ-equality is an equivalence relation for every type Δ (together with the fact that equality on histories is an equivalence relation). □

Second, we introduce a *one-step relation* $\rightarrow$ *between pairs* in $H \times Var_\Sigma$, given by putting:

$$(h,x) \rightarrow (h',x') \text{ iff } \exists y \in Var_\Sigma \exists A \supseteq \mathscr{A}(y) \text{ s.t. } h \rightarrow_A h', (x=y) \in last(h) \;\&\; (x'=y) \in last(h').$$

Finally, we define our main equivalence relation $\approx$ on pairs $(history, variable)$ in $H \times Var_\Sigma$, as the *smallest equivalence relation on $\mathscr{D}$, that includes the relations $\sim$ and $\rightarrow$*. It is useful to have a more concrete characterization of $\approx$. A first such characterization is immediate from the reflexivity and symmetry of $\sim$: if we denote the converse of $\rightarrow \subseteq H \times Var_\Sigma$ by $\leftarrow$, then it is easy to see that $\approx$ *coincides with the transitive closure* $(\sim \cup \rightarrow \cup \leftarrow)^+$ *of the union of the relations* $\sim$, $\rightarrow$ *and* $\leftarrow$. To obtain a more useful characterization, note that the relational compositions $\rightarrow;\sim$ and $\sim;\rightarrow$ coincide with $\rightarrow$. This, together with the existence of the non-redundant path from h to h', immediately give us the following:

Lemma 7. *("Path Lemma") Let $(h,x),(h',x') \in H \times Var_\Sigma$, and let $h = h_0 \leftarrow h_1 \leftarrow \ldots \leftarrow h_i \rightarrow \ldots \rightarrow h_n = h'$ be the non-redundant path from h to h'. Then the following are equivalent:*

1. *$(h,x) \approx (h',x')$;*
2. *there exist terms $x_0, x_1, \ldots, x_i, \ldots, x_n \in Var_\Sigma$ such that we have:*

$$(h,x) = (h_0,x_0) \leftarrow (h_1,x_1) \leftarrow (h_2,x_2) \leftarrow \cdots \leftarrow (h_i,x_i) \rightarrow \cdots \rightarrow (h_n,x_n) = (h',x')$$

From Quasi-Model to Model We are first defining our *first-order data model* $\mathbf{D} = (D, I)$ for the restricted vocabulary $\mathscr{V}_\Sigma$: as announced, the *set of 'values' D* is the quotient

$$D := (H \times Var_\Sigma)/\approx = \{[h,x] : (h,x) \in H \times Var_\Sigma\},$$

consisting of equivalence classes modulo $\approx$

$$[h,x] := \{(h',x') \in H \times Var_\Sigma : (h,x) \approx (h',x')\} \text{ (for any given pair } (h,x) \in H \times Var_\Sigma).$$

The *interpretation function I* will map every constant $c \in C \cap Var_\Sigma$ into $I(c) = c_{\mathbf{D}} := [(\Delta_0), c] = [h,c]$ (for all $h \in H$), and map all $c \in C - Var_\Sigma$ into $I(c) := \uparrow_{\mathbf{D}}$; it will also map n-ary predicate symbols $P \in Pred_\Sigma$ into n-ary relations $I(P) \subseteq D^n$ given by

$$I(P) := \{([h,x_1],\ldots,[h,x_n]) \in D^n : h \in H, \overline{x} = (x_1,\ldots,x_n) \in Var_\Sigma^n \text{ s.t. } P\overline{x} \in last(h)\};$$

and it will map n-ary functional symbols $F \in Funct_\Sigma$ into an n-ary functions $I(F) : D^n \rightarrow D$ given by

$$I(F)([h,x_1],\ldots,[h,x_n]) := [h, F(x_1,\ldots,x_n)] \text{ if } F(x_1,\ldots,x_n) \in Var_\Sigma, \text{ and}$$
$$I(f)([h,x_1],\ldots,[h,x_n]) := \uparrow_{\mathbf{D}} \text{ otherwise.}$$

Next, *our epistemic data model* $\mathbf{M} = (H, \sim, \bullet(\bullet))$ is given by taking: as set of states, the set H of all histories; the indistinguishability relations $\sim_a \subseteq H \times H$ as defined above on histories[10]; and the basic assignment function $\bullet(\bullet) : H \times V_\Sigma \rightarrow D$ is given by putting $h(v_i) := [h, v_i]$, for all $h \in H$ and $v \in V_\Sigma$.

Lemma 8. *("Interpretation Lemma") The interpretation I is well-defined, i.e. we have the following:*

1. *$(h,\overline{x}) \approx (h',\overline{y})$ implies $(h,F(\overline{x})) \approx (h',F(\overline{y}))$;*

[10] Note that this is meant to be a (standard) model, so the group indistinguishability relations $\sim_A$ are just the intersections $\bigcap_{a \in A} \sim_a$, thus coinciding with the general relations $\sim_A$ introduced above.

2. $(h,\overline{x}) \approx (h',\overline{y})$ *implies that:* $(P\overline{x}\overline{z}) \in last(h)$ *iff* $(P\overline{y}\overline{z}) \in last(h')$*;*

3. $I(E)$ *is really the identity relation on* D*:*

$$([[h,x],[h',x']) \in I(E) \text{ iff } [h,x] = [h',x'] \text{ iff } (h,x) \approx (h',x')$$

Proof. The proofs are by induction on the length of the non-redundant path from h to h', using our conditions on types. □

Lemma 9. *("Preservation of Values") If* $h \sim_A h'$ *are histories and* $x \in Var_\Sigma$ *is s.t.* $\mathscr{A}(x) \subseteq A$*, then* $[h,x] = [h',x]$.

Proof. Let $h = h_0 \leftarrow_{A^1} h_1 \leftarrow_{A^2} \ldots \leftarrow_{A^i} h_i \rightarrow_{A^{i+1}} \ldots \rightarrow_{A^n} h_n = h'$ be the non-redundant path from h to h'. Since $h \sim_A h'$, we know that $A \subseteq A_k$ for all k (by Lemma 5). Using this together with $\mathscr{A}(x) \subseteq A$ (and the definition of the relation $(h,x) \rightarrow (h',x')$ on history-variable pairs), we can see that we have $(h,x) \leftarrow (h_1,x) \leftarrow \cdots (h_i,x) \rightarrow \cdots \rightarrow (h_n,x) = (h',x)$. Thus, we have $(h,x) \approx (h',x)$, and hence $[h,x] = [h',x]$. □

Lemma 10. *("Knowledge-of-Value Lemma") If* $(x_A^\varphi\downarrow) \in last(h)$*, then there exists some history* $h' \in H$*, satisfying:*

1. $h' \sim_A h$*;*

2. $\varphi \in last(h')$*;*

3. $[h'',x] = [h',x] = [h,x_A^\varphi]$ *for all* $h'' \sim_A h$ *s.t.* $\varphi \in last(h'')$.

Proof. From $(x_A^\varphi\downarrow) \in last(h)$, we infer that $\langle K_A\rangle\varphi \in last(h)$ (by condition 8 in the definition of types), so by the Diamond Lemma there exists some $h' \sim_A h$ s.t. $\varphi \in last(h')$.

To prove the third claim of the lemma, let $h'' \in H$ be s.t. $h'' \sim_A h$ and $\varphi \in last(h'')$. From $(x_A^\varphi\downarrow) \in last(h)$ and $h'',h' \sim_A h$, we obtain that $(x_A^\varphi\downarrow) \in last(h') \cap last(h'')$ (by the definition of $\sim_A$ on histories and the fact that $\mathscr{A}(x_A^\varphi\downarrow) = A$). From this together with $\varphi \in last(h') \cap last(h'')$, we obtain that $(x = x_A^\varphi) \in last(h') \cap last(h'')$ (by condition 9 on types), hence $(h',x) \approx (h',x_A^\varphi)$ and $(h'',x) \approx (h'',x_A^\varphi)$. On the other hand, by Lemma 9 and the fact that $\mathscr{A}(x_A^\varphi) = A$, we have $(h',x_A^\varphi) \approx (h,x_A^\varphi) \approx (h',x_A^\varphi)$. Putting these together and using the transitivity and symmetry of the relation $\approx$, we obtain that $(h'',x) \approx (h',x) \approx (h,x_A^\varphi)$, i.e. $[h'',x] = [h',x] =]h,x_A^\varphi]$, as desired. □

Lemma 11. *("Unknown Proper Value Lemma") If* $(x_A^\varphi\uparrow), \langle K_A\rangle(\varphi \wedge x\downarrow) \in last(h)$*, then there exist histories* $h' \sim_A h'' \sim_A h$*, with* $\varphi \in last(h') \cap last(h'')$ *and* $[h',x] \neq [h,x]$.

Proof. From $\langle K_A\rangle(\varphi \wedge x\downarrow) \in last(h)$, we infer by the Diamond Lemma that there exists $h' \sim_A h$ with $\varphi, x\downarrow \in last(h')$. If we put $\Delta := last(h')$, we have $last(h) \sim_A \Delta$ and $\varphi, x\downarrow \in \Delta$, and also $x_A^\varphi\uparrow \in \Delta$ (by the definition of $\sim_A$ on histories and the fact that $x_A^\varphi\uparrow \in last(h) \sim_A \Delta$, together with the fact that $\mathscr{A}(x_A^\varphi\uparrow) = A$). Thus, the assumptions in clause 12 on quasi-models are satisfied, except that we need to find some suitable y to apply clause 12.

Let us choose y to be some randomly chosen variable $y \in Var_\Sigma$ s.t. $(x = y) \in \Delta$ and $\dagger \subseteq A$, if such a variable exists (Case 1); while otherwise (Case 2), we let $y := \uparrow$. In both cases, we have $\mathscr{A}(y) \subseteq A$. Applying now clause 12 on types (to Δ and y), we must have that $\langle K_A\rangle(\varphi \wedge x \neq y) \in \Delta$. By condition (**) in the definition of pseudo-models, there exists some $\Delta' \in S$ with $\Delta' \sim_A \Delta$ and $\varphi, (x \neq y) \in \Delta'$. Take now the history $h'' := (h',A,\Delta') = (h,A,\Delta,A,\Delta')$. It is clear that $h \sim_A h' \sim_A h''$ and that $\varphi \in \Delta \cap \Delta' = last(h') \cap last(h'')$. So, to prove our Lemma, we only need to check that $(h',x) \not\approx (h'',x)$.

For this, assume towards a contradiction that we have $(h',x) \approx (h'',x)$. Using the characterization of $\approx$ in the Path Lemma 7 and the fact that the non-redundant path from h' to h'' is $(h', \rightarrow^A, h'')$, we infer that

there exists $y' \in Var_\Sigma$ with $\mathscr{A}(y') \subseteq A$, $(x = y') \in last(h') = \Delta$ and $(x = y') \in last(h'') = \Delta'$. The fact that we have $(x = y') \in \Delta$ means that we are in Case 1 above, so that our previously chosen y has the property that $(x = y) \in \Delta$. Using the fact that Δ-equality is an equivalence relations on variables in Var_Σ, we obtain that $(y' = y) \in \Delta$. Since $\mathscr{A}(y), \mathscr{A}(y') \subseteq A$, we have $\mathscr{A}(y' = y) \subseteq A$, and so from $(y' = y) \in \Delta$ and $\Delta \sim_A \Delta'$ we infer that $(y' = y) \in \Delta'$ (by the definition of $\sim_A$ on histories). Putting this together with the fact that $(x = y') \in \Delta'$ and using again the $=$-transitivity condition on types, we conclude that $(x = y) \in \Delta'$. But this contradicts the fact that $(x \neq y) \in \Delta'$ (given condition 1 on types). □

Lemma 12. *("Truth Lemma") For every formula $\varphi \in \Sigma$ and every variable $x \in Var_\Sigma$, the following hold for all histories $h \in H$:*

1. *$h \models \varphi$ iff $\varphi \in last(h)$;*
2. *$h(x) = [h,x]$.*

Proof. We need a *non-standard notion of "complexity" order on both formulas $\varphi \in \Sigma$ and terms $x \in Var_\Sigma$*, according to which: (a) a formula is more complex than its subformulas and than the terms occurring in it; (b) a term is more complex than its subterms (variables); (c) the term x_A^φ is more complex than the formulas $\langle K_A\rangle(\varphi \wedge x{\downarrow}), K_A(\varphi \to x{=}{\uparrow})$. It is easy to see that *there indeed exists a well-founded partial order on Σ satisfying these properties*.

We now prove the two claims *simultaneously by induction on our non-standard "complexity" order*.

Proof of the Atomic Cases for Claims 1 and 2:

(i) *Propositional Atoms*: $\varphi = P\overline{x}$, with $\overline{x} = (x_1, \ldots, x_n)$. We have the following equivalences:
$h \models P\overline{x}$ iff $h(\overline{x}) \in I(P)$ iff (by induction hypothesis for claim (2)) $([h,x_1], \ldots, [h,x_n]) \in I(P)$ iff (by definition of $I(P)$) $\exists h' \in H$ s.t. $((h,\overline{x}) \approx (h',\overline{x}) \& (P\overline{x}) \in last(h'))$ iff (by Lemma 8) $(P\overline{x}) \in last(h)$.

(ii) *Constants*: By definition, $w^h(c) = c_{\mathbf{D}} = [h,c]$.

(iii) *Atomic variables*: $x := v_i \in V_\Sigma$ is a basic variable. This is trivial: by definition, $w^h(v_i) = [h, v_i]$.

Proof of the Inductive Cases for Claims 1 and 2:

(iv) *Boolean Cases*: $\neg\varphi, \varphi \wedge \psi$. These is trivial, using conditions 1 and 2 on types.

(v) *K_A-modal Case*: $(K_A\varphi) \in \Sigma$. From *left-to-right*: assume (towards a contradiction) that we have $h \models K_A\varphi$ but $(K_A\varphi) \notin last(h)$. Then, by the closure condition 1 on types, we have $(\langle K_A\rangle{\sim}\varphi) \in last(h)$, and so by the property (**) of quasi-models, there exists some type $\Delta' \in S$ s.t. $last(h) \sim_A \Delta'$ and $({\sim}\varphi) \in \Delta'$. Consider the history $h' := (h, A, \Delta')$: this is a well-defined history in H, satisfying $h \sim_A h'$ and $last(h') = \Delta'$. From $h \models K_A\varphi$ we infer that $h' \models \varphi$, and so by the induction hypothesis we have $\varphi \in last(h') = \Delta'$. But this contradicts the fact that $({\sim}\varphi) \in \Delta'$, given condition 1 on types.
For the *right-to-left* direction: we assume that $(K_A\varphi) \in last(h)$, and we have to prove that $h \models K_A\varphi$. For this, let $h' \in H$ be s.t. $h \sim_A h'$, and we need to show that $h' \models \varphi$. From $h \sim_A h'$ we obtain that $last(h) \sim_A last(h')$ (by Lemma 6), which together with $h \models K_A\varphi$ gives us $\varphi \in last(h')$ (by Proposition 4). Applying the induction hypothesis, we conclude that $h' \models \varphi$, as desired.

(vi) *$C_{\mathfrak{A}}$-modal Case*: $(C^\theta_{\mathfrak{A}}\varphi) \in \Sigma$. From *left-to-right*: assume (towards a contradiction) that we have $h \models C^\theta_{\mathfrak{A}}\varphi$ but $(C^\theta_{\mathfrak{A}}\varphi) \notin last(h)$. Then, by the closure condition 1 on types, we have $(\langle C^\theta_{\mathfrak{A}}\rangle{\sim}\varphi) \in last(h)$, and so by the property (**) of quasi-models, there exists some finite chain of types $\Delta_0 \sim_{A^1} \Delta_1 \ldots \sim_{A^n} \Delta_n$, with $A^1, \ldots, A^n \in \mathfrak{A}$, $\Delta_0 = last(h)$, $\theta \in \Delta_1, \ldots, \Delta_n \in S$ and $({\sim}\varphi) \in \Delta_n$. Consider the histories $h_1 = (h, A^1, \Delta_1)$, $h_2 = (h_1, A^2, \Delta_2) = (h, A^1, \Delta_1, A^2, \Delta_2)$, …, $h_n := (h, A_1, \Delta_1, \ldots, A_n, \Delta_n)$: these are well-defined histories in H, satisfying $h = h_0 \sim_{A^1} h_1 \sim_{A^2} h_2 \cdots \sim_{A^n} h_n$ and $\theta \in last(h_1), \ldots, last(h_n)$. By the induction hypothesis, we have $h_1, h_2, \ldots, h_n \models \theta$. Using this and the assumption that $h \models C^\theta_{\mathfrak{A}}\varphi$ (as well as the semantics of $C_{\mathfrak{A}}$), we infer that $h_n \models \varphi$. Applying again the induction hypothesis, we have $\varphi \in last(h_n) =$

Δ_n, but this contradicts the fact that $(\sim\varphi)\in\Delta$ (given the consistency condition 1 on types).
For the *right-to-left* direction: we assume that $(C^\theta_{\mathfrak{A}}\varphi)\in last(h)$, and we have to prove that $h\models C^\theta_{\mathfrak{A}}\varphi$. For this, let $h=h_0\sim_{A^1}h_1\cdots h_{n-1}\sim_{A^n}h_n$ be any chain of histories (of any length $n\geq 0$) with $A^1,\ldots,A^n\mathfrak{A}$ and $h_1,\ldots,h_n\theta$, and we need to show that $h_n\models\varphi$. For this, we first prove the auxiliary claim that $(C^\theta_{gA}\varphi)\in last(h_n)$, by induction on the length n of the chain. For $n=0$, we have $h_0=h$, so we already have $(C^\theta_{\mathfrak{A}}\varphi)\in last(h)=last(h_0)$. For the inductive step for $n\geq 1$, we can assume by induction that $(C^\theta_{\mathfrak{A}}\varphi)\in last(h_{n-1})$. Using condition 13 on types, we infer that $(K_{A^n}C^\theta_{\mathfrak{A}}\varphi)\in last(h_{n-1})$. Using the fact that $h_{n-1}\sim_{A^n}h_n$ and Proposition 4, we obtain that $(C^\theta_{\mathfrak{A}}\varphi)\in last(h_n)$, so we have established our auxiliary claim. From this, we infer that $\varphi\in last(h_n)$ (using again condition 13 on types), and so the induction hypothesis for Claim 1, we conclude that $h_n\models\varphi$, as desired.

(vii) *Terms defined by cases* $x|_\varphi y$. Given $(x|_\varphi y)\in\Sigma$, we have $\varphi\in\Sigma$, so we have that either $\varphi\in last(h)$, or else $(\sim\varphi)\in last(h)$. In the first case, we have $(x|_\varphi y=x)\in last(h)$ (by condition 6 on types), hence $[h,x|_\varphi]=[h,x]$; and on the other hand, $\varphi\in last(h)$ yields $h\models\varphi$ (by the induction hypothesis for Claim 1), so by definition we have $h(x|_\varphi y)=h(x)$, and by the induction hypothesis for Claim 2 we have $h(x)=[h,x]$, thus obtaining $h(x|_\varphi y)=h(x)=[h,x]=h(x|_\varphi)$, as desired. The second case is similar: from $(\sim\varphi)\in last(h)$ we get $(x|_\varphi y=y)\in last(h)$ (by condition 6 on types), hence $[h,x|_\varphi]=[h,y]$; and on the other hand, $(\sim\varphi)\in last(h)$ implies $\varphi\notin last(h)$ (by the consistency of types), which by the induction hypothesis for Claim 1 yields $h\not\models\varphi$, so by definition we have $h(x|_\varphi y)=h(y)$, and by the induction hypothesis for Claim 2 we have $h(y)=[h,y]$, thus obtaining $h(x|_\varphi y)=h(y)=[h,y]=h(x|_\varphi)$, as desired.

(viii) *Functional terms* $F(\overline{x})\in Var_\Sigma$ for some $\overline{x}=(x_1,\ldots,x_n)$. We assume by the induction hypothesis that $h(\overline{x})=[h,\overline{x}]$, and using the semantics of functional terms and the definition of $I(F)$ in our history model, we obtain $h(F(\overline{x}))=(I(F))(h(\overline{x}))=(I(F))[h,\overline{x}]=[h,F(\overline{x})]$, as desired.

(ix) *Conditional terms* $x^\varphi_A\in Var_\Sigma$. We distinguish three subcases.

Case 1: $x^\varphi_A{\uparrow},\langle K_A\rangle(\varphi\wedge x{\downarrow})\in last(h)$. From $(x^\varphi_A=\bot)\in last(h)$, we get that $(h,x^\varphi_A)\sim(h,\bot)$, hence $[h,x^\varphi_A]=[h,\bot]=\bot$. So, to prove that $h(x^\varphi_A)=[h,x^\varphi_A]$, we need to show that $h(x^\varphi_A)=\bot$. For this, we apply Lemma 11 and the assumptions that $x^\varphi_A{\uparrow},\langle K_A\rangle(\varphi\wedge x{\downarrow})\in last(h)$, to obtain histories h',h'' with $h\sim_A h'\sim_A h''$, $\varphi\in last(h')\cap last(h'')$ and $[h',x]\neq[h'',x]$. By the induction hypothesis for both Claims 1 and 2, we get that $h'\models\varphi$ and $h''\models\varphi$ and $h'(x)=[h',x]\neq[h'',x]=h''(x)$. By definition, the existence of histories h',h'' with these properties implies that $h(x^\varphi_A)=\uparrow$, as desired.

Case 2: $x^\varphi_A{\uparrow}\in last(h)$, but $\langle K_A\rangle(\varphi\wedge x{\downarrow})\notin last(h)$. Since $\langle K_A\rangle(\varphi\wedge x{\downarrow})=\langle K_A\rangle(\varphi\wedge x\neq\uparrow)\in\Sigma$ (because of the closure conditions on Σ and the fact that $x^\varphi_A{\uparrow}\in last(h)\subseteq\Sigma$), we must have that $K_A(\varphi\to x=\uparrow)\in last(h)$ (since $last(h)$ is a type). By condition 10 on types, we have $(x^\varphi_A=\uparrow^\varphi_A)\in last(h)$, which together with condition 11 on types gives us $(x^\varphi_A=\uparrow)\in last(h)$, i.e. $(h,x^\varphi_A)\sim(h,\uparrow)$, so $[h,x^\varphi_A]=[h,\uparrow]=\uparrow_{\mathbf{D}}$. To prove that $[h,x^\varphi_A]=h(x^\varphi_A)$ as desired, we only need to check now that $h(x^\varphi_A)=\uparrow_{\mathbf{D}}$. For this, let $h'\in H$ s.t. $h\sim_A h'$ and $h'\models\varphi$, and we need to show $h'(x)=\uparrow_D$. By the induction hypothesis for Claim 1 and the fact that $K_A(\varphi\to x=\uparrow)\in last(h)$, we have $h\models K_A(\varphi\to x=\uparrow)$, hence $h'\models(\varphi\to x=\uparrow)$, which together with $h'\models\varphi$ gives us $h'\models(x=\uparrow)$. Applying again the induction hypothesis for Claim 1, we have $(x=\uparrow)\in last(h)$, hence $(h,x)\sim(h,\uparrow)$ and so $[h,x]=[h,\uparrow]=\uparrow_D$, as desired.

Case 3: $(x^\varphi_A{\downarrow})\in last(h)$. Using Lemma 10 and the Induction Hypothesis for Claims 1 and 2, we infer the existence of some history $h'\in H$, s.t: $h'\sim_A h$; $h'\models\varphi$; and $h''(x)=h'(x)=[h,x^\varphi_A]$ for all $h''\sim_A h$ s.t. $h''\models\varphi$. This establishes the existence of a unique value $d:=[h,x^\varphi_A]\in D^\bot$ satisfying the conditions in the first semantic clause for $h(x^\varphi_A)$, so by definition we have $h(x^\varphi_A)=d=[h,x^\varphi_A]$, as desired. □

Proof of Proposition 7: If S is a quasi-model for φ_0, then by applying claim 1 of the Truth Lemma 12 to φ_0 and to the history $h_0:=(\Delta_0)$, we conclude that $h_0\models\varphi_0$ in our history-based data model $\mathbf{M}$ above.

6 Completeness and Decidability for the dynamic logic *DLKV*

In this section we prove Theorem 2. We first establish our co-expressivity result: *DLKV and LKV are provably co-expressive*. For this, we need a few preliminary notions and results.

Expressions and subexpression-complexity An *expression* ε of *DLKV* is any formula φ, term x or event e of *DLKV*. The *sub-expression complexity order* $>$ is the natural extension to expressions of the usual notion of subformula-complexity; more precisely, $>$ is the least transitive relation on expressions, s.t.: (1) every formula is $>$ all its subformulas and all terms occurring in it; (2) every term is $>$ all its subterms; (3) every event $e = \Phi/\sigma$ is $>$ than all formulas $\varphi \in \Phi$ and all postconditions $\sigma(v)$ (for all $v \in V$); and (4) the formula $[e]\varphi$ is $> e$.

It is easy to see that $<$ *is a well-founded partial order*.

Reducible expressions A formula θ in *DLKV* is *reducible* if it is provably equivalent to a 'static' formula; i.e., if there exists some formula θ' in the static fragment *LKV* s.t. $\vdash \theta \leftrightarrow \theta'$ is provable in **DLKV**. A term $x \in Var_{DLKV}$ is *reducible* if it is provably equal to a static term; i.e., if there exists some $x' \in Var_{LKV}$ s.t. $\vdash x = x'$ is provable in **DLKV**. Finally, an event $e = \Phi/\sigma$ is *reducible* if all formulas $\varphi \in \Phi$ and all its post-conditions $\sigma(v)$ are reducible (for all $v \in V$).

Observation. *If $e = \Phi/\sigma$ is a reducible event, then its precondition $pre_e := \bigwedge \Phi$ is a reducible formula.*

Lemma 13. *The following are provable in* **DLKV***:*

1. (Derived Reduction Axiom for Conjunction) $\vdash [e](\varphi \wedge \psi) \leftrightarrow ([e]\varphi \wedge [e]\psi)$
2. (Existential Dynamic Modality) $\vdash \langle e\rangle\theta \leftrightarrow (pre_e \wedge [e]\theta)$

Proof. The first is a standard consequence of $[e]$-Necessitation and $[e]$-Distributivity, that holds for all normal modalities. The second follows immediately from the Partial Functionality reduction axiom, together with the definition of $\langle e\rangle\theta$ as an abbreviation for $\neg[e]\neg\theta$. □

Lemma 14. (Replacement of Equivalents) *Suppose that $\vdash \varphi \leftrightarrow \varphi'$, $\vdash \theta \leftrightarrow \theta'$, $\vdash x = x'$, $\vdash y = y'$ and $\vdash x_i = x_i'$ (for all $i = 1,n$) are provable in* **DLKV**. *Then the following are also provable in* **DLKV***:*

1. $\vdash x|_\varphi y = x'|_{\varphi'} y'$
2. $\vdash F(x_1,\ldots,x_n) = F(x_1',\ldots,x_n')$
3. $\vdash x_A^\varphi = (x')_A^{\varphi'}$
4. $\vdash e(x) = e(x')$
5. $\vdash Px_1\ldots x_n \leftrightarrow Px_1'\ldots x_n'$
6. $\vdash \neg\varphi \leftrightarrow \neg\varphi'$
7. $\vdash (\varphi \wedge \theta) \leftrightarrow (\varphi' \wedge \theta')$
8. $\vdash K_A\varphi \leftrightarrow K_A\varphi'$
9. $\vdash [e]\varphi \leftrightarrow [e]\varphi'$
10. $\vdash C_{\mathfrak{A}}^\theta\varphi \leftrightarrow C_{\mathfrak{A}}^{\theta'}\varphi'$

We can now prove a preliminary "one-step reduction" result:

Lemma 15. *Let e be any reducible event. Then, for every 'static' formula θ in LKV and every 'static' term $x \in Var_{LKV}$, the formula $[e]\theta$ and the term $e(x)$ are also reducible.*

Proof. Since $e = \Phi/\sigma$ is reducible, its precondition $pre_e = \bigwedge\Phi$ is also reducible. Let us fix some static formula ρ that is provably equivalent to pre_e. Throughout the proof, we'll make liberal use of *Replacement of Equivalents* Lemma 14, without explicitly mentioning it.

We now prove both claims simultaneously, by induction on sub-expression complexity:

For $\theta := Px_1 \dots x_n$: by the induction hypothesis, for all $i = 1,n$ we have $\vdash e(x_i) = x_i'$ for some static terms $x_1', \dots, x_n' \in Var_{LKV}$. Using also the Indiscernability axiom and the Atomic Change reduction axiom, $[e]\theta$ is provably equivalent to $pre_e \to Px_1' \dots x_n'$, and hence also to $\rho \to Px_1' \dots x_n'$.

The case $\theta := \neg\varphi$: by the induction hypothesis, φ is reducible, so there exists some static formula φ_e s.t. $\vdash [e]\varphi \leftrightarrow \varphi_e$ is provable. This, together with our Lemma's assumption and the Partial Functionality reduction axiom, gives us $\vdash [e]\theta \leftrightarrow (\rho \to \neg\varphi_e)$.

The case $\theta := \varphi \wedge \theta$ is similar: by induction, there exist static formulas φ_e and θ_e s.t. $\vdash [e]\varphi \leftrightarrow \varphi_e$ and $\vdash [e]\theta \leftrightarrow \theta_e$ are provable. Using now Lemma 13.1 (the Derived Reduction Axiom for Conjunction) we obtain $\vdash [e]\theta \leftrightarrow (\varphi_e \wedge \theta_e)$.

For $\theta := K_A\varphi$: by induction, there exists some static formula φ_e s.t. $\vdash [e]\rangle\varphi \leftrightarrow \varphi_e$ is provable. Putting this together with the Knowledge Update reduction axiom, we get $\vdash [e]\theta \leftrightarrow (\rho \to K_{e(A)}\varphi_e)$.

For $\theta := C^{\theta}_{\mathfrak{A}}\varphi$: we use the induction hypothesis to get static formulas φ_e and θ_e s.t. $\vdash [e]\varphi \leftrightarrow \varphi_e$ and $\vdash [e]\theta \leftrightarrow \theta_e$ are provable. Using now the $C^{\theta}_{\mathfrak{A}}\varphi$-Update reduction axiom, as well as Lemma 13.2, we obtain $\vdash [e]\theta \leftrightarrow (\rho \to C^{\rho\wedge\theta_e}_{e[\mathfrak{A}]}\varphi_e)$.

Moving to terms: the case $x := c \in C$ is trivial, since the Preservation of Constants reduction axiom implies that we have $\vdash e(c) = c|_{pre_e}\uparrow$, so $e(c)$ is reducible: $\vdash e(c) = c|_{\rho}\uparrow$.

For $x := v$ (basic variable): since e is reducible, there exists some static term x' s.t. $\vdash post_e(v) = x'$. Using this and the Change of Basic Variables reduction axiom, we obtain that $\vdash e(x) = x'|_{\rho}\uparrow$, so $e(x)$ is reducible.

For $x := y|_{\varphi}z$: by the induction hypothesis, there exist static terms y', z' and static formula φ', s.t. $\vdash e(y) = y'$, $\vdash e(z) = z'$ and $\vdash [e]\varphi \leftrightarrow \varphi'$ are provable. Using these, as well as the Change of Disjunctive Terms reduction axiom and Lemma 13.2, we obtain $\vdash e(x) = (y'|_{\rho\wedge\varphi'}z')|_{\rho}\uparrow$.

For $x := F(x_1, \dots, x_n)$: by the induction hypothesis, there exist static terms $x_1', \dots, x_n'$ s.t. $\vdash e(x_i) = x_i'$ is provable for all $i \leq n$. Using the Change of Functional Terms reduction axiom, we obtain $\vdash e(x) = F(x_1', \dots, x_n')|_{\rho}\uparrow$.

For $x := y^{\varphi}_A$: by induction, there exists a static formula φ', s.t. $\vdash [e]\varphi \leftrightarrow \varphi'$ is provable. Using also the Change of Hypothetical Values reduction axiom, as well as Lemma 13.2, we conclude that $\vdash e(x) = e(y)^{\rho\wedge\varphi'}_{e(A)}|_{\rho}\uparrow$. □

We are now in the position to establish our full reduction result:

Lemma 16. (Co-expressivity of *DLKV* and *LKV*) *All expressions (formulas, terms and events) ε of DLKV are reducible. As a consequence, DLKV and LKV have the same expressive power.*

Proof. We prove this by induction on the subexpression-complexity of the expression ε.

The base cases $\varepsilon := c \in C$ and $\varepsilon := v \in V$ are trivial, since these are already static terms.

The cases $\varepsilon := x|_{\varphi}y$, $\varepsilon := F(\overline{x})$ and $\varepsilon := x^{\varphi}_A$ are straightforward: we use the induction hypothesis and parts 1,2 and 3 of Lemma 14.

For the case $\varepsilon := e(x)$: by induction, e and x are reducible; so there exists a static term x' s.t. $\vdash x = x'$ is provable. By Lemma 14.4, we have $\vdash \varepsilon = e(x')$, and by Lemma 15 $e(x'$ is reducible (since e is reducible and x' is static) so we have $\vdash e(x') = x''$ for some static term x''. Putting these together and using transitivity of equality, we obtain $\vdash \varepsilon = x''$, so ε is reducible.

The cases $\varepsilon := P\overline{x}$, $\varepsilon := \neg\varphi$, $\varepsilon := \varphi \wedge \theta$, $\varepsilon := K_A\varphi$ and $\varepsilon := C^{\theta}_{\mathfrak{A}}\varphi$ are again straightforward: use the induction hypothesis and parts 5,6,7,8 and 10 of Lemma 14.

For the case $\varepsilon := [e]\varphi$: by induction, e and φ are reducible; so there exists a static formula φ' s.t. $\vdash \varphi \leftrightarrow \varphi'$ is provable. By 14.9, we have $\vdash \varepsilon \leftrightarrow [e]\varphi'$, and by 15 $[e]\varphi'$ is reducible (since e is reducible and φ' is static) so we have $\vdash [e]\varphi' \leftrightarrow \varphi''$ for some static formula φ''. Putting these together and using transitivity of equivalence, we obtain $\vdash \varepsilon \leftrightarrow \varphi''$, hence ε is reducible.

Finally, the case $\varepsilon := e = \Phi/\sigma$ is straightforward: since all the formulas $\varphi \in \Phi$ and all postconditions $\sigma(v)$ (with $v \in V$) are $< e$ in the subexpression complexity order, they are all reducible (by the induction hypothesis), and hence e is also reducible (by definition). □

Finally, we can prove Theorem 2 (completeness and decidability of **DLKV**):

Proof of Theorem 2: For completeness, let θ be a valid formula in the language *DLKV*. By Lemma 16, there exists some static formula θ' in *LKV* s.t. $\vdash \theta \leftrightarrow \theta'$ is a theorem in the axiomatic system **DLKV**. By the soundness of **DLKV**, θ' is also valid. By the completeness of the system **LKV** (Theorem 1), $\vdash \theta'$ is provable in the system **LKV** (since θ' is a static formula), and hence it is also a theorem in the system **DLKV**. Using this, together with the **DLKV**-theorem $\vdash \theta \leftrightarrow \theta'$ again, we conclude that $\vdash \theta$ is provable in **DLKV**, as desired. Decidability follows immediately from the decidability of the static logic *LKV* (Theorem 1) and the co-expressivity of *LKV* and *DLKV* (Lemma 16).

7 Conclusions

In this paper we axiomatize a decidable but richly expressive logic, that can deal with both "knowledge that" and "knowledge what", in a multi-agent setting in which data is modeled using local variables that can be shared, exchanged or modified. Our investigation builds on prior work on knowledge of values and on data-exchange events, as well as more classical work in Epistemic Logic on public announcements, common and distributed knowledge. But our setting has a number of innovative features, such as definite descriptions for hypothetical values, conditional distributed knowledge of values, a conditional version of common distributed knowledge, etc. The proofs also present a number of challenges (due to the presence of the above features). Due to page limitations, we confined ourselves to semi-public events, but in future work we plan to extend this study to arbitrary data-exchange events (such as secret hacking).

References

[1] Agotnes, T., Wang, Y.N., "Resolving Distributed Knowledge", in: *Artificial Intelligence*, 252, pp. 1-21, (2017). doi:10.1016/j.artint.2017.07.002.

[2] Baltag, A., "To Know is to Know the Value of a Variable", in: *Adv. in Modal Logic 2016*, pp. 135-155, (2016).

[3] Baltag, A., van Benthem, J., "A Simple Logic of Functional Dependence", in: *Journal of Philosophical logic*, vol. 50, pp. 939-1005, (2021). doi:10.1007/S10992-020-09588-Z.

[4] Baltag, A., van Benthem, J., "Updates, Generalized p-Morphisms, and (Co-)Recursive Equations", To appear in van Benthem, J. and F. Liu (eds.), *Graph Games and Logic Design - Recent developments and further directions*, Springer, (2024).

[5] Baltag, A., Moss, L.S., "Logics for Epistemic Programs", in: *Synthese*, 139(2), pp. 165-224, (2004). doi:10.1023/B:SYNT.0000024912.56773.5E.

[6] Baltag, A., Moss, L.S, Solecki, S., "The Logic of Public Announcements, Common Knowledge, and Private Suspicions", in: *Proceedings TARK 98*, pp. 43-56, (1998).

[7] Baltag, A., Moss, L.S., Solecki, S., "Logics for epistemic actions: completeness, decidability, expressivity", in: *Logics*, 1 (2), pp. 97-147, (2023). doi:10.3390/logics1020006.

[8] Baltag, A., Renne, B., "Dynamic Epistemic Logic", in: *Stanford Encyclopedia of Philosophy*, (2016).

[9] Baltag, A., Smets, S., "Learning what Others Know", in: Kovacs, L. and E. Albert (eds.), *LPAR23 proceedings of the International Conference on Logic for Programming AI and Reasoning, EPiC Series in Computing*, Volume 73, pp. 90-110, (2020). doi:10.29007/PLM4.

[10] Baltag, A., Smets, S., "Logics for Data Exchange and Communication", Conference proceedings of Advances in Modal Logic 2024, vol. 15, 147-169, (2024).

[11] van Benthem, J., "One is a lonely number", in: Koepke, P.; Z. Chatzidakis and W. Pohlers, (eds.) *Logic Colloquium02*, pp.96-129, ASL and A.K. Peters, Wellesley MA, (2002). doi:10.1017/9781316755723.006.

[12] van Benthem, J., *Logical Dynamics of Information and Interaction*, Cambridge University Press, Cambridge UK, (2011). doi:10.1017/CBO9780511974533.

[13] van Benthem, J., van Eijck, J., Kooi, B., "Logics of Communication and Change"', in: *Information and Communication*, 204:11, pp. 1620-1662, (2006). doi:10.1016/j.ic.2006.04.006.

[14] Blackburn, P., de Rijke, M., Venema, Y., *Modal Logic*, Cambridge University Press, Cambridge, (2001).doi:10.1017/CBO9781107050884.

[15] van Ditmarsch, H., van der Hoek, W., Kooi, B., *Dynamic Epistemic Logic*, Springer, Dordrecht, (2007). doi:10.1007/978-1-4020-5839-4.

[16] van Eijck, J., Gattinger, M., Wang, Y., "Knowing Values and Public Inspection", in: *Proceedings 7th Indian Conference on Logic and its Applications*, Kanpur, 77–90, (2007). doi:10.1007/978-3-662-54069-5_7.

[17] Fagin, R., Halpern, J., Moses, Y., Vardi, M., *Reasoning About Knowledge*, The MIT Press, Cambridge MA, (1995). doi:10.7551/mitpress/5803.001.0001.

[18] Gerbrandy, J., *Bisimulations on Planet Kripke*, Doctoral Dissertation, ILLC, UvA. ILLCS DS Series DS-1999-01, (1998).

[19] Parikh, R., "Levels of Knowledge, Games and Group Action", in: *Research in Economics*, 57, pp. 267-281, (2003). doi:10.1016/S1090-9443(03)00030-9.

[20] Parikh, R., Ramanujam, R., "A Knowledge-Based Semantics of Messages", in: *Journal of Logic, Language and Information*, 12, pp. 453-467, (2003). doi:10.1023/A:1025007018583.

[21] Plaza, J., "Logics of Public Communication", in: *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*, pp. 201-216, (1989). doi:10.1007/S11229-007-9168-7.

[22] van Wijk, S., *Coalitions in Epistemic Planning.MSc Thesis*, ILLC, UvA, (2015). *https://www.illc.uva.nl/Research/Publications/Reports/MoL/*

[23] Wang, Y., "Beyond Knowing That: A New Generation of Epistemic Logics", in *Hintikka Volume*, Springer, Dordrecht, 499-533, (2018).

[24] Wang, Y., Fan, J., "Knowing that, Knowing what, and Public Communication: Public Announcement Logic with Kv Operators", in Proceedings of IJCAI 2013: 1147-1154. AAAI press, (2013).

[25] Wang, Y., Fan, J., "Conditionally knowing what", in Advances in Modal Logic Vol. 10: 569-587, College Publications, (2014).

[26] Gu, T., Wang, Y., " "Knowing value" logic as a normal modal logic", in Advances in Modal Logic Vol. 11:362-381, College Publications, (2016).

[27] Ding, Y., "The axiomatization and complexity of Knowing-What-Logic on model class K", doi:10.48550/arXiv.1609.07684, arXiv:1609.07684v2 (2016).

[28] Hong, B., "Knowing the Value of a Predicate". Conference LORI 2023, 149-166, (2023). doi:10.1007/978-3-031-45558-2_12.

Causality Without Causal Models

Joseph Y. Halpern
Cornell University
halpern@cs.cornell.edu

Rafael Pass
Cornell University
rafael@cs.cornell.edu

Perhaps the most prominent current definition of (actual) causality is due to Halpern and Pearl. It is defined using causal models (also known as *structural equations models*). We abstract the definition, extracting its key features, so that it can be applied to any other model where counterfactuals are defined. By abstracting the definition, we gain a number of benefits. Not only can we apply the definition in a wider range of models, including ones that allow, for example, backtracking, but we can apply the definition to determine if A is a cause of B even if A and B are formulas involving disjunctions, negations, beliefs, and nested counterfactuals (none of which can be handled by the Halpern-Pearl definition). Moreover, we can extend the ideas to getting an abstract definition of explanation that can be applied beyond causal models. Finally, we gain a deeper understanding of features of the definition even in causal models.

1 Introduction

Perhaps the most prominent current definition of (actual) causality is due to Halpern and Pearl [7, 8] (HP from now on). It is defined using causal models (also known as *structural equations models*). The fact that the definition is so tied to causal models has benefits—for example, the fact that causal models can be represented graphically makes them relatively easy to work with—but also leads to a number of limitations; specifically, the language in which we can express counterfactuals is limited (so we cannot express causal statements that involve more complex counterfactuals), and we cannot deal with *backtracking*. (See below for further discussion.) The goal of this paper is to abstract the definition of causality in causal models, extracting its key features, so that it can be applied to any other model where counterfactuals are defined.

Doing so lets us get around these limitations. For one thing, it lets us consider more general causal statements. The HP definition of "A is a cause of B" in causal models can be applied only when A is a conjunction of primitive events $X = x$, where X is a variable and x is a value in the range of X, and B is a Boolean combination of primitive events. It cannot be applied if A is a disjunction, or if either A or B involves modalities like belief or nested counterfactuals. But such statements are important in practice! For example, Sartorio [19] gives examples where disjunctive causes play a significant role; Soloviev and Halpern [20] give examples where nested counterfactuals are critical for expressing security properties. And we clearly want to be able to reason about (and intervene on) agent's beliefs. For example, we might want to say "If were to intervene on Alice's beliefs so that she believed that the vaccine was effective, then she would take it".

Causal models have been extended to allow counterfactuals that involve disjunctions and nested counterfactuals [2], and reasoning about both knowledge/belief and causality [1] (although not interventions on beliefs). These extensions involving disjunction and nested counterfactuals are not at all straightforward, and lead to some arguably unreasonable axioms. For example, the axiom given by Briggs that characterizes interventions involving disjunctions is $[A \vee A']B \Leftrightarrow ([A]B \wedge [A']B \wedge [A \wedge A']B)$.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 160–174, doi:10.4204/EPTCS.437.16

It is not clear why this is appropriate. Briggs does not try to justify it and, indeed, admits that the axiom validates "the controversial rule *Simplification of Disjunctive Antecedents* ... [which] most authors reject". Dealing with interventions on beliefs also seems quite nontrivial. All this means that a definition of causality involving these constructs may well be controversial. The advantage of the approach that we are suggesting here is that, as long as we start with a framework that has a definition of counterfactuals that a user is comfortable with and supports these constructs, we can essentially import the HP definition to that framework. In particular, we get a definition of causality in the counterfactual structures considered by Stalnaker [21] and Lewis [14] that are standard in the philosophy literature. Moreover, we can show that this definition generalizes that in causal models: as shown by Halpern [6], causal models correspond to a subset of counterfactuals; we show here (see Theorem 4.5) that, in a causal model M, the original definition of causality agrees with the new definition in the counterfactual structures M' corresponding to M.

As we said above, the second key limitation of defining causality in causal models is that we cannot deal with backtracking. As has often been observed, when considering the effect of an intervention such as $X = x$ in a causal model, no variables "upstream" of X in the causal graph are affected; only descendants of X can be affected. This is referred to as "no backtracking".

The argument for no backtracking is not hard to explain. If we intervene to set, say $X = 2$, then, intuitively, all that should change are the values of variables that are descendants of X, since these are the only ones that can be affected by the intervention. But, as argued by von Kügelen et al. [13], there are cases where we want to ascribe causality but do not want to consider counterfactuals that result from interventions. Intuitively, these are cases where the causal laws, not the background conditions, are shared between the actual and counterfactual worlds. Consequently, the "upstream" variables must be allowed to differ to accommodate possibly contradictory facts. As discussed by von Kügelen et al. (who also give references to work in the cognitive psychology literature on the topic), people certainly consider backtracking counterfactuals when evaluating causality. This makes it useful to have a definition of causality that allows them. Causal models have been extended so as to deal with backtracking [13, 16], but the definitions of causality used are quite different from the HP definition. Our approach lets us import the HP definition "for free", so to speak, as long as we start with a definition of counterfactual that allows backtracking (again, as do the standard definitions of Stalnaker [21] and Lewis [14]).

This approach can be further extended to give an abstract definition of *explanation*. There are many definitions of explanations in the literature. The modern literature goes back to the work of Hempel [10] and Salmon [17], but this work is well known to have problems because it does not take causality into account. Here we focus on causal explanations, again taking as our basis the work of Halpern and Pearl [7, 9]. The definition is based on the defintion of causality (which is what will allow us to apply the approach of this paper), but it also takes into account the well-known observation that what counts as an explanation is relative to what an agent knows [3, 18]. As Gärdenfors [3] observes, an agent seeking an explanation of why Mr. Johansson has been taken ill with lung cancer will not consider the fact that he worked for years in asbestos manufacturing an explanation if he already knew this fact. It is relatively straightforward to take an agent's knowledge into account in our abstract approach. Doing so gives us the tools we need to handle explanation as well.

Our Definition in a Nutshell The key intuition for the definition is the classic notion of *but-for* causality used in the law literature: A is a cause of B if, but for A, B would not have happened. As is well known (see Example 2.2), but-for causality does not suffice to give a definition of causality. There are times we would like to call A a cause of B even if A is not a but-for cause of B. The definition of actual

causality in causal models and our new abstract definition can both be viewed as essentially saying that A is a cause of B if, conditional on some appropriate C, A is a but-for cause of B. Like the definition of [7], we just require C to be a formula that is true (in the current state of the world).

In more detail, our notion says that *A is an actual cause of B* with respect to a language $\mathscr{L}$ if, roughly speaking, the following three conditions (which are formalized in Section 4) hold:

AC1′. A and B both hold (in the true state of the world);

AC2′. There is a formula $C \in \mathscr{L}$ such that C holds (in the true state of the world), and if $\neg A \wedge C$ were to hold (counterfactually) then $\neg B$ would hold.

AC3′. A is minimal; that is, there is no formula A' that implies A yet is not implied by A that also satisfies AC2′.

Note that the definition above is parametrized by a language $\mathscr{L}$. As we show, in case this language $\mathscr{L}$ consists only of conjunctions of primitive events (and A, B are also restricted to this language), then our definition collapses down to the HP definition of causality (see Theorem 4.2). Furthermore, if the language is slightly richer, then the same is true even in *recursive counterfactual structures*, counterfactual structures that essentially capture causal models (see Theorem 4.5).

Paper Outline The rest of the paper is organized as follows. In Section 2, we review the HP definition of actual causality, as modified by Halpern [7]. In Section 3, we discuss *causal-counterfactual families (of models)—ccfs*, a general class of frameworks to which our abstract definition applies. We show that causal models and the counterfactual structures considered in the philosophy literature are instances of ccfs. In Section 4, we give our abstract definition of causality, show that the HP definition is a special case of it, and consider its implications in counterfactual structures. We consider backtracking in Section 5, and explanation in Section **??**. Proofs are deferred to the appendix.

2 Actual causality in causal models

Here we briefly review the definition of causal models introduced by Halpern and Pearl [8], and the definition of actual causality in causal models. In fact, three variants of the definition have been proposed, called the "original", "updated", and "modified" definitions [7]. We focus here on the modified definition, both because it is the simplest to state and, as shown by Halpern [7], the easiest to work with, and because it lends itself naturally to generalization. The material in the first three subsections of this section is largely taken from [7]. While there are a number of other definitions of causality in causal models (e.g., [4, 5, 11, 12, 22]), we focus on the HP definition here because our generalization is based on it and because it has been the most influential (judging by Google scholar citations). We believe that our approach should be applicable to other ways of defining causality, but we have not checked details.

2.1 Causal models

We assume that the state is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how the latter values are determined.

Formally, a *causal model M* is a pair $(\mathscr{S}, \mathscr{F})$, where $\mathscr{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathscr{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathscr{S}$ is a tuple $(\mathscr{U}, \mathscr{V}, \mathscr{R})$,

where $\mathscr{U}$ is a set of exogenous variables, $\mathscr{V}$ is a set of endogenous variables, and $\mathscr{R}$ associates with every variable $X \in \mathscr{U} \cup \mathscr{V}$ a nonempty set $\mathscr{R}(X)$ of possible values for X (i.e., the set of values over which X *ranges*). For simplicity, we assume here that $\mathscr{V}$ is finite, as is $\mathscr{R}(X)$ for every endogenous variable $X \in \mathscr{V}$. $\mathscr{F}$ associates with each endogenous variable $X \in \mathscr{V}$ a function denoted F_X (i.e., $F_X = \mathscr{F}(X)$) such that $F_X : (\times_{U \in \mathscr{U}} \mathscr{R}(U)) \times (\times_{Y \in \mathscr{V} - \{X\}} \mathscr{R}(Y)) \to \mathscr{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathscr{U} \cup \mathscr{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u,y,z) = u+y$, which is usually written as $X = U+Y$. Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

The structural equations define what happens in the presence of external interventions. Setting the value of some variables $\vec{X}$ to $\vec{x}$ in a causal model $M = (\mathscr{S}, \mathscr{F})$ results in a new causal model, denoted $M_{\vec{X} \leftarrow \vec{x}}$, which is identical to M, except that, for each $X \in \vec{X}$, the equation in $\mathscr{F}$ for X is replaced by $X = x$, where x is the component of X in $\vec{x}$.

The dependencies between variables in a causal model $M = ((\mathscr{U}, \mathscr{V}, \mathscr{R}), \mathscr{F})$ can be described using a *causal network* (or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in M, with one node for each variable in $\mathscr{U} \cup \mathscr{V}$. The roots of the graph are (labeled by) the exogenous variables. There is a directed edge from variable X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathscr{U} \cup \mathscr{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x,\vec{z}) \neq F_Y(x',\vec{z})$. A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. If M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathscr{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). We can determine these values by starting at the top of the graph and working our way down. In this paper, following most of the literature, we restrict to recursive models.

2.2 Reasoning about causality

Given a signature $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathscr{V}$ and $x \in \mathscr{R}(X)$. A *basic causal formula (over $\mathscr{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathscr{V}$, and $y_i \in \mathscr{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$. A *causal formula* is a Boolean combination of basic causal formulas. Let $\mathscr{L}(\mathscr{S})$ consist of all causal formulas over the signature $\mathscr{S}$. (We typically omit $\mathscr{S}$ when it is clear from context.)

A pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ is a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values for the variables that simultaneously satisfies all equations in M with the variables in $\mathscr{U}$ set to $\vec{u}$); $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} = \vec{y}}, \vec{u}) \models \varphi$; and Boolean combinations are defined in the standard way.

2.3 The modified definition of actual causality

We are now in a position to state the (modified) definition of actual causality in causal models.

Definition 2.1: $\vec{X} = \vec{x}$ is an *actual cause of φ in the causal setting* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$.

AC2. There is a set $\vec{W}$ of variables in $\mathcal{V}$ disjoint from $\vec{X}$, $\vec{w}^* \in \mathcal{R}(\vec{W})$, and $\vec{x} \in \mathcal{R}(\vec{X})$ such that $(M,\vec{u}) \models \vec{W} = \vec{w}^*$ and $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$.[1]

AC3. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies condition AC2, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.[2]

∎

2.4 The role of AC2

The motivation for AC2 was to be able to deal with examples like the following, due to Lewis [15]. (The following analysis is largely taken from [7].)

Example 2.2: Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

We would like to call Suzy's throw the cause of the bottle shattering. This is the case in an appropriate causal model. The following causal model M^{rt} does the job. There are five endogenous variables:

- *ST* for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- *BT* for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- *BS* for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).
- *BH* for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- *SH* for "Suzy's rock hits the bottle", again with values 0 and 1.

For simplicity, assume that there is one exogenous variable U, with values ij, $i,j \in \{0,1\}$, which determines whether Billy and Suzy throw.

We consider the following equations for the variables:

- $ST = 1$ iff $U \in \{10,11\}$;
- $BT = 1$ iff $U \in \{01,11\}$;
- $SH = 1$ iff $ST = 1$;
- $BH = 1$ iff $BT = 1$ and $SH = 0$ (this builds into the equation the fact that Billy rock does not hit the intact bottle if Suzy's rock already hit it);
- $BS = 1$ iff $SH = 1$ or $BH = 1$.

Now consider the context $U = 11$, where both Suzy and Billy throw rocks. In this context, most people want to view $ST = 1$ as a cause of the bottle shattering, while $BT = 1$ is not. But $ST = 1$ is not a but-for cause of the bottle shattering; had Suzy not thrown, the bottle would have shattered anyway (Billy's rock would have hit it). Nevertheless, it is a cause. To show that $ST = 1$ is a cause, we take W in AC2 to be $\{BH\}$, and fix BH at its actual value of 0 (in fact, Billy 's throw does not hit the bottle). Note that $(M,11) \models [ST \leftarrow 0, BH \leftarrow 0]BS = 0$, so AC2 holds for $ST = 0$; clearly AC1 and AC3 hold as well. Thus, $ST = 1$ is a cause of $BS = 1$ in the context $U = 11$. It is easy to see that the symmetric argument does *not* show that $BT = 1$ is a cause. We cannot fix $SH = 0$, because in the actual state, SH is 1, not 0. We can think of $ST = 1$ as a but-for cause of $BS = 1$, conditional on $BH = 0$. ∎

[1] Although the definition in [7] does not explicitly say that $\vec{W}$ is disjoint from $\vec{X}$, since interventions are performed on distinct variables, this follows from the notation.

[2] Note that $\vec{X}' = \vec{x}'$ is guaranteed to satisfy AC1, which is why only AC2 is mentioned here.

The idea of fixing some variables at their actual values, as embodied by AC2, seems at first somewhat *ad hoc*. While it does give us the "right" answer in the example above, why should we want to do it in general? The idea of keeping some variables fixed at their actual value turns out to be quite a powerful one. This is perhaps best seen if we consider causality along a path. To take a simple example, suppose that someone's gender affects the outcome of a loan decision in two ways. First, gender affects salary, which in turns affects the decision. Second, gender affects how reliable someone is viewed as being (say, for example, that men are viewed as less likely to repay loans, even fixing everything else). That is, we have a simple causal graph with variables G (gender) that affects both S (salary) and R (reliability), while S and R both affect L (loan repayment).

We might decide that it is legitimate to take salary into account when considering loan repayment, but not the perception of reliability. That is, we might decide that it is legitimate to consider the effect of gender along the path $G-S-L$, but not along the path $G-R-L$. We can model causality along a path by simply fixing the values of variables at their actual values off the path. In this case, since we do not want to consider the impact of perception of reliability, we fix R at its actual value, which allows us to consider the impact of changing G along the path $G-S-L$.[3]

Fixing the values of some variables $\vec{W}$ is useful beyond being able to capture path causality. It allows us to consider causality conditional on certain features of the state being kept fixed—the key feature of our more abstract generalized definition.

3 Causal-counterfactual families

We want to define causality for all families of structures that satisfy certain minimal conditions. To that end, we consider (for lack of a better name) *causal-counterfactual families (ccfs)*. A ccf $\mathscr{M}$ is a family of models or structures (we use the latter two words more or less interchangeably) such that each model $M \in \mathscr{M}$ has a set of possible *states* or worlds, and there is a semantics that allows us to define $(M,w) \models \varphi$ for all formulas in a language $\mathscr{L}^{\mathscr{M}}$. We require that $\mathscr{L}^{\mathscr{M}}$ include primitive events in some signature $\mathscr{S}$, and be closed under conjunction and negation (so contains $\mathscr{L}(\mathscr{S})$) and *basic counterfactuals*, so that if φ and ψ are in $\mathscr{L}(\mathscr{S})$, then $\varphi \rightarrow \psi \in \mathscr{L}^{\mathscr{M}}$. But $\mathscr{L}^{\mathscr{M}}$ may also be far more expressive and include, for example, nested counterfactuals, and modalities such as belief and time. The richness of $\mathscr{L}^{\mathscr{M}}$ is exactly what will allow our abstract definition of causality to apply to quite expressive languages.

3.1 Causal models as a ccf

We can almost view causal models (over $\mathscr{S}$) as a ccf $\mathscr{M}$. The states are the contexts. Let the language $\mathscr{L}^+(\mathscr{S})$ be the minimal language allowed, namely, the result of starting with $\mathscr{L}(\mathscr{S})$ and then extending it to include formulas of the form $\varphi \rightarrow \psi$, where φ and ψ are in $\mathscr{L}(\mathscr{S})$. (Thus, $\mathscr{L}^+(\mathscr{S})$ has nested counterfactuals, while $\mathscr{L}(\mathscr{S})$ does not.) We take $\mathscr{L}^{\mathscr{M}} = \mathscr{L}^+(\mathscr{S})$.

We might hope to identify $\varphi \rightarrow \psi$ with $[\varphi]\psi$. This works fine as long as φ is a conjunction of primitive events. But $[\varphi]\psi$ is not defined in causal models for an arbitrary Boolean combination φ of primitive events. So, in order to view causal models as a ccf, we have to define the semantics of $\varphi \rightarrow \psi$ for an arbitrary Boolean combination φ of primitive events. To give semantics to a counterfactual of the form $\varphi \rightarrow \psi \in \mathscr{L}^+(\mathscr{S})$, we proceed as follows: Let $\vec{Y}$ be the endogenous variables that appear in φ. Then we take $(M,\vec{u}) \models \varphi \rightarrow \psi$ iff there exists a vector $\vec{y}$ of values such that $\varphi \wedge \vec{Y} = \vec{y}$ is consistent (when viewed as a propositional formula) and $(M,\vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\psi$. It is easy to check that $\varphi \rightarrow \psi$ agrees

[3] Thanks to Sander Beckers for pointing this out to us.

with $[\varphi]\psi$ if φ is a conjunction of primitive events, so this is a generalization of the standard definition of counterfactuals in causal models.

This definition seems somewhat *ad hoc* and, indeed, has some counter-intuitive properties (see Example 4.4). For example, why restrict $\vec{Y}$ to consisting only of variables that appear in φ? Indeed, there are other definitions of counterfactuals in causal models that agree with $[\varphi]\psi$ if φ is a conjunction of primitive propositions. For example, we could take $\varphi \rightarrow \psi$ to be $[\varphi]\psi$ if φ is a conjunction of primitive propositions, and *false* otherwise. The current definition has the benefit of allowing us to prove that the generalized definition of causality agrees with the HP definition in causal models (see Theorem 4.2).

3.2 Counterfactual structures as a ccf

For our purposes, a (Lewis-style) *counterfactual structure* M over $\mathscr{S}$ is a tuple (S,R,π), where S is a finite set of states, π is an *interpretation* that maps each state to a truth assignment on the primitive events over $\mathscr{S}$, and R is a ternary relation over S. Intuitively, $(s,t,u) \in R$ if t is at least as close to s as u is.[4] Let $t \preceq_s u$ be an abbreviation for $(s,t,u) \in R$. Define $t \prec_s u$ if $t \preceq_s u$ and $u \not\preceq_s t$. For simplicity, we require that $\preceq_s$ be reflexive and transitive and assume that $s \prec_s t$ for all $t \neq s \in S$ (so, intuitively, s is the unique state closest to itself according to $\preceq_s$). These are standard assumptions in the literature.

As for the language, we again fix a signature $\mathscr{S}$. Taking $\mathscr{M}^c(\mathscr{S})$ to be the counterfactual structures over signature $\mathscr{S}$ (so that the interpretation π in these structures gives semantics to primitive events over $\mathscr{S}$), we assume that $\mathscr{L}^{\mathscr{M}^c(\mathscr{S})}$ includes $\mathscr{L}(\mathscr{S})$ and is closed under conjunction, negation, counterfactuals (so that if $\varphi,\psi \in \mathscr{L}^{\mathscr{M}^c(\mathscr{S})}$, then so is $\varphi \rightarrow \psi$, which means that we have nested counterfactuals). As usual, given a counterfactual structure $M \in \mathscr{M}^c(\mathscr{S})$ and state s, we define $(M,s) \models \varphi \rightarrow \psi$ if for all states t such that $(M,t) \models \varphi$ and there is no state t' such that $(M,t') \models \varphi$ and $t' \prec_s t$, we also have $(M,t) \models \psi$. That is, $(M,s) \models \varphi \rightarrow \psi$ if in all the closest states to s satisfying φ, ψ also holds. (It follows that $(M,s) \models \varphi \rightarrow \psi$ vacuously if there are no states t such that $(M,t) \models \varphi$.)

The careful reader will have noticed that this means we are treating formulas with negations differently in counterfactual structures and in causal models. For example, in the former case, the formula $(X \neq x) \rightarrow \psi$ is taken to be true at a state s if at the closest state to s such that $X \neq x$, ψ is true. In the latter case, the formula is true if, roughly speaking, there exists some value x' such that at the closest state where $X = x'$, ψ is true. If X is a binary variable and takes on only two values, say x and x', at the closest state where where $X \neq x$, we must have $X = x'$, and the two approaches agree. But if X is not binary, the two approaches may not agree. As we shall see, this distinction disappears given how we deal with causality.

4 A more general abstract definition of actual causality

We now define actual causality in arbitrary ccfs. While the focus of this paper is on causal models and counterfactual structures, we stress that the definition applies far more broadly. For example, rather than Lewis-style counterfactual structures, we could consider Stalnaker-style counterfactual structures [21], where there is a function f that associates with each state s and formula φ a unique state $f(s,\varphi)$ closest to s that satisfies φ. We can also apply the definition to subclasses of counterfactual structures (and do), alternate definitions of causal models, and the extended causal models of Lucas and Kemp [16]. In the definition, the set of possible witnesses is a parameter. As we shall see, this provides useful flexibility.

[4] The presentation here follows that of Halpern [6], with some simplifications for ease of exposition. Although we focus on Lewis's definition [14], largely the same approach handles that of Stalnaker [21], who assumes that there is a unique closest state, while Lewis allows for a set of closest states.

Definition 4.1: Given a ccf $\mathscr{M}$ and set $\mathscr{C}_{\mathscr{X}} \subseteq \mathscr{L}^{\mathscr{M}}$ of formulas, φ is an *actual cause of ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in a model $M \in \mathscr{M}$ and state s* if the following three conditions hold:

AC1′. $(M,s) \models \varphi \wedge \psi$;

AC2′. There is a formula $\tau \in \mathscr{C}_{\mathscr{X}}$ such that $(M,s) \models \tau$ and $(M,s) \models (\neg\varphi \wedge \tau) \rightarrow \neg\psi$.

AC3′. φ is minimal; there is no formula $\varphi' \in \mathscr{C}_{\mathscr{X}}$ such that $\models \varphi \Rightarrow \varphi'$ and $\not\models \varphi' \Rightarrow \varphi$ such that AC2′ holds with φ replaced by φ'. ∎

4.1 Applying the abstract definition in actual causal models

This definition is, by design, quite similar to the definition of actual causality in causal models, but more general. Causes no longer have to have the form $\vec{X} = \vec{x}$, and we allow the witnesses in $\vec{W}$ to come from some arbitrary set $\mathscr{C}_{\vec{X}}$ of formulas that may depend on $\vec{X}$. But, as we observed above, there is a nontrivial difference in how negation is dealt with. As we hinted above, despite this difference, in causal models, we defined $\rightarrow$ so that AC2 and AC2′ lead to identical conclusions (with the appropriate choice of $\mathscr{C}_{\vec{X}}$).

Theorem 4.2: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of arbitrary (non-negated) primitive events (so $\mathscr{C}_{\vec{X}}$ is in fact independent of $\vec{X}$). Then $\vec{X} = \vec{x}$ is a cause of ψ in a causal setting $(M,\vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in $(M,\vec{u})$ according to AC1′-3′.*

The choice of $\mathscr{C}_{\vec{X}}$ in Theorem 4.2 ensures that, just as in AC2, in whatever corresponds to the "closest states s' to s where $\vec{X} \neq \vec{x} \wedge \tau$ is true, a formula of the form $\vec{X} = x' \wedge \tau$ is true, where τ is the conjunction of all the primitive events in τ (but does not involve the negations of primitive events in τ). It easily follows that $(M,s) \models \tau$, which shows that AC2′ ends up with a formula much like that in AC2.

It is also worth noting that if $\mathscr{C}'_{\vec{X}} \supseteq \mathscr{C}_{\vec{X}}$, then the forward direction of Theorem 4.2 holds, but the converse may not. Theorem 4.2 does continue to hold if we allow negated primitive events as conjuncts, but once we allow disjunctions of primitive events, the converse fails in general. For example, if τ is $SH = 0 \vee BH = 0$ in Example 2.2, then clearly $(M^{rt}, U_{11}) \models \tau$. Moreover, $BT = 0 \wedge SH = 0 \wedge BH = 0$ is consistent with $BT \neq 1 \wedge \tau$, and $(M^{rt}, U_{11}) \models [BT \leftarrow 0, SH \leftarrow 0, BH \leftarrow 0](BS = 0)$, so with this choice of $\mathscr{C}'_{\mathscr{X}}$, $BT = 1$ would be a cause of $BS = 1$.

This example (and Theorem 4.2, for that matter) very much depend on the fact that $\vec{Y}$ consists precisely of the variables that appear on the left of the $\rightarrow$.

4.2 Applying the abstract definition of causality in recursive counterfactual structures

As we now show, the abstract definition of causality, when applied to a subfamily of counterfactual structures, can be viewed as generalizing the definition in causal structures. The subfamily of counterfactual structures corresponds in a precise sense to recursive causal models. Formally, a counterfactual structure M' *corresponds* to a causal model M if M and M' agree on all the equations; more precisely, for each endogenous variable Y, if $S_Y = (\mathscr{U} \cup \mathscr{V}) - \{Y\}$, then for each setting $\vec{s}_Y$ of the variables in S_Y, and each state s, in the closest states to s such that $S_Y = \vec{s}_Y$ holds, we have that $Y = F_Y(\vec{s}_Y)$. M' *strongly corresponds* to M if (a) M' corresponds to M, (b) for each assignment v of values to variables, there is a state in M' where v is the assignment, and (c) for each state s, if $(M',s) \models \mathscr{U} = \vec{u}$ and ψ is a propositional formula such that $\mathscr{U} = \vec{u} \wedge \psi$ is consistent, then in the closest states to s such that ψ holds, we have that $\mathscr{U} = \vec{u}$.[5] Roughly speaking, $\mathscr{U} = \vec{u}$ continues to hold in the closest states to s unless some value in $\mathscr{U}$

[5] Although formulas of the form $\mathscr{U} = \vec{u}$ are not in the language $\mathscr{L}(\mathscr{S})$, for the purposes of this discussion, we assume that we have extended the language to include them, with the obvious semantics.

is explicitly changed. (M',s) is (strongly) consistent with $(M,\vec{u})$ if M' (strongly) corresponds to M and $(M',s) \models \mathscr{U} = \vec{u}$.

A counterfactual structure is *recursive* if it strongly corresponds to a recursive causal model. In general, there may be many recursive counterfactual structures that strongly correspond to a given recursive causal model; the definition of strong consistency does not specify how the closest-state relation R in the counterfactual structure should work for formulas that are not structural equations. Nevertheless, as the following result shows, recursive counterfactual structures are closely related to (recursive) causal models, at least as far as counterfactual reasoning is concerned. We say that (M',s) is a *recursive setting* if M' is a recursive counterfactual structure and s is a state in M'.

Proposition 4.3: *[6, Theorem 3.4] If $(M,\vec{u})$ is a recursive causal setting and (M',s) is a recursive setting that is strongly consistent with $(M,\vec{u})$, then for all formulas $\varphi \in \mathscr{L}(\mathscr{S})$, $(M,\vec{u}) \models \varphi$ iff $(M',s) \models \varphi$ (where we identify $\psi \rightarrow \psi'$ with $[\psi]\psi'$).*

The restriction to $\mathscr{L}(\mathscr{S})$ is critical here, as the following example shows.

Example 4.4: Consider a recursive causal model M with two endogenous variables, X and Y, and one exogenous variable U, such that $\mathscr{R}(U) = \mathscr{R}(X) = \mathscr{R}(Y) = \{0,1,2\}$. Suppose that U is the unique parent of X, and X is the parent of Y; the equation for X is $X = U$, and the equation for Y is $Y = X$. Let ψ be the formula $(X \neq 0 \rightarrow Y = 1) \wedge (X \neq 0 \rightarrow Y = 2)$. Note that $\psi \notin \mathscr{L}(\mathscr{S})$ due to the antecedent $X \neq 0$. It is easy to see that $(M,U = 0) \models \psi$, but there can be no counterfactual structure M' and state s such that $(M',s) \models \psi$. ∎

Despite this example, we have the following result (whose proof is deferred to the appendix).

Theorem 4.5: *Let $\vec{X}$ be a set of endogenous variables, and let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X} = \vec{x} \vee \vec{X} = \vec{x}^*$, where $\vec{x}^* \neq \vec{x}$. If $(M,\vec{u})$ is a recursive causal setting and (M',s) is a recursive counterfactual structure that is strongly consistent with $(M,\vec{u})$ then $\vec{X} = \vec{x}$ is a cause of φ in $(M,\vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of φ with respect to $\mathscr{C}_{\vec{X}}$ in (M',s) according to AC1'-3'.*

Note that $\mathscr{C}_{\vec{X}}$ differs in Theorems 4.2 and 4.5. These results show the power of allowing $\mathscr{C}_{\vec{X}}$ as a parameter.

On the Role of the Language: We emphasize that for Theorem 4.5 to hold, we must allow the set $\mathscr{C}_{\vec{X}}$ of formulas that we condition on to contain (a quite restricted form of) disjunctions. If we were to allow arbitrary disjunctions, in particular, disjunctions that contain $\neg\varphi$, then any formula would become a cause of φ, and the definition becomes trivial. What happens if we restrict $\mathscr{C}_{\vec{X}}$ to contain only a conjunction of primitive events? Then our definition and the HP definition would diverge: Suppose that Bob is placed in a room with a ticking bomb, which can be defused by entering a secret 100-digit combination. Bob can decide to run away, or try pushing in a number combination. If Bob runs away and the bomb explodes, should his action of running away be considered the cause of it? According to the HP definition, it would (since if Bob had just entered the right combination, then the bomb would be defused). According to our definition (and disallowing disjunctions and negations in $\mathscr{C}_{\vec{X}}$), it would depend on what number combination Bob would enter in the closest world where he does not run away. For instance, if Bob knows the right combination, then him running away would be considered a cause, but if he does not, then it would not. Arguably, the new definition (without disjunctions) gives the more reasonable answer in this situation. On the other hand, if the number combination is short (say, 1 digit), then arguably the HP definition would give the right answer whereas, without disjunctions, our new definition would not.

5 Backtracking

As has often been observed, when considering the effect of an intervention such as $X = x$ in a causal model, no variables "upstream" of X in the causal network (more precisely, no variables that are not descendants of X) are affected; only descendants of X can be affected. By way of contrast, even in recursive counterfactual structures, in the state(s) closest to (M,s) where $X = x$ is true, it may well be the case that ancestors of X have changed their value. Such backtracking is allowed in Lewis-style models, when considering closest states. Kügelgen et al. [13] introduced a whole mechanism for allowing backtracking counterfactuals in causal models that involved changing the values of exogenous variables. Note that changing the values of an exogenous variable is not viewed as an intervention; rather, it is assumed that the agent considering whether $\vec{X} = \vec{x}$ is a cause of φ is uncertain about the value of the exogenous variable, and is considering what would happen to φ if the value of the exogenous variable were different. Here we show that our abstract definition of causality allows backtracking in a natural way, without having to introduce a special mechanism for it. Indeed, we can move smoothly from allowing backtracking to forbidding it, just by fixing an appropriate set of variables.

It follows from Proposition 4.5 that we have no backtracking in recursive counterfactual structures that strongly correspond to causal models. But once we allow the context to change when considering causality, we can get backtracking. This is exactly what happens when we move from counterfactual structures that strongly correspond to a causal model to ones that just correspond to a causal model.

Example 5.1: Recall the causal setting $(M^{rt}, U = 11)$ for Example 2.2, the rock-throwing example. Let $M' = (S,R,\pi)$ be consistent with M^{rt}. Let s be a state such that (M',s) is consistent with $(M^{rt}, U = 11)$; in particular, $(M',s) \models ST = BT = SH = BS = 1 \wedge BH = 0$. Define R so that the closest state s' to s is such that $(M',s') \models U = 00$, so that $(M',s') \models ST = BT = SH = BH = BS = 0$. With these choices, M' does not strongly correspond to M. It is easy to see that $(M',s) \models BT = 0 \rightarrow BS = 0$, so $BT = 1$ is a cause of $BS = 1$ in (M',s). By allowing the context to change in the closest state to s, we get backtracking. ∎

By working with a counterfactual setting (M',s) that corresponds to a causal setting $(M,\vec{u})$, we can capture causality with and without backtracking in $(M,\vec{u})$, simply by adding the appropriate conjuncts to the formula τ in AC2′. For example, if we want capture why $\vec{X} = \vec{x}$ is a cause of φ without backtracking, we simply add the conjunct $\mathscr{U} = \vec{u}$ to ψ. If we want to allow some backtracking, we may want to fix the values of some ancestors of $\vec{X}$ to their actual value, without fixing $\mathscr{U}$. This makes it clear that whether or not we have backtracking depends on the set of formulas $\mathscr{C}_{\mathscr{X}}$ from which τ is drawn.

6 An Abstract Definition of Explanation

Our approach can be extended to deal with causal explanation, using the HP definition [7, 9]. We start by reviewing the HP definition of explanation; for more details and intuition, see [7]. As noted in the introduction, the HP definition of causality is relative to an agent and, specifically, what the agent knows. In causal models, an agent's knowledge is captured by the agent's *epistemic state* $\mathscr{K}$, where $\mathscr{K}$ is a set of contexts settings with a fixed causal model M.[6] Intuitively, the contexts in $\mathscr{K}$ are the ones that the agent considers possible, and reflects the agent's uncertainty regarding how the world works (represented by the

[6] In [7] there is also assumed to be a probability on $\mathscr{K}$; for simplicity, we ignore the probability here. Also, in [7], an agent may be uncertain about the model as well as the context. Assuming that the model is fixed, as we are doing here, is without loss of generality. If an agent is uncertain about the model, that uncertainty can be encoded into the context by having a special exogenous variable u^* that determines the model; that is, the value of u^* gives the model. Given a model M, the equations in M can be encoded using equations that hold when $u^* = M$.

equations involving context $\vec{u}$) and what is currently true (represented by the context $\vec{u}$). As is standard, we say that an agent *knows* φ if φ is true at all the settings in $\mathscr{K}$.

We next give the formal definition of explanation in causal models, and then give intuition for the clauses, particularly EX1(a).

Definition 6.1: $\vec{X} = \vec{x}$ *is an explanation of* ψ *in a model* M *relative to a set* $\mathscr{K}$ *of contexts in* M if the following conditions hold:

EX1(a). If $\vec{u} \in \mathscr{K}$ and $(M, \vec{u}) \models \vec{X} = \vec{x} \wedge \psi$, then there exists a conjunct $X = x$ of $\vec{X} = \vec{x}$ and a (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is a cause of ψ in $(M, \vec{u})$.

EX1(b). $(M, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\psi$ for all contexts $\vec{u}' \in \mathscr{K}$.

EX2. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies EX1(a) and EX1(b), where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.

EX3. For some $\vec{u} \in \mathscr{K}$, we have that $(M, \vec{u})) \models \vec{X} = \vec{x} \wedge \psi$. (The agent considers possible a setting where the explanation and explanandum both hold.)

The explanation is *nontrivial* if it also satisfies

EX4. $(M, \vec{u}') \models \neg(\vec{X} = \vec{x}) \wedge \psi$ for some $\vec{u}' \in \mathscr{K}$ (The explanation is not already known given the observation of ψ.) ∎

The key part of the definition is EX1(b). Roughly speaking, it says that the explanation $\vec{X} = \vec{x}$ is a *sufficient cause* for ψ: for all settings that the agent considers possible, intervening to set $\vec{X}$ to $\vec{x}$ results in ψ. (See [7, Chapter 2.6] for a formal definition of sufficient cause.) EX2 and EX3 should be fairly clear. EX4 is meant to capture the intuition (discussed in the introduction) that what counts as an explanation depends on what the agent knows. In the formal model, this is captured by the set $\mathscr{K}$. That leaves EX1(a). Roughly speaking, it says that the explanation causes the explanandum. But there is a tension between EX1(a) and EX1(b) here: we may need to add conjuncts to the explanation to ensure that it suffices to make ψ true in all contexts (as required by EX1(b)). But these extra conjuncts may not be necessary to get causality in all contexts. At least one of the conjuncts of $\vec{X} = \vec{x}$ must be part of a cause of ψ, but the cause can include extra conjuncts. To understand why, it is perhaps best to look at an example.

Example 6.2: Consider a version of the rock-throwing example where now in the causal model M it is possible that (a) Suzy's rock gets to the bottle first, (b) Billy's rock gets to the bottle first, or (c) both rocks arrive simultaneously. We capture this using 12 contexts of the form u_{ijk} where, as before $i = 1$ if Suzy throws and $j = 1$ if Billy throws. If $k = 1$ (resp., $k = 2$, $k = 3$) and both Billy and Suzy throw, then Suzy's rock gets to the bottle first (resp., Billy's rock gets to the bottle first; the rocks arrive simultaneously). In all cases, just one rock hitting the bottle suffices for the bottle to shatter.

Consider $\mathscr{K}_1 = \{u_{111}, u_{112}, u_{101}\}$. An explanation for the bottle shattering relative to $\mathscr{K}_1$ is $ST = 1 \wedge BT = 1$. Each of $ST = 1$ and $BT = 1$ satisfies EX1(b). But while $ST = 1$ is a cause of the bottle shattering in $(M, \vec{u}_{111})$ and $(M, \vec{u}_{101})$, it is not a cause of the bottle shattering in $(M, \vec{u}_{111})$, while $BT = 1$ is. On the other hand, $BT = 1$ is not a cause of the bottle shattering in $(M, \vec{u}_{112})$. Thus, EX1(a) holds for $ST = 1 \wedge BT = 1$, taking $ST = 1$ to be the relevant conjunct in $(M, \vec{u}_{111})$ and $(M, \vec{u}_{111})$ and $BT = 1$ to be the relevant conjunct in $(M, \vec{u}_{111})$. (Note that we do not have to take u_{101} into account for EX1(a), because $(M, \vec{u}_{111}) \not\models ST = 1 \wedge BT = 1$.) It is easy to see that EX2, EX3, and EX4 hold in this case.

Next consider $\mathscr{K}_2 = \{u_{111}, u_{112}\}$. In this case, $ST = 1 \wedge BT = 1$ is an explanation of the bottle shattering relative to $\mathscr{K}_2$, but a trivial one; it was already known.

Finally, consider $\mathscr{K}_3 = \{u_{003}, u_{103}, u_{013}, u_{113}\}$. In this case, each of $ST = 1$ and $BT = 1$ is an explanation of the bottle shattering relative to $\mathscr{K}_3$. However, note that in the context (M, u_{113}), $ST = 1$ is *not*

a cause of the bottle shattering; if Suzy doesn't throw, then the bottle still shatters (since Billy's rock hit the bottle). However, $ST = 1 \wedge BT = 1$ is a cause, and EX1(a) allows adding a conjunct $Y = y$ (in this case, $BT = 1$) to get a cause. Note that $ST = 1 \wedge BT = 1$ is not an explanation of the bottle shattering; it violates the minimality condition EX2. ∎

We now show how to translate each of these conditions to our more abstract framework. To start with, the obvious analogue of $\mathscr{K}$ is just a set of worlds in some model in a ccf.

Definition 6.3: Given a ccf $\mathscr{M}$ and set $\mathscr{C}_{\mathscr{X}} \subseteq \mathscr{L}^{\mathscr{M}}$ of formulas, φ is an *explanation of* ψ *with respect to* $\mathscr{C}_{\mathscr{X}}$ *in a model* $M \in \mathscr{M}$ *and state* s *relative to a set* $\mathscr{K}$ *of states in* M if the following three conditions hold:

EX1(a)′. If $s \in \mathscr{K}$ and $(M,s) \models \varphi \wedge \psi$, then there exist formulas $\tau_1, \tau_2 \in \mathscr{C}_{\mathscr{X}}$ such that (i) τ_1 is not valid, (ii) $\models \varphi \Rightarrow \tau_1$, (iii) $\models \tau_2 \Rightarrow \tau_1$, and (iv) τ_2 is a cause of ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in (M,s) (according to Definition 4.1).

EX1(b)′. $(M,s') \models \varphi \rightarrow \psi$ for all states $s' \in \mathscr{K}$.

EX2′. φ is minimal; there is no formula φ' such that $\models \varphi \Rightarrow \varphi'$ and $\not\models \varphi' \Rightarrow \varphi$ such that EX1(a) and EX1(b) hold with φ replaced by φ'.

EX3′. For some $s \in \mathscr{K}$, we have that $(M,s) \models \varphi \wedge \psi$.

Again, the explanation is nontrivial if

EX4′. $(M,s') \models \neg\varphi \wedge \psi$ for some $s' \in \mathscr{K}$. ∎

All the primed conditions except for EX1(a)′ are fairly direct translations of their unprimed counterparts. For EX1(a)′, note that the effect of adding a conjunction $\vec{Y} = \vec{y}$ to $X = x$ in EX1(a) is to get a formula that implies $X = x$; that is why we have $\models \tau_2 \Rightarrow \tau_1$ in part (iii) or EX1(a)′. Similarly, the conjunct $X = x$ in EX1(a) (which corresponds to τ_1 in EX1(a)′) is implied by $\vec{X} = \vec{x}$ (which correspond to φ in EX1(a)′). To understand the requirement that τ_1 not be valid, note that if it were valid, then conditions (ii) and (iii) would hold vacuously, so (iv) would hold as long as we can find a formula that causes ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in (M,s). We want τ_1 to have a nontrivial connection to φ. In EX1(a), this is accomplished by making it a conjunct of φ. In EX1(a)′, this is accomplished by having it be a nonvalid formula that is implied by φ.

In any case, we now get the analogue of Theorem 4.2. showing that the abstract definition of explanation generalizes the original HP definition.

Theorem 6.4: *Let* $\mathscr{C}_{\vec{X}}$ *consist of all formulas* τ *such that* τ *is a conjunction of arbitrary (non-negated) primitive events, and let* $\mathscr{K}$ *be a set of contexts in a causal model* M*. Then* $\vec{X} = \vec{x}$ *is an explanation of* ψ *in a causal setting* $(M,\vec{u})$ *relative to* $\mathscr{K}$ *according to EX1–3 (resp., EX1–4) iff* $\vec{X} = \vec{x}$ *is an explanation of* ψ *with respect to* $\mathscr{C}_{\vec{X}}$ *relative to* $\mathscr{K}$ *in* $(M,\vec{u})$ *according to EX1′-3′ (resp., EX1′-4′).*

We also get an analogue of Theorem 4.5, but to get this we need an even tighter connection between recursive counterfactual structures and recursive causal models. A recursive causal model M and a recursive counterfactual structure M' are *compatible* if for every context $\vec{u}$ in M there is a state s in M' such that $(M,\vec{u})$ is strongly consistent with (M,s), and for every state s in M' there is a context $\vec{u}$ in M such that $(M,\vec{u})$ is strongly consistent with (M,s). A set $\mathscr{K}$ of contexts in M and a set $\mathscr{K}'$ of states in M' are *compatible* if for every context $\vec{u}$ in $\mathscr{K}$ there is a state s in $\mathscr{K}'$ such that $(M,\vec{u})$ is strongly consistent with (M,s), and for every state s in $\mathscr{K}'$ there is a context $\vec{u}$ in $\mathscr{K}$ such that $(M,\vec{u})$ is strongly consistent with (M,s). Finally, a structure M is *acceptable* if (M,s) is acceptable for all states s in M.

Theorem 6.5: *Let $\vec{X}$ be a set of endogenous variables, and let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X} = \vec{x} \vee \vec{X} = \vec{x}^*$, where $\vec{x}^* \neq \vec{x}$. Given a recursive causal model M and a recursive counterfactual structure M' such that M and M' are compatible, and compatible sets $\mathscr{K}$ of contexts in M and $\mathscr{K}'$ of states in M', if $(M,\vec{u})$ is a recursive causal setting and (M',s) is strongly consistent with $(M,\vec{u})$, then $\vec{X} = \vec{x}$ is an explanation of ψ relative to $\mathscr{K}$ in $(M,\vec{u})$ according to EX1–3 (resp., EX1-4) iff $\vec{X} = \vec{x}$ is an explanation of ψ with respect to $\mathscr{C}_{\vec{X}}$ relative to $\mathscr{K}'$ in (M',s) according to EX1′-3′ (resp., EX1′-4′).*

7 Conclusions

We have given an abstract definition of causality that applies to arbitrary ccfs. In particular, it applies to causal models (where it agrees with the standard HP definition) and to Lewis-style counterfactual structures (where it generalizes the HP definition). As we have observed, having such a general definition lets us apply the definition far more broadly, and to richer, more expressive languages.

Roughly speaking, the abstract definition says that A is a cause of B if, conditional on C, A is a but-for cause of B. We have restricted C to come from some set that we denoted $\mathscr{C}_{\vec{X}}$. We are currently trying to understand the impact of $\mathscr{C}_{\vec{X}}$ on the definition. While in causal models it sufficed to require that $\mathscr{C}_{\vec{X}}$ be a conjunction of primitive events to capture the HP notion of causality, in counterfactual structures we also required $\mathscr{C}_{\vec{X}}$ to contain (restricted) disjunctions. What happens if we were to restrict $\mathscr{C}_{\vec{X}}$ to contain only a conjunction of primitive events? At this point, the two definitions would diverge: Consider a scenario where Bob is placed in a room with a ticking bomb, which can be defused by entering a secret 100-digit combination. Bob can decide to run away, or try pushing in a number combination. If Bob runs away and the bomb explodes, should his action of running away be considered the cause of it? According to the HP definition, it would (since if Bob had just entered the right combination, then the bomb would be defused). According to our definition (and disallowing disjunctions and negations in $\mathscr{C}_{\vec{X}}$), it would depend on what number combination Bob would enter in the closest world where he does not run away. For instance, if Bob knows the right combination, then him running away would be considered a cause, but if he does not, then it would not. Arguably, the new definition (without disjunctions) gives the more reasonable answer in this situation. On the other hand, if the number combination is short (say 1 digit), then arguably the HP definition would give the right answer whereas, without disjunctions, ours would not (even if Bob does not know the secret, assuming that he is allowed multiple trials to enter it).

We can, to some extent, deal with the problem by considering *normality* [7], where, intuitively, we want to consider $\vec{X} = \vec{x}$ a cause of φ only if the formula $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}^*$ in AC2 is not abnormal. As argued in [7], this restriction leads to an arguably more reasonable notion of causality. We are currently exploring the possibility of capturing normality by allowing disjunctions in $\mathscr{C}_X$, but restricting all the disjuncts to be normal, in some appropriate sense, or by restricting to formulas using a small number of disjunctions. Again, this would show that by conditioning on the appropriate set of formulas, we can capture important intuitions.

A Proofs

For the reader's convenience, we repeat the statement of the results.

Theorem 4.2: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of arbitrary (non-negated) primitive events (so $\mathscr{C}_{\vec{X}}$ is in fact independent of $\vec{X}$). Then $\vec{X} = \vec{x}$ is a cause of ψ in a causal setting $(M,\vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in $(M,\vec{u})$ according to AC1′-3′.*

Proof: Suppose that $\vec{X} = \vec{x}$ is a cause of ψ in $(M, \vec{u})$ according to AC1-3. Clearly AC1′ and AC3′ are essentially equivalent to AC1 and AC3, so to show that $\vec{X} = \vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ according to AC1′-3′, it suffices to show that AC2′ holds.

Since AC2 holds, there exist a set $\vec{W}$ of endogenous variables, $\vec{w}^* \in \mathscr{R}(\vec{W})$, and $\vec{x}' \in \mathscr{R}(\vec{X})$ such that $(M, \vec{u}) \models \vec{W} = \vec{w}^*$ and $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*]\neg\psi$. We must have $\vec{x}' \neq \vec{x}$, since $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}^*]\psi$. Let τ be $\vec{W} = \vec{w}^*$. Clearly, $\tau \in \mathscr{C}_{\vec{X}}$; by assumption, $(M, \vec{u}) \models \tau$. Moreover, $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}^*$ is consistent with $\vec{X} \neq \vec{x} \wedge \tau$ (since $\vec{W}$ and $\vec{X}$ are disjoint). Let $\vec{Y}$ consist of the endogenous variables in $\neg(\vec{X} = x) \wedge \tau$; note that $\vec{Y} = \vec{W} \cup \vec{X}$. Choose $\vec{y} \in \mathscr{R}(\vec{Y})$ so that $\vec{Y} = \vec{y}$ is $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}^*$. Thus, $\vec{X} \neq \vec{x} \wedge \tau \wedge \vec{Y} = \vec{y}$ is consistent. By assumption, $(M, \vec{u}) \models [\vec{Y} = \vec{y}]\neg\psi$. By definition, $(M, \vec{u}) \models (\vec{X} \neq \vec{x} \wedge \tau) \rightarrow \neg\psi$, so AC2′ holds, as desired.

For the converse, suppose that $\vec{X} = \vec{x}$ is a cause of φ with respect to $\mathscr{C}_{\vec{X}}$ in $(M, \vec{u})$ according to AC1′-3′. Again, it suffices to show that AC2 holds in $(M, \vec{u})$. Since AC2′ holds, there must exist a formula $\tau \in \mathscr{C}_{\vec{X}}$ such that $(M, \vec{u}) \models \tau$ and $(M, \vec{u}) \models (\vec{X} \neq \vec{x} \wedge \tau) \rightarrow \neg\psi$. Let $\vec{Y}$ be the endogenous variables that appear in $\vec{X} \neq \vec{x} \wedge \tau$. By definition, there must be a vector $\vec{y}$ of values such that $\tau \wedge (\vec{Y} = \vec{y})$ is consistent and $(M, \vec{u} \models [\vec{Y} = \vec{y}]\neg\psi$. $\vec{Y}$ must include $\vec{X}$. Let $\vec{x}'$ be the restriction of $\vec{y}$ to the variables in $\vec{X}$.

For each variable $Y \in \vec{Y}' = \vec{Y} - \vec{X}$, there is a value y such that $Y = y$ is a conjunct of τ; $Y = y$ must also be a conjunct of $\vec{Y} = \vec{y}$, since $\vec{Y} = \vec{y}$ is consistent with τ. Thus, $(M, \vec{u}) \models Y = y$ and, more generally, if $\vec{y}'$ is the restriction of the values in $\vec{y}$ to the variables in $\vec{Y}'$, we have that $(M, \vec{u}) \models \vec{Y}' = \vec{y}'$. It follows that AC2 holds, taking $\vec{W} = \vec{Y}'$ and $\vec{x}'$ as defined above. ∎

Theorem 4.5: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X} = \vec{x} \vee \vec{X} = \vec{x}^*$, where $\vec{x}^* \neq \vec{x}$. If $(M, \vec{u})$ is a recursive causal setting and (M', s) is strongly consistent with $(M, \vec{u})$, then $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of φ with respect to $\mathscr{C}_{\vec{X}}$ in (M', s) according to AC1′-3′.*

Proof: Suppose that $\vec{X} = \vec{x}$ is a cause of φ in a causal setting $(M, \vec{u})$ (according to AC1-3), and (M', s) is a recursive counterfactual structure strongly consistent with $(M, \vec{u})$. We want to show that $\vec{X} = \vec{x}$ is a cause of φ in (M', s). The argument is similar in spirit to the first half of the proof of Theorem 4.2.

Again, AC1′ and AC3′ are immediate; we need to show that AC2′ holds. Since AC2 holds, there is a set $\vec{W}$ of endogenous variables, a setting $\vec{w}^*$ such that $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow w^*]\neg\varphi$. Let τ be $\vec{W} = \vec{w}^* \wedge (\vec{X} = \vec{x} \vee \vec{X} = \vec{x}')$. Clearly, $\tau \in \mathscr{C}_{\vec{X}}$ and $(M, \vec{u})) \models \tau$. Since (M', s) is strongly consistent with $(M, \vec{u})$, by Proposition 4.3, we have that $(M', s) \models \tau$. Moreover, $(M', s) \models (\vec{X} \neq \vec{x} \wedge \tau) \rightarrow \neg\varphi$. The result follows.

For the converse, suppose that $\vec{X} = \vec{x}$ is a cause of φ with respect to $\mathscr{C}_{\vec{X}}$ in (M', s). To show that $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$, yet again, it suffices to show that AC2 holds. Since AC2′ holds, there must exist a formula $\tau \in \mathscr{C}_{\vec{X}}$ such that $(M', s) \models \tau$, and $(M', s) \models (\vec{X} \neq \vec{x} \wedge \tau) \rightarrow \neg\psi$. There is a conjunction of the form $\tau' \wedge \vec{X} = \vec{x}^*$ of non-negated primitive events that is equivalent to $\vec{X} \neq \vec{x} \wedge \tau$, where τ' does not mention any variable in $\vec{X}'$. Thus, $(M', s) \models (\vec{X} = \vec{x}^* \wedge \tau') \rightarrow \neg\psi$. By Proposition 4.3, it follows that $(M, \vec{u}) \models [\vec{X} = \vec{x}^* \wedge \tau']\neg\psi$. Since $(M', s) \models \tau$, we must have $(M', s) \models \tau'$, so by Proposition 4.3, $(M, \vec{u}) \models \tau'$. Thus, taking $\vec{W} = \vec{w}^*$ in AC2 to be τ', it follows that AC2 holds, as desired. ∎

Acknowledgments: Joe Halpern's work was suppprted in part by NSF grant FMitF-2319186, ARO grant W911NF-17-1-0061, MURI grant W911NF-19-1-0217 from the ARO, and a grant from Open Philanthropy. Rafael Pass's work was supported in part by NSF Award CNS 2149305, AFOSR Award FA9550-23-1-0387, AFOSR Award FA9550-24-1-0267 and ERC Advanced Grant KolmoCrypt - 1 01142322. Any opinions, findings, 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] F. Barbero, K. Schulz, F. R. Velázquez-Quesada & K. Xie (2023): *Observing interventions: A logic for thinking about experiments*. Journal of Logic and Computation 33, pp. 1152–1185, doi:10.1093/logcom/exac011.

[2] R. Briggs (2012): *Interventionist counterfactuals*. Philosophical Studies 160, pp. 139–166, doi:10.1007/s11098-012-9908-5.

[3] P. Gärdenfors (1988): *Knowledge in Flux*. MIT Press, Cambridge, MA.

[4] C. Glymour & F. Wimberly (2007): *Actual causes and thought experiments*. In J. Campbell, M. O'Rourke & H. Silverstein, editors: Causation and Explanation, MIT Press, Cambridge, MA, pp. 43–67.

[5] N. Hall (2007): *Structural equations and causation*. Philosophical Studies 132, pp. 109–136, doi:10.1007/s11098-006-9057-9.

[6] J. Y. Halpern (2013): *From causal models to possible-worlds models of counterfactuals*. Review of Symbolic Logic 6(2), pp. 81–101, doi:10.1017/s1755020312000305.

[7] J. Y. Halpern (2016): *Actual Causality*. MIT Press, Cambridge, MA.

[8] J. Y. Halpern & J. Pearl (2005): *Causes and explanations: a structural-model approach. Part I: causes*. British Journal for Philosophy of Science 56(4), pp. 843–887, doi:10.1093/bjps/axi147.

[9] J. Y. Halpern & J. Pearl (2005): *Causes and explanations: a structural-model approach. Part II: explanations*. British Journal for Philosophy of Science 56(4), pp. 889–911.

[10] C. G. Hempel (1965): *Aspects of Scientific Explanation*. Free Press, New York.

[11] C. Hitchcock (2001): *The intransitivity of causation revealed in equations and graphs*. Journal of Philosophy XCVIII(6), pp. 273–299, doi:10.2307/2678432.

[12] C. Hitchcock (2007): *Prevention, preemption, and the principle of sufficient reason*. Philosophical Review 116, pp. 495–532, doi:10.1215/00318108-2007-012.

[13] J. von Kügelgen, A. Mohamed & S. Beckers (2023): *Backtracking counterfactuals*. In: Proc. 2nd Conference on Causal Learning and Reasoning, pp. 9584–9594. Available in *Proceedings of Machine Learning Research*, Vol 213, pp. 1–20, 2023.

[14] D. Lewis (1973): *Counterfactuals*. Harvard University Press, Cambridge, MA.

[15] D. Lewis (2000): *Causation as influence*. Journal of Philosophy XCVII(4), pp. 182–197, doi:10.2307/2678389.

[16] C. G. Lucas & C. Kemp (2015): *An improved probabilistic account of counterfactual reasoning*. Psychological Review 122(4), pp. 700–734.

[17] W. C. Salmon (1970): *Statistical explanation*. In R. Kolodny, editor: The Nature and Function of Scientific Theories, University of Pittsburgh Press, Pittsburgh, PA, pp. 173–231.

[18] W. C. Salmon (1984): *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, N.J.

[19] C. Sartorio (2006): *Disjunctive causes*. Journal of Philosophy 103(10), pp. 521–538, doi:10.5840/jphil20061031029.

[20] M. Soloviev & J. Y. Halpern (2021): *Security properties as nested causal statements*. Available at http://arxiv.org/abs/2104.00872.

[21] R. C. Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: Studies in Logical Theory, Blackwell, Oxford, U.K., pp. 98–112.

[22] J. Woodward (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, U.K.

Optimality of Simultaneous Consensus with Limited Information Exchange (Extended Abstract)*

Kaya Alpturer
Princeton University
New Jersey, USA
kalpturer@princeton.edu

Ron van der Meyden
School of Computer Science and Engineering
UNSW Sydney
Sydney, Australia
r.vandermeyden@unsw.edu.au

Sushmita Ruj
School of Computer Science and Engineering
UNSW Sydney
Sydney, Australia
sushmita.ruj@unsw.edu.au

Godfrey Wong
School of Computer Science and Engineering
UNSW Sydney
Sydney, Australia
godfrey.wong@student.unsw.edu.au

Work on the development of optimal fault-tolerant Agreement protocols using the logic of knowledge has concentrated on the "full information" approach to information exchange, which is costly with respect to message size. Alpturer, Halpern, and van der Meyden (PODC 2023) introduced the notion of optimality with respect to a limited information exchange, and studied the *Eventual* Agreement problem in the sending omissions failure model. The present paper studies the *Simultaneous* Agreement problem for the crash failures model, and a number of limited information exchanges from the literature. In particular, the paper considers information exchanges from a FloodSet protocol (Lynch, Distributed Algorithms 1996), a variant of this in which agents also count the number of failures (Castañeda et al, NETYS 2017), and a variant in which agents associate each agent with a value (Raynal, PRDC 2002). A new information exchange is also introduced that enables decisions to be made at worst one round later than the optimal protocol of Dwork and Moses (I&C 88), but with lower computation cost and space requirements. By determining implementations of a knowledge based program, protocols are derived that are optimal amongst protocols for each of these information exchanges.

1 Introduction

Epistemic logic has been shown to play a useful role in the analysis of distributed algorithms [6]. By characterizing the knowledge that an agent needs to have in order to take an action, it is possible to obtain distributed algorithms that are optimal in the sense of terminating as soon as any other algorithm that solves the same problem.

In particular, variants of the Byzantine Agreement problem have been fruitfully studied using this methodology. (The term Byzantine agreement is used in literature even when the failure model is weaker than Byzantine failures [4, 10, 5, 1].) In Byzantine Agreement, a group of agents is required to make a common decision, in a setting where there could be faulty agents in the system that might be unreliable or even malicious. In the Simultaneous Byzantine Agreement (SBA) version of this problem, nonfaulty agents are required to make the decision at the same time (in the same round of computation). Reaching simultaneous agreement was shown to be related to common knowledge [4], and this

*This research is supported by an Australian Government Research Training Program (RTP) Scholarship. Research supported by AFOSR grant FA23862114029. The Commonwealth of Australia (represented by the Defence Science and Technology Group) supported this research through a Defence Science Partnerships agreement.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 175–189, doi:10.4204/EPTCS.437.17

© K. Alpturer, R. van der Meyden, S. Ruj and G. Wong
This work is licensed under the
Creative Commons Attribution License.

characterization has been used to derive optimal protocols for a variety of failure models, including crash [4] and omission failures [10].

In the present paper, we reconsider this type of optimality result in the particular case of crash failures and a synchronous message passing communications model. We assume synchronized clocks and a known upper bound on message delivery time, so that the computation can be structured into rounds. In a distributed system prone to crash failures, a faulty agent can crash during a round and send an arbitrary subset of the messages it was supposed to send. In the rounds following the crash, the crashed agent no longer sends any messages.

In the quest for optimality, most existing literature on using epistemic logic to study distributed algorithms has worked with *full information protocols*, in which each agent broadcasts its complete local state in each round, and stores all messages received. This results in states that grow exponentially with time. Even when the state size can be optimized, there are still cases where the computation required to be performed in each round is intractable. For example, in the general omissions model, optimal implementations require *NP*-hard computations [10, 9]. This means that practical protocols will often exchange less information than does the full information protocol. This raises the following question: having made a compromise on the information that is exchanged by the agents, is the protocol making full use of *that* information. That is, is the protocol optimal amongst protocols that use the same pattern of information exchange, terminating no later than does any other such protocol?

To date, only one paper has studied optimality of Byzantine agreement protocols with respect to limited information exchange [1]. There, the authors studied the *eventual* Byzantine agreement problem in synchronous message passing systems that are prone to sending omission failures. In the present paper, we will study the Simultaneous Byzantine Agreement (SBA) problem in synchronous message passing systems that are prone to crash failures.

We study a number of limited information exchanges derived from protocols in the literature, as well as one new information exchange "SendWaste". The main contribution of the paper is to derive protocols that are optimal with respect to these information exchanges, by identifying, in each run, the earliest time that agents are able to make a decision. In a number of cases, this analysis establishes decision times that are earlier than the times stated in the source publications, showing that the protocols in the literature can be optimized.

For FloodSet [7], in which the information exchanged in each round is the set of initial values an agent has heard about, we are able to terminate as soon as $m \geq \min\{t+1, n-1\}$, where m is time, t is maximum possible number of failures in a run, and n is the total number of agents. (When $t = n-1$, this is earlier than the time $t+1$ usually stated.)

For Counting FloodSet [3], which is an extension of FloodSet that counts the number of messages an agent has received, the earliest stopping condition is the same as FloodSet, except for the case when an agent does not receive any message from other agents.

We also show that even by extending Counting FloodSet to store a history of number of missing messages in each round, we are unable to obtain a stopping condition that terminates earlier. This protocol is of interest because in EBA, an earlier stopping condition can be obtained by retaining the count of the 2 previous rounds. We show that for SBA, there is no such advantage. However, we pay a price of more space used, for no advantage in the stopping time. So, while this protocol is optimal with respect to its own information exchange, it is still possible to improve this protocol by reducing its space usage.

For another variant, Vectorized FloodSet [11], in which agents keep not just a set of initial values, but also know which agents have those initial values, and in which a nonfaulty agent sends a message only when it learns a new agent-value pair, we establish an interesting nontrivial early stopping condition,

Protocol	Computation	Message size	Space
FloodSet [7]	$O(n)$	$O(1)$	$O(1)$
Counting FloodSet [3]	$O(n)$	$O(1)$	$O(1)$
Vectorized FloodSet [11]	$O(n^2)$	$O(n)$	$O(n)$
SendWaste [this paper]	$O(n)$	$O(1)$	$O(1)$
Dwork and Moses [4]	$O(nt)$	$O(t)$	$O(n)$

Table 1: Summary of computational complexity per round, message size, and space complexity for each agent.

that improves upon the condition $m = t+1$ given in [11]. In Vectorized FloodSet, we can decide as soon as the time is greater than $\min\{t+1, n-1\} - \max\{1, \beta_i(r,m)\}$, where $\beta_i(r,m)$ is the number of messages not received by agent i during round 1 of run r.

The most interesting protocol we consider is a new protocol SendWaste that we introduce. We show that it can terminate no later than one round after the protocol of Dwork and Moses [4], which is optimal with respect to full-information exchange, but at lower computation cost and space usage, particularly when the number of possible failures t is large. The protocol of Dwork and Moses maintains several sets of agents (one of which it transmits), from which it computes a notion of "Waste". Our new protocol SendWaste estimates the waste and transmits the estimate. This reduces both message size and computation costs.

The computational complexity per round, message size, and space complexity for each agent for the protocols we have studied is summarised in Table 1. (For the protocols that send or store numerical values, we assume that the number n can be stored in a single machine word, and that arithmetic operations have cost 1, since we expect this to be the case for practically feasible systems. Beyond this, some logarithmic adjustments would be required.)

Among the protocols studied in this paper, FloodSet performs the worst in every run, which terminates in $\min\{t+1, n-1\}$ rounds. Counting FloodSet has the same optimal early stopping condition except for the case where an agent detects $n-1$ missing messages. Counting FloodSet with Perfect Recall has the same optimal early stopping condition as Counting FloodSet in every run. Vectorized FloodSet is able to terminate as soon as Counting FloodSet except for the case where an agent does not receive $n-1$ messages. In this case, Counting FloodSet is able to decide while Vectorized FloodSet is unable to decide. SendWaste is able to terminate as early as any other FloodSet variants studied in this paper. The protocol of Dwork and Moses [4] is optimal with respect to full-information exchange and hence it is at least as good as any other protocol for any information exchange.

We begin by introducing our model of the problem in Section 2. Then, in Section 3 we discuss what it means for a Simultaneous Byzantine Agreement protocol to be optimal relative to an information exchange protocol. We give an analysis of the FloodSet protocol in Section 4 and its counting variants in Section 5. Section 6 is an analysis of the Vectorized FloodSet protocol from [11]. Our new protocol SendWaste is treated in Section 7. Finally, we conclude this paper in Section 8.

2 Framework

We will mostly follow the framework from [1] and [8]. In the Byzantine agreement problem, agents from a set $Agt = \{1, \ldots, n\}$ communicate using a synchronous, round-based, message-passing network

in order to decide on a value. Each agent starts with an initial value from the set $Values = \{0,1\}$. We assume that $t < n$ is the maximum number of failures that can occur in a run. We focus on *crash* failures, in which a faulty agent j acts correctly until a crash occurs in some round m in which an arbitrary subset of its messages in round m are delivered. Moreover, agent j never sends a message again in rounds $m' > m$.

The specification of Eventual Byzantine Agreement (EBA) can be stated as follows:

- **Agreement:** if nonfailed agents i and j decide on values v and v', respectively, then $v = v'$
- **Validity:** if a nonfailed agent decides on value v, then some agent has v as its initial value.

We remark that, in Agreement, we could have that $i = j$, and the times at which a decision is made could be different. Thus, the specification implies that once an agent has made a decision, it never makes a different decision. We allow that an agent makes the same decision at different times. In this paper, we do not adopt the "Unique Decision" requirement that an agent makes a decision at most once. The specification of the Simultaneous Byzantine Agreement (SBA) problem involves all the rules of Eventual Byzantine Agreement problem together with simultaneity:

- **Simultaneity:** if a nonfailed agent decides in a round, then every nonfailed agent decides in the same round.

Various formulations of a termination condition have been considered in the literature. The protocols we study satisfy

- **Termination:** every agent eventually decides or fails,

We decompose SBA protocols into two components $(P, \mathscr{E})$, comprised of a decision protocol P and an information exchange protocol $\mathscr{E}$. We represent the states of the environment in which agents operate by a set L_e, defined more precisely below. An information exchange protocol describes what information the agents record in their local states, and what messages an agent sends in a given local state. Formally, an information exchange protocol is a tuple $\langle \mathscr{E}_1, \ldots, \mathscr{E}_n \rangle$ containing a local information exchange protocol $\mathscr{E}_i$ for each agent i. A local information exchange protocol $\mathscr{E}_i$ of agent i is a tuple $\langle L_i, I_i, A_i, M_i, \mu_i, \delta_i \rangle$ that consists of:

- a set of local states L_i,
- a set of initial states $I_i \subseteq L_i$,
- a set of allowed actions A_i for agent i
- a set M_i of messages that are allowed to be sent by agent i,
- a message selection function $\mu_i : L_i \times A_i \to \prod_{j \in Agt}(M_i \cup \{\perp\})$,
- a transition function $\delta_i : L_i \times L_e \times A_i \times \prod_{j \in Agt}(M_j \cup \{\perp\}) \to L_i$.

For SBA protocols, we assume that there exist, for each agent i, a function $init_i : I_i \to Values$ that identifies an agent's initial value, and a function $time_i : L_i \to \mathbb{N}$ that identifies the time on the agent's clock. We assume that the function $init_i$ is a bijection, that is, that there is a unique initial state for each value in $Values$. The message selection function μ_i takes a local state and an action performed by an agent in a round to a tuple of messages that the agent sends in that round. The transition function δ_i updates the local state depending on the state of the environment, the action an agent performs in the round and the set of messages it receives in that round.

In the full information exchange, the local initial states are the initial values of the agent. In every round, each agent sends their local state to every other agent, and records every message it receives in its local state.

A decision protocol describes what actions the agents perform in a given situation. For the SBA problem, the action set for each agent i is $A_i = \{\mathtt{decide}_i(v) \mid v \in \mathit{Values}\} \cup \{\mathtt{noop}\}$. A *local decision protocol* $P_i : L_i \to A_i$ for agent i maps a local state of agent i to an action that i can perform. A *decision protocol* P is a tuple $\langle P_1, \ldots, P_n \rangle$ of local decision protocols for all agents.

It will help when comparing two information exchanges below to assume that agent's decisions do not affect the information that they transmit. We say that an information exchange $\mathscr{E}$ *does not transmit information about actions* if, for all agents i, local states $s \in L_i$, and actions $a_1, a_2 \in A_i$,

- $\mu_i(s, a_1) = \mu_i(s, a_2)$ (that is, the messages the agent sends do not depend on their action), and
- for all message vectors $\mathtt{m}$ and environment state $s_e \in L_e$, we have $\delta_i(s, s_e, a_1, \mathtt{m}) = \delta_i(s, s_e, a_2, \mathtt{m})$ (that is, the agent's state update does not depend on their action).

The definition does not allow an agent to record the particular decision it has made in its local state, since it could then transmit information about this decision in the next step, since μ_i still depends on the local state. All the concrete information exchanges we consider in this paper have the property that they do not transmit information about actions.

A failure model describes the failures that could occur in a run of a protocol. We use the hard crash failure model Crash_t, depending on a parameter t that is a number less than the number of agents. In this failure model, there may be up to t faulty agents, who can crash at any time. During the round in which an agent crashes, it sends a subset of the messages it is supposed to send. After an agent has crashed it does not send any message and no longer runs the protocol.

We represent Crash_t as a set of *adversaries*, where an adversary is a function $\mathscr{F} : \mathbb{N} \times \mathit{Agt} \times \mathit{Agt} \to \{\top, \bot\}$. If agent i crashes in round m, then we require that $F(m, i, i) = \bot$. If the message that agent i was supposed to send to agent j is not sent in round m, we have $\mathscr{F}(m, i, j) = \bot$, otherwise $\mathscr{F}(m, i, j) = \top$. If $\mathscr{F}(m, i, i) = \bot$ then $\mathscr{F}(m', i, k) = \bot$ for all times $m' > m$ and every $k \in \mathit{Agt}$. There may be at most t agents i such that $\mathscr{F}(m, i, j) = \bot$ for some m and j. We take the set of states of the environment L_e to be the set of all such adversaries. Furthermore, for this model, we assume that the set of local states of agent i contains a special state *crashed*, representing that the agent has crashed. The state update function operates so that in a round m in which the adversary $s_e = \mathscr{F}$ and $\mathscr{F}(m, i, i) = \bot$, we have $\delta_i(s_i, s_e, a, v) = \mathit{crashed}$ for all local states s_i of agent i, actions a and vectors v of messages.

A global state is an element of $L_e \times \prod_{i \in \mathit{Agt}} L_i$ consisting of a tuple comprised of the state of the environment and a local state of each agent. A *run* is a mapping $r : \mathbb{N} \to L_e \times \prod_{j \in \mathit{Agt}} L_j$ from times to a global states. If $r(m) = \langle s_e, s_1, \ldots, s_n \rangle$, we write $r_e(m)$ for s_e and, for each agent i, $r_i(m)$ for s_i. The time between (r, m) and $(r, m+1)$ is *round* $m+1$. A pair (r, m) comprised of a run r and a time m is a *point*.

A *system* is a set $\mathscr{R}$ of runs. An *interpreted system* is a pair $\mathscr{I} = (\mathscr{R}, \pi)$, where $\mathscr{R}$ is a system and π is an *intepretation* mapping each point (r, m) of $\mathscr{I}$ to a function $\pi(r, m) : \mathit{Prop} \to \{\mathit{true}, \mathit{false}\}$ that determines whether or not each atomic proposition from a set Prop is true at the point.

Given an information exchange $\mathscr{E}$ and decision protocol P for the crash failure model Crash_t, we construct an interpreted system $\mathscr{I}_{P,\mathscr{E},\mathit{Crash}_t} = (\mathscr{R}, \pi)$ as follows. (Generally, Crash_t will be implicit, and we write $\mathscr{I}_{P,\mathscr{E}}$ for brevity.) The runs in $\mathscr{R}$ are generated by selecting an initial global state $r(0) = \langle s_e, s_1, \ldots, s_n \rangle$, where $s_e \in L_e = \mathit{Crash}_t$ is an environment state encoding an adversary from the crash failure model and $s_i \in I_i$ for all $i \in \mathit{Agt}$ are initial local states of agents. The remainder of the run is uniquely determined from the initial global state by the following induction. The state of the environment $r_e(m)$ will be the same for all times m. For each round $k+1$, the local state $r_i(k+1)$ of agent i is determined as follows. First, each agent i uses the decision protocol P_i to select its action $a_i = P_i(r_i(k))$, and attempts to send the messages $\mu_i(r_i(m), a_i)$. The adversary $r_e(k)$ determines which of these messages are delivered, so that each agent i receives an agent indexed vector v_i of messages (or

$\perp$ in case of a message that was not sent or is not delivered because the agent sending it crashed). Agent i's local state $r_i(k+1)$ is then equal to $\delta_i(r_i(k), r_e(k), a_i, v_i)$. We consider only *synchronous* protocols, for which $time_i(r_i(m)) = m$ for all points (r,m) of runs of the protocol.

Note that if $\mathscr{E}$ does not transmit information about actions, then the set of runs of $\mathscr{I}_{P,\mathscr{E}}$ does not depend on P. We may therefore write simply $\mathscr{I}_{\mathscr{E}}$ when not referring to propositions about the protocol.

The atomic propositions *Prop* in the interpretation π consist of:

- $\exists v$, which is true if an agent in the current run has v as its initial value,
- $i \in \mathscr{N}$, which is true if agent i has not failed up to the current time, and
- *clean*, which is true if a clean round has occurred (defined below).

We work with a modal logic that uses the atomic propositions along with the standard boolean operators $\neg, \vee, \wedge$. For an agent i and an *indexical* set $\mathscr{N}$ of agents (which may depend on the point at which we evaluate the formula) we also have unary modal operators $K_i, E_{\mathscr{N}}, C_{\mathscr{N}}$ which we will define later. The semantics of the logic in an interpreted system $\mathscr{I}$ is given by a relation $\models$, such that $(\mathscr{I}, r, m) \models \phi$, for a point (r,m) of $\mathscr{I}$ and a formula ϕ, represents that ϕ is true at the point (r,m) of $\mathscr{I}$. The definition of this relation is given by an induction on the construction of the formula ϕ. For an atomic proposition p, we have $(\mathscr{I}, r, m) \models p$, if $\pi(r,m)(p) = \mathbf{True}$.

The semantics of the modal operators is given using an indistinguishability relation $\sim_i$ on points for each agent i. The points (r,m) and (r',m') are indistinguishable to agent i, written as $(r,m) \sim_i (r',m')$, if $r_i(m) = r'_i(m')$. Note that the indistinguishability relation is an equivalence relation. The intuition behind the definition of knowledge is that agent i knows φ if φ is true in all the points that are indistinguishable to i.

For a formula φ and point (r,m) of an interpreted system $\mathscr{I}$, we say:

- agent i knows φ at (r,m), written as $(\mathscr{I}, r, m) \models K_i\varphi$, if for all points (r',m') of $\mathscr{I}$ satisfying $(r,m) \sim_i (r',m')$, we have $(\mathscr{I}, r', m') \models \varphi$,
- everyone in $\mathscr{N}$ knows φ at (r,m), written as $(\mathscr{I}, r, m) \models E_{\mathscr{N}}\varphi$, if for every $i \in \mathscr{N}(r,m)$, we have $(\mathscr{I}, r, m) \models K_i(\varphi)$,
- φ is common knowledge among $\mathscr{N}$ at the point (r,m), written as $(\mathscr{I}, r, m) \models C_{\mathscr{N}}\varphi$, if for every $k \in \mathbb{Z}^+$, we have $(M, s) \models E^k_{\mathscr{N}}\varphi$.
- φ is distributed knowledge amongst group $\mathscr{N}$, written as $(\mathscr{I}, r, m) \models D_{\mathscr{N}}\varphi$, if $(\mathscr{I}, r', m) \models \varphi$ for all $(r',m) \in \bigcap_{i \in \mathscr{N}(r,m)} S_i(r,m)$, where $S_i(r,m)$ is the set of points of $\mathscr{I}$ that are indistinguishable to agent i from (r,m).

A formula φ is valid in $\mathscr{I}$, written as $\mathscr{I} \models \varphi$, if for every point (r,m) of $\mathscr{I}$, we have $(\mathscr{I}, r, m) \models \varphi$.

In this paper, the indexical set $\mathscr{N}$ represents the set of nonfailed agents. Formally, $i \in \mathscr{N}(r,m)$ iff $r_i(m) \neq crashed$.

3 Optimality and Knowledge-based Programs

In this section we study knowledge-based programs and what it means for a protocol to be optimal with respect to a given information exchange.

To compare two protocols, we consider their behaviour in runs in which the environment has the same behaviour. Two runs $r \in \mathscr{I}$ and $r' \in \mathscr{I}'$ *correspond* if they have identical initial values and the

failure pattern is the same in both runs, that is, for all agents i, we have $init_i(r_i(0)) = init_i(r'_i(0))$ and $r_e(0) = r'_e(0)$.

To define optimality, we define a partial order on protocols. Let P and P' be two decision protocols for the same information exchange $\mathscr{E}$. We write $P \leq_{\mathscr{E}} P'$ if for all corresponding runs r of P and r' of P' with respect to $\mathscr{E}$, and for all times m, if $P_i(r_i(m)) = \mathtt{decide}_i(v)$ for some value v, then $P'_i(r'_i(m)) = \mathtt{decide}_i(w)$ for some value w. That is, at any time when P makes a decision, P' also makes a decision (but possibly a different one). We define P to be an *optimal SBA protocol with respect to* $\mathscr{E}$ if P is an SBA protocol with respect to $\mathscr{E}$, and for all SBA protocols P' with respect to $\mathscr{E}$, if $P \leq_{\mathscr{E}} P'$ then $P' \leq_{\mathscr{E}} P$. That is, there does not exist an SBA protocol P' with respect to $\mathscr{E}$ that makes decisions strictly more often than P. Protocol P is an *optimum SBA protocol with respect to* $\mathscr{E}$ if it is an SBA protocol with respect to $\mathscr{E}$ such that for all SBA protocols P' with respect to $\mathscr{E}$, we have $P' \leq_{\mathscr{E}} P$.

These definitions of optimality are related to other definitions of optimality in the literature [1, 2], which focus on optimizing the earliest time at which a decision is made, as follows. Note that if $P \leq_{\mathscr{E}} P'$, then protocol P never makes its first decision earlier than P' makes its first decision. Thus, an optimal SBA protocol cannot be improved with respect to the earliest time at which it makes a decision on any run, and an optimum protocol always decides as least as early as any other protocol.

The following lemma is taken from Moses and Tuttle [10]. Its proof does not assume a full information protocol. The proposition $\mathrm{deciding}_i(v)$ is defined to be true at a point (r,m) if agent i decides v in round $m+1$ of run r, i.e., $P_i(r_i(m)) = \mathtt{decide}_i(v)$.

Lemma 1. *[10] Let r be a run of an SBA protocol generating interpreted system $\mathscr{I}$ and let $i \in \mathscr{N}(r,m)$. If $(\mathscr{I},r,m) \models \mathrm{deciding}_i(v)$ for $v \in \{0,1\}$, then $(\mathscr{I},r,m) \models C_{\mathscr{N}}(\exists v)$.*

It is useful to know that a failure-free run requires at least $\min\{t+1, n-1\}$ rounds to attain common knowledge of initial values. Therefore, when time $m < \min\{t+1, n-1\}$ and r is a run of any SBA protocol, if (r,m) is $\mathscr{N}$-reachable from a failure-free run then we do not have common knowledge of any initial value at (r,m).

Lemma 2. *[4] Let r be a failure-free run of an SBA protocol generating system $\mathscr{I}$. When $m < \min\{t+1, n-1\}$ we have $(\mathscr{I},r,m) \models \neg C_{\mathscr{N}}(\exists 0) \wedge \neg C_{\mathscr{N}}(\exists 1)$.*

Moses and Tuttle [10] have proven that a full information protocol attains common knowledge no later than any arbitrary SBA protocol. A *proposition φ is about the environment* if its truth value in interpreted systems for the SBA problem, as defined above, depends only on the adversary and the initial values of the agents. That is, for corresponding runs r and r' of systems $\mathscr{I}$ and $\mathscr{I}'$, respectively, and all times m, we have $(\mathscr{I},r,m) \models \varphi$ iff $(\mathscr{I}',r',m) \models \varphi$.

Lemma 3. *[10, Corollary 6] Let φ be a proposition about the environment. Let r and r' be corresponding runs of an arbitrary protocol $(P,\mathscr{E})$ and a full-information protocol $(P',\mathscr{E}')$ respectively. If $(\mathscr{I}_{P,\mathscr{E}},r,m) \models C_{\mathscr{N}}\varphi$, then $(\mathscr{I}_{P',\mathscr{E}'},r',m) \models C_{\mathscr{N}}\varphi$.*

Let $\mathscr{E}^1$ and $\mathscr{E}^2$ be two information exchange protocols that do not transmit information about actions. We say $\mathscr{E}^1$ *stores at least as much information as* $\mathscr{E}^2$ if, for every pair of corresponding runs r^1 and r^2 of $\mathscr{I}_{\mathscr{E}_1}$ and $\mathscr{I}_{\mathscr{E}_2}$, respectively, and every pair of corresponding runs r^3 and r^4 of $\mathscr{I}_{\mathscr{E}_1}$ and $\mathscr{I}_{\mathscr{E}_2}$, respectively, $(r^1,m) \sim_i (r^3,m)$ implies $(r^2,m) \sim_i (r^4,m)$. This relation is reflexive and transitive.

Lemma 4. *Let $\mathscr{E}^1$ be an information exchange that stores at least as much information as information exchange $\mathscr{E}^2$ and let r^1 and r^2 be corresponding runs of $\mathscr{I}_{\mathscr{E}^1}$ and $\mathscr{I}_{\mathscr{E}^2}$ respectively. If r^3 and r^4 are corresponding runs, of $\mathscr{I}_{\mathscr{E}^1}$ and $\mathscr{I}_{\mathscr{E}^2}$ respectively, such that (r^3,m) is $\mathscr{N}$-reachable from (r^1,m) in k steps, then (r^4,m) is $\mathscr{N}$-reachable from (r^2,m) in k steps.*

We now get the following generalization of Lemma 3.

Proposition 1. *Suppose φ is a proposition about the environment. Suppose $\mathscr{E}^1$ stores at least as much information as protocol $\mathscr{E}^2$ and let r^1 and r^2 be corresponding runs of $\mathscr{I}_{\mathscr{E}^1}$ and $\mathscr{I}_{\mathscr{E}^2}$, respectively. Then $(\mathscr{I}_{\mathscr{E}^1}, r^1, m) \models \neg C_{\mathscr{N}}\varphi$ implies $(\mathscr{I}_{\mathscr{E}^2}, r^2, m) \models \neg C_{\mathscr{N}}\varphi$.*

We remark that Proposition 1 does not make assumptions about the failure model and would work even with arbitrary failures.

Knowledge based programs [5] can be understood as specifications, in a code-like form, that describe how each agent's actions relate to its knowledge. Syntactically, they are like standard programs for an agent, except that the formulas that occur in as the conditions of conditional statements may be a formula of the logic of knowledge stating a property of the agent's knowledge, rather than just a predicate of the agent's local state, as in standard programs. Program 1 gives an example, which will be the specific knowledge based program that we study in the present paper. This program is to be interpreted as the rule used to determine agent i's action in each round, until one of the actions $\texttt{decide}_i(v)$ is selected, after which the program terminates. In this program, $\exists 0$ and $\exists 1$ means there exists an agent with initial value of 0 and 1 respectively.

Program 1: $\mathbf{P}_i^{\mathrm{OPT}}$

if $K_i(C_{\mathscr{N}}(\exists 0))$ **then** $\texttt{decide}_i(0)$
else if $K_i(C_{\mathscr{N}}(\exists 1))$ **then** $\texttt{decide}_i(1)$
else `noop`

The semantics of knowledge based programs is given by relating concrete protocols to a knowledge based program. Since we only consider one specific program in this paper, we do not give this definition in general, and just explain this relation for the specific program $\mathbf{P}_i^{\mathrm{OPT}}$. Intuitively, a decision protocol P implements the knowledge based program with respect to an information exchange $\mathscr{E}$, if, in each round, it selects for each agent i the same action as would be obtained for agent i from the rule in the knowledge based program. "If-then-else" statements are interpreted according to the usual semantics, but to interpret the knowledge formulas, we need an interpreted system. For this, we use the interpreted system obtained from running P with respect to $\mathscr{E}$.

Formally, given an interpreted system $\mathscr{I}$, for each agent i we define the decision protocol $(\mathbf{P}_i^{\mathrm{OPT}})^{\mathscr{I}}$ on a local state $s \in L_i$ by taking $(\mathbf{P}_i^{\mathrm{OPT}})^{\mathscr{I}}(s)$ to be the action a selected the rule $\mathbf{P}_i^{\mathrm{OPT}}$, with the truth values of the knowledge formulas evaluated at any point (r,m) of $\mathscr{I}$ with $r_i(m) = s$. Note that the truth values of these formulas are independent of the choice of such an (r,m), since they depend only on the local state of agent i at that point. Using this, we define a decision protocol *P to be an implementation of Program $\mathbf{P}^{OPT}$ with respect to $\mathscr{E}$ in the crash failure context $Crash_t$* if for all points (r,m) of $\mathscr{I}_{\mathscr{E},Crash_t}$ and all agents $i \in \mathscr{N}(r,m)$, we have $(\mathbf{P}_I^{\mathrm{OPT}})^{\mathscr{I}}(r_i(m)) = P_i(r_i(m))$.

To obtain SBA protocols, we use a special case of the Agreement property:

Self-Agreement: For all agents i, if i decides value v and later decides value v', then $v = v'$.

The following result shows that implementations of the knowledge based program $\mathbf{P}^{\mathrm{OPT}}$ satisfying this property are SBA protocols.

Theorem 1. *Suppose P is an implementation of $\mathbf{P}^{OPT}$ with respect to information exchange $Crash_t$ and P satisfies Self-Agreement. Then P is an SBA protocol, that is, $\mathscr{I}_{P,\mathscr{E},Crash_t}$ satisfies SBA.*

The following result shows that implementations of the knowledge based program $\mathbf{P}^{\mathrm{OPT}}$ with respect to an information exchange yield SBA protocols that are optimum with respect to that information

exchange. In the following sections, we apply this result by considering a number of information exchanges and deriving protocols that are implementations of $\mathbf{P}^{\text{OPT}}$ with respect to these information exchanges. The resulting protocols will therefore be optimum SBA protocols with respect to their information exchanges.

Theorem 2. *Let an SBA protocol P be an implementation of $\mathbf{P}^{OPT}$ with respect to information exchange $\mathscr{E}$ in the crash failure context $Crash_t$. Suppose that $\mathscr{E}$ does not transmit information about actions. Then, P is an optimum SBA protocol with respect to $\mathscr{E}$ in the crash failure context $Crash_t$.*

Using Theorem 2, we can show an SBA protocol to be an optimum SBA protocol by showing it is an implementation of $\mathbf{P}^{\text{OPT}}$. In the following sections, We do this for a number of concrete protocols. In particular, we consider FloodSet in Section 4, Counting FloodSet in Section 5, Vectorized FloodSet in Section 6, and SendWaste in Section 7.

4 FloodSet

Lynch introduced the FloodSet protocol [7] as an example of a simple protocol that solves SBA under crash failures. In this section, we present a refined analysis of the Floodset protocol in terms of the agents' knowledge, showing that after a minor modification, the Floodset protocol is an optimum SBA protocol, given the amount of information the agents are exchanging.

We now define the FloodSet protocol in two parts using our framework, the information-exchange protocol $\mathscr{E}^{flood}$ and the decision protocol P^{flood}. The local states of agent i in $\mathscr{E}^{flood}$ will be either the special state $crashed_i$, or a tuple $\langle W_i, m, v_i\rangle$, where W_i is a set of distinct values agent i has seen so far, m is the current time, and v_i is the initial value of i. The information-exchange protocol $\mathscr{E}^{flood}$ has the set of initial local states I_i for each agent i consisting of all states of the form $\langle\{v_i\}, 0, v_i\rangle$ for some $v_i \in Values$. In each round, every agent i sends W_i to all other agents. That is, $\mu_i((W_i, m, v_i), a_i) = (W_i, \ldots, W_i)$ for all $(W_i, m, v_i) \in L_i$ and actions a_i, and $\mu_i(crashed_i, a_i) = (\bot, \ldots, \bot)$. At the end of a round $m+1$, v_i stays the same, and the local state of each agent is updated to have a new W_i which is the union of all messages received in the current round. That is, the function δ_i is defined by

$$\delta_i\left((W_i, m, v_i), \mathscr{F}, a_i, \left(W_1', \ldots, W_n'\right)\right) = \left(W_i \cup \bigcup_j W_j', m+1, v_i\right)$$

provided the environment does not crash agent i, i.e., $\mathscr{F}(m, i, i) \neq \bot$, otherwise

$$\delta_i((W_i, m, v_i), \mathscr{F}, a_i, (W_1', \ldots, W_n')) = crashed_i$$

and $\delta_i(crashed_i, \mathscr{F}, a_i, (W_1', \ldots, W_n')) = crashed_i$.

Throughout this section, we fix $t \leq n$ and write $\mathscr{I}$ for $\mathscr{I}_{\mathscr{E}_{flood}, Crash_t}$. When agent i's local state at point (r, m) of $\mathscr{I}$ is $r_i(m) = (W, m, v)$, we define $W_i(r, m)$ to be W, and $time_i(r, m)$ to be m.

The standard program from [7] representing the FloodSet decision protocol for an agent i such that $P^{flood} = (P_1^{flood}, \ldots, P_n^{flood})$ is Program 2.

A crucial definition used in the analysis of SBA protocols is the notion of a *clean round*, in which every nonfailed agent receives the same messages.

Definition 1 (Clean Round)**.** *A round is clean if for any nonfailed agents i and j, agent i receiving a message from agent k implies that agent j also received a message from agent k.*

The following show the importance of clean rounds.

Program 2: P_i^{flood} from [7]

if $m = t+1$ **then** $\texttt{decide}_i(\min W_i)$
else `noop`

Lemma 5. *[7, Lemma 6.1] If round m of run r of FloodSet is clean, then $W_i(r,m) = W_j(r,m)$ for all nonfailed agents i, j in run r.*

Lemma 6. *[7, Lemma 6.2] Suppose $W_i(r,m) = W_j(r,m)$ for any nonfailed agents i, j in run r of FloodSet. Then, $W_i(r,m') = W_j(r,m')$ for all $m \leq m' \leq t+1$.*

Therefore, since we can only have up to t crashes, by round $t+1$ we are guaranteed to have a clean round. Moreover, all agents $i \in Agt$ have the same $W_i(r,t+1)$ in any run r in $\mathscr{I}$. Note that however, when $t = n-1$, either we have a clean round by time t, or exactly one agent failed in every round until round t, so that only one active agent is remaining. In both cases, we should be able to decide without waiting until $t+1$. This observation leads to the following refinement to the Floodset decision protocol. Let $P^{flood+} = (P_1^{flood+}, \ldots, P_n^{flood+})$ be the decision protocol defined by the following standard program:

Program 3: P_i^{flood+}

if $m \geq \min\{t+1, n-1\}$ **then** $\texttt{decide}_i(\min W_i)$
else `noop`

Theorem 3. *Let (r,m) be a point in $\mathscr{I}$. We have $(\mathscr{I}, r, m) \models \bigvee_{v \in Values} K_i(C_{\mathscr{N}}(\exists v))$ for all $i \in \mathscr{N}(r,m)$ if and only if $m \geq \min\{t+1, n-1\}$.*

By Theorem 2, this result implies that the protocol P^{flood+} is optimal with respect to the information-exchange $\mathscr{E}^{flood}$.

5 Counting Extensions

The aim of this section is to explore what happens if an agent stores the number of missing messages during a round. We aim to find out whether we can achieve a better early stopping condition by making the above change to the FloodSet information-exchange protocol $\mathscr{E}^{flood}$. This modification was studied in [3] and was shown to lead to an early stopping condition for EBA. However, there is no literature on this modification for SBA. Furthermore, we will extend Counting FloodSet to store the entire history of number of missing messages to investigate if we are able to obtain an earlier stopping condition. We consider this variant because [3] shows that a predicate based on the *difference* between two successive counts of missing messages leads to an EBA protocol that stops earlier than the protocol based just on a single count. Again, we are interested in whether this information is helpful in the setting of SBA.

The local states of agent i in the Counting FloodSet information-exchange protocol $\mathscr{E}^{count}$ will be either the state $crashed_i$ or a tuple $\langle W_i, h_i, m, v_i \rangle$, where W_i is the set of distinct values agent i has seen so far, h_i is the number of missing messages during round m, m is the current time, and v_i is the initial value of i. The initial state of an agent i which starts with initial value of v_i is $\langle \{v_i\}, 0, 0, v_i \rangle$. As in the FloodSet protocol, in each round, each nonfailed agent i sends a message comprised of the set W_i to all other agents. The state update function adds any values in a message W_j received from another agent

to the set W_i, but additionally sets h_i to the number of agents from which i did not receive a message in the current round. The time m is incremented in each round, but the value v_i is unchanged.

The local states of agent i in the Counting FloodSet with perfect recall information-exchange protocol $\mathscr{E}^{count(pr)}$ will be a tuple $\langle W_i, h_i, m, v_i\rangle$, where W_i is the set of distinct values agent i has seen so far, h_i is a vector in which $h_i[k]$ is the number of missing messages during round k, m is the current time, and v_i is the initial value of i. The initial state of an agent i which starts with initial value of v_i is $\langle\{v_i\},[0],0,v_i\rangle$. Messages and state updates are as in the Counting FloodSet protocol, except that the count of missing messages is appended to h_i.

The decision rules used in the early stopping condition in both information-exchanges can be found in the following programs:

Program 4: Counting FloodSet decision rule

if $m \geq \min\{t+1, n-1\}$ **then** decide$(\min W_i)$
else if $h_i \geq n-1$ **then** decide$(\min W_i)$
else noop

Program 5: Counting FloodSet (perfect recall) decision rule

if $m \geq \min\{t+1, n-1\}$ **then** $\texttt{decide}_i(\min W_i)$
else if $\exists k \leq m(h_i[k] \geq n-1)$ **then** $\texttt{decide}_i(\min W_i)$
else `noop`

Lemma 7. *The following relations hold amongst the versions of the FloodSet information exchange:*

1. $\mathscr{E}^{count(pr)}$ *stores at least as much information as* $\mathscr{E}^{count}$.
2. $\mathscr{E}^{count}$ *stores at least as much information as* $\mathscr{E}^{flood}$.

The following results characterize the situations in which agents obtain common knowledge when using the counting variants of FloodSet.

Theorem 4. *Let r be a run of the Counting FloodSet protocol, let $\mathscr{I} = \mathscr{I}_{\mathscr{E}^{count}, Crash_t}$, and let m be a time. When the local state of agent i at point (r,m) is $r_i(m) = \langle W, h, m, v_i\rangle$ we have $(\mathscr{I}, r, m) \models \bigvee_{v \in Values} C_{\mathscr{N}}(\exists v)$ if and only if $m \geq \min\{t+1, n-1\}$ or $h \geq n-1$.*

Theorem 5. *Let r be a run of Counting FloodSet with perfect recall, let $\mathscr{I} = \mathscr{I}_{\mathscr{E}^{count(pr)}, Crash_t}$, and let m be a time. When the local state of agent i at point (r,m) is $r_i(m) = \langle W, h, m, v_i\rangle$ we have $(\mathscr{I}, r, m) \models \bigvee_{v \in Values} C_{\mathscr{N}}(\exists v)$ if and only if $m \geq \min\{t+1, n-1\}$ or there exists $1 \leq k \leq m$ such that $h[k] \geq n-1$,*

We have proven that Counting FloodSet and Counting FloodSet with perfect recall are optimum SBA protocols with respect to their information exchange, due to Theorem 2.

6 Vectorized FloodSet

Raynal [11] describes a protocol that differs from the FloodSet protocol by recording not just the initial values received but also which agents had these initial values. Instead of sending a message every round, an agent only sends newly received information, immediately after the round in which the agent receives the new information. As in the Floodset protocol, agents do not learn whether any other agent

has detected a failure. Agents also do not record the round number in which a message is received. The local states of agent i at a point (r,m) have the form $\langle V_i(r,m), New_i(r,m), time_i(r,m)\rangle$ where $V_i(r,m)$ is an agent-indexed array of initial values (or $\perp$, representing an unknown value), $New_i(r,m)$ is a set of tuples of type $Values \times Agt$ recording new information that agent i received in round m, and $time_i(r,m) = m$ is the current time. Intuitively, a tuple $(v,j) \in Values \times Agents$ represents the information that agent j has initial preference v. This information is new to agent i in round k if at the start of the round (time $k-1$), the agent has $V_i[k](r,k-1) = \perp$ and the agent receives a message $New_{i'}$ in round k containing the tuple (v,j).

Code for Vectorized FloodSet is given in the following figure. From this, the information exchange protocol $\mathscr{E}^R$ for this protocol can be easily constructed. (We leave the details for the reader to complete.) For the remainder of this section, we write $\mathscr{I}$ for $\mathscr{I}_{\mathscr{E}^R,Crash_t}$.

Program 6: Vectorized FloodSet for crash failures for agent i

```
V_i ← [⊥,...,v_i,...,⊥]
New_i ← {(v_i,i)}
for r = 1,2,...,t+1 do
    begin round
    if New_i ≠ ∅ then
        for each agent j ≠ i do
            send New_i to agent j
    New_i ← ∅
    W_j ← message received from agent j or ∅ if none
    for each agent j ≠ i do
        for each (v,k) ∈ W_j do
            if V_i[k] = ⊥ then
                V_i[k] ← v
                New_i ← New_i ∪ {(v,k)}
    W_j ← ∅
    end round
if ∃0 ∈ V_i then decide 0
else decide 1
```

In Raynal's presentation, the protocol waits until the end of round $t+1$ before making a decision. We show that an optimal protocol using the same information exchange can make earlier decisions. Specifically, define $\beta_(r,m)$ be the number of $\perp$ values in the array $V_i(r,m)$. We show that the implementation of the knowledge based program uses the following predicate of agent i's local state to determine the time m at which a decision can be made: $m > \min\{t+1, n-1\} - \max\{1, \beta_i(r,m)\}$.

The following theorem concludes when have common knowledge of initial values in Vectorized FloodSet.

Theorem 6. *Let r be any run in Vectorized FloodSet. For any time m, we have $(\mathscr{I}, r, m) \models (C_{\mathscr{N}}(\exists 0) \vee C_{\mathscr{N}}(\exists 1))$ if and only if $m > \min\{t+1, n-1\} - \max\{1, \beta_i(r,m)\}$.*

We have found that Vectorized FloodSet is not optimal with respect to its information exchange. An optimum early stopping condition is when $m > \min\{t+1, n-1\} - \max\{1, \beta_i(r,m)\}$.

7 Sending the Waste

Dwork and Moses [4] showed that an early decision can be made at time $m \geq \min\{t+1, n-1\} - \mathscr{W}(r)$, where the waste $\mathscr{W}(r) = \max_{k\geq 0}\{diff(r,k)\}$ and difference $diff(r,k) = \#KF(r,k) - k$ for

$$\#KF(r,k) = \max_{0\leq \ell \leq t} \{\ell \mid (\mathscr{I}, r, k) \models D_{\mathscr{N}}(\text{"}\ell \text{ agents have failed"})\} .$$

They use this to develop a protocol that is optimal with respect to the full information exchange. This protocol transmits a set of agents (newly detected to be faulty) in each round, and maintains a number of variables, each of type a set of (faulty) agents. To determine the set of agents from which it did not receive a message in a given round, the protocol requires a temporary variable recording the agents from which a message was received. This implies that the protocol has message size $O(t)$, space requirement $O(n)$ and computation cost per round of $O(nt)$.

In this section, we develop a new protocol, that is able to make decisions almost as soon as the Dwork and Moses protocol, but at lower cost $O(1)$ in message size, and $O(n)$ in computation per round, and $O(1)$ in space costs. Whereas [4] need to track which agents have been heard from, SendWaste need only *count* the messages received in a round. The protocol is based on an information exchange in which agents maintain, in addition to a set of known values, a numerical estimate of the waste, which they transmit to other agents in each round. In this information exchange protocol, each agent estimates the waste at time m by $h - m$, where h is the number of missing messages in round m. The value d stores the agent's estimate of the waste, determined from the agent's own detection of missing messages and the waste estimates that it receives from other agents.

The local states of agent i in the SendWaste information-exchange protocol $\mathscr{E}^{send}$ will be a tuple $\langle W_i, h_i, d_i, m, v_i\rangle$, where W_i is the set of distinct values agent i has seen so far, h_i is the number of messages missing from the last round, d_i is the waste estimate, m is the current time, and v_i is the initial value of the agent. The initial states are $\langle\{v\}, 0, 0, m, v\rangle$, where $v \in Values$.

In every round of SendWaste, agent i sends (W_i, d_i) to every agent. Upon receiving messages in round m, agent i updates W_i to be the union of all W_j's it receives and d_i is updated to be the maximum of the agent's previous value of d_i, the values d_j in messages (W_j, d_j) received from other agents j, and $h_i - m$, where h_i is the number of missing messages in the current round.

For this information exchange, we define the SendWaste standard decision protocol P_i^{send} for an agent i by

Program 7: P_i^{send}

if $m \geq \min\{t+1, n-1\} - d_i$ **then** $\texttt{decide}_i(\min W_i)$
else `noop`

and take $P^{send} = (P_1^{send}, \ldots, P_n^{send})$. We will show that that P^{send} implements the knowledge based program $\mathbf{P}^{OPT}$ with respect to $\mathscr{E}^{send}$, and is therefore an optimal SBA protocol with respect to that information exchange.

For the remainder of this section, we write $\mathscr{I}$ for $\mathscr{I}_{\mathscr{E}^{send}, Crash_t}$. We write $d_{\mathscr{N}}(r,m)$ for the value $\max_{i\in\mathscr{N}(r,m)} d_i(r,m)$. The following theorem concludes when do we have and not have common knowledge of initial values in SendWaste.

Theorem 7. *Let i be a nonfailed agent at (r,m), where r is a run of SendWaste. We have $(\mathscr{I}, r, m) \models \bigvee_{v\in Values} C_{\mathscr{N}}(\exists v)$ if and only if $(\mathscr{I}, r, m) \models m \geq \min\{t+1, n-1\} - d_{\mathscr{N}}$.*

We now show that SendWaste decides at most one round later than the protocol from Dwork and Moses [4, Figure 3], which is an optimum with respect to full-information exchange.

Theorem 8. *Let r' be a run of Dwork and Moses' protocol [4, Figure 3] and r be its corresponding run of SendWaste. If r' decides at round m', then r decides at some m, where $m' \leq m \leq m' + 1$.*

We have proven that SendWaste is an optimum SBA protocol with respect to its information exchange. In addition, the protocol is able to decide at most one round later than a protocol that is an optimum with respect to full-information exchange. We can further improve the space complexity and message size of SendWaste by only storing and sending $\min W$ instead of the whole set W. With an $O(\log t)$ bit increase to message size when compared to FloodSet, SendWaste has an early stopping condition close to Dwork and Moses' protocol.

8 Conclusion

We have studied the FloodSet, Counting FloodSet, Counting FloodSet with perfect recall and Vectorized FloodSet information exchanges for simultaneous Byzantine agreement in synchronous message-passing systems prone to crash failures. We have also introduced a new information exchange SendWaste in which agents transmit their known values and their estimate of the waste. For each of these information exchanges, we have determined the earliest possible time that common knowledge about the initial values is attained, giving the earliest time at which agents are able to decide. This yields protocols that are an optimum for each of these information exchanges.

For FloodSet and its variants, the early stopping condition is trivial. On the other hand, Vectorized FloodSet has an interesting nontrivial early stopping condition.

In addition, we came up with a limited information-exchange protocol that decides at most one round later than the Dwork and Moses [4] protocol that is optimum with respect to full-information exchange. The new protocol, SendWaste, has less computational complexity and space complexity than the Dwork and Moses protocol. The message size is also reduced when compared to the protocol from Dwork and Moses.

We have used a version of the SBA specification in the present work that does not require that agents make a decision at most once, and have, indeed, allowed that agents continue to perform decision actions after they have first made a decision. Related to this, the local states in the concrete information exchange protocols we study do not record the fact that an agent has decided. This facilitates comparisons of protocols that make different decisions, or decide at different times. It also supports comparison of information exchange protocols with respect to the amount of information that they transmit, which facilitates transfer of conclusions from one protocol to another in our proofs.

A downside of these assumptions is that our protocols do not satisfy the Unique Decision property that is often taken to be part of the SBA specification. We conjecture that, were we to modify these protocols so that they decide at most once, by adding a record of the fact that an agent has decided, then we would still be able to show that these protocols are the optimum SBA protocol for the given information exchanges, in the sense that they decide at least as early as all other SBA protocols for these information exchanges. Results of van der Meyden [8] show that a terminating variant of the knowledge based program of the present paper yields *optimum* SBA protocols in this sense, for information exchanges that do not transmit any information about actions, but may record decision information in a separate component of the local state. To apply this to our present protocols requires some different techniques to show that the decision protocols are implementations of the knowledge based program. We leave this for future work.

In this paper, we considered the problem only under crash failures. In future work, we intend to study the problem under general omission failures. For this failure model, a polynomial time Simultaneous Byzantine Agreement protocol, optimal with respect to its own information exchange, would be particularly interesting, because an optimal protocol with respect to the full information exchange is known to require NP-hard computations in each round [10].

References

[1] Kaya Alpturer, Joseph Y. Halpern & Ron van der Meyden (2023): *Optimal Eventual Byzantine Agreement Protocols with Omission Failures.* In: *Proceedings of the 2023 ACM Symposium on Principles of Distributed Computing*, PODC '23, Association for Computing Machinery, New York, NY, USA, p. 244–252, doi:10.1145/3583668.3594573.

[2] Armando Castañeda, Yannai A. Gonczarowski & Yoram Moses (2022): *Unbeatable consensus. Distributed Comput.* 35(2), pp. 123–143, doi:10.1007/S00446-021-00417-3.

[3] Armando Castañeda, Yoram Moses, Michel Raynal & Matthieu Roy (2017): *Early Decision and Stopping in Synchronous Consensus: A Predicate-Based Guided Tour.* In Amr El Abbadi & Benoît Garbinato, editors: *Networked Systems*, Springer International Publishing, Cham, pp. 206–221. Available at https://doi.org/10.1007/978-3-319-59647-1_16.

[4] Cynthia Dwork & Yoram Moses (1990): *Knowledge and common knowledge in a Byzantine environment: Crash failures. Information and Computation* 88(2), pp. 156–186, doi:10.1016/0890-5401(90)90014-9. Available at https://www.sciencedirect.com/science/article/pii/0890540190900149.

[5] R. Fagin, J.Y. Halpern, Y. Moses & M.Y. Vardi (1995): *Reasoning About Knowledge.* The MIT Press.

[6] Joseph Y. Halpern & Yoram Moses (1990): *Knowledge and Common Knowledge in a Distributed Environment. J. ACM* 37(3), p. 549–587, doi:10.1145/79147.79161.

[7] Nancy A. Lynch (1996): *Distributed Algorithms.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[8] Ron van der Meyden (2024): *Optimal Simultaneous Byzantine Agreement, Common Knowledge and Limited Information Exchange.* Submitted for publication. https://cgi.cse.unsw.edu.au/~meyden/research/common-knowledge.pdf.

[9] Yoram Moses (2009): *Optimum Simultaneous Consensus for General Omissions Is Equivalent to an NP Oracle.* In Idit Keidar, editor: *Proc. Distributed Computing, 23rd International Symposium, DISC 2009*, *Lecture Notes in Computer Science* 5805, Springer, pp. 436–448, doi:10.1007/978-3-642-04355-0_45.

[10] Yoram Moses & Mark R. Tuttle (1988): *Programming Simultaneous Actions Using Common Knowledge. Algorithmica* 3, pp. 121–169, doi:10.1007/BF01762112.

[11] M. Raynal (2002): *Consensus in synchronous systems: a concise guided tour.* In: *2002 Pacific Rim International Symposium on Dependable Computing, 2002. Proceedings.*, pp. 221–228, doi:10.1109/PRDC.2002.1185641.

Common Knowledge, Sailboats, and Publicity

Sena Bozdag* Olivier Roy†

We revisit a recent puzzle about common knowledge, the "sailboat" case [22], and argue that Lewisian common knowledge allows us to reconcile the pre-theoretical intuition that certain facts are "public" in such situations, while these facts cannot be common knowledge in the classical, iterative sense. The crux of the argument is to understand Lewisian common knowledge as an account of what it means for an event to be public. We first formulate this argument informally to clarify its philosophical commitment and then propose one way to capture it formally in epistemic-plausibility models. Taken together, we take the philosophical and the formal arguments as providing evidence that Lewisian common knowledge is a plausible account of what it means for an event to be public.

The various mathematical definitions of common knowledge, most prominently the iterative and fixed point definitions [4, 5], have often been associated, explicitly or implicitly, with the pre-theoretical notions of something being "public information", "out in the open", or what "any fool" [24] would know. Barwise's [4] argument that "the fixed point is the best conceptual analysis of the pre-theoretic notion" implicitly takes the latter to be some form of publicity. Geanakoplos' early survey [15] explicitly states that "public events are the most obvious candidates for common knowledge." Similar remarks can be found in [26, 11] and in informal presentations of the characterizations of iterative common knowledge in terms of "self-evident events" [15, 28]. This association between the pre-thereotical notion of publicity and the mathematical definitions of common knowledge is also salient outside epistemic logic and game theory, for instance, in philosophy of collective action [16, 8, 7], or in Chwe's work on social rituals [10].

Many of the classical and more recent criticisms of the iterative or the fixed-point definitions of common knowledge can be diagnosed as questioning this association, c.f. [31] for an overview. A common criticism is, for example, that iterative common knowledge is unattainable for resource-bounded agents due to its infinitary character [30]. This fact is most worrying when one assumes that information can be public among such agents, and that iterative common knowledge should somehow capture this notion of publicity. One can similarly diagnose the "Halpern-Moses problem" [20, 27, 18]. The paradoxical nature of the problem is most salient when one grants, on the one hand, that successful asynchronous coordination often rests on the publicity of some relevant facts and, on the other hand, that the iterative or the fixed-point definition of common knowledge should capture the latter.

A natural response to these worries, which we follow in this paper, is to propose an alternative account of publicity and argue that it meets some of the challenges facing the iterative or fixed-point definitions of common knowledge [16, 17, 37]. Here we follow a suggestion made by Paternotte [31] to use *Lewisian* common knowledge [23, 12, 33, 36] to flesh out the pre-theoretical understanding of publicity. While Paternotte provides general, philosophical arguments for understanding publicity in terms of Lewisian common knowledge, in this paper we pitch this suggestion against a structurally related "paradox" of iterated common knowledge: Lederman's "Sailboat" example [22]. In doing so, we argue that Lewisian common knowledge allows us to reconcile the pre-theoretical intuition that some obvious facts are public with the observation that iterative common knowledge of these facts do not hold, even for ideal reasoners.

*Universität Bayreuth
†Universität Bayreuth

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 190–202, doi:10.4204/EPTCS.437.18

The rest of this extended abstract is organized as follows. After briefly introducing the iterated and the Lewisian definitions of common knowledge, we sketch Lederman's Sailboat example and our philosophical argument to the effect that Lewisian common knowledge addresses it satisfactorily. We then provide a formalization of our argument in a variant of epistemic-plausibility models [3, 35]. Section 5 concludes that Lewisian common knowledge constitutes a plausible account of publicity.

1 Iterated and Lewisian Common Knowledge

For the present paper, only the iterative definition of common knowledge is relevant. We use its standard version: A proposition A is shared knowledge in a group G whenever every member of G knows A. A is level-$n+1$ shared knowledge in group G whenever it is shared knowledge in G that A is level-n shared knowledge. Taking plain shared knowledge as level 0, iterative common knowledge of A is defined as level-n shared knowledge of A, for all $n \in \mathbb{N}$.

Lewisian common knowledge differs from the iterative definition in two ways. First and foremost, despite its name, it does *not* feature higher-order knowledge, but rather generates higher-order *reasons to believe*.[1] Second, Lewisian common knowledge requires the existence of a specific *state of affairs*, the "basis" for common knowledge, which grounds the group members' higher-order reasons to believe. In that sense, Lewisian common knowledge has a much more "externalist" flavor [31], since it refers to shared aspects of the informational environment [5], instead of referring to the members' mental states.

We use Cubitt and Sugden's [12] rendering of the Lewisian definition. Let $R_i(...)$ stand for "i has a reason to believe...". A proposition B is then said to be *Lewisian common knowledge* in group G whenever there exists a state of affairs A such that:

C1 If A then $R_i(A)$ for any agent i

C2 For all agents i, j if $R_i(A)$ then $R_iR_j(A)$

C3 For any agent i, if $R_i(A)$ then $R_i(B)$

C4 For all agents i, j and proposition C, from the fact that $R_i(A)$ implies $R_i(C)$ it follows that $R_i(R_j(A)$ implies $R_j(C))$

Conditions C1 to C3 encapsulate the idea that A is the basis for common knowledge, or what Cubitt and Sugden call a reflexive common indicator of B. These conditions do most of the work in generating the infinite hierarchy of higher-order reasons to believe B. C4 states that the agents share reasoning standards.[2]

Lewis argued informally that these conditions together entail that every member of G has iterated higher-order reasons to believe B, which we write $r^n(B)$ following Cubitt and Sugden's [12] notation. In other words, the existence of the relevant state of affairs A satisfying C1 to C4 entails that both agents have a reason to believe B, they have a reason to believe that they have a reason to believe B, and so on. Lewis' argument, sometimes also called "Lewis' Theorem," has been formalized using a variety of tools [12, 33, 36].

[1] Vromen [36] provides textual evidence that David Lewis himself thought *ex post* that "common knowledge" was a misnomer for the notion he defined.

[2] Observe that A is a state of affairs and not a proposition. $R_i(A)$ is thus somewhat an abuse of notation and should be read as "i has a reason to believe that A holds." Note also that our formulation of C1-C4 slightly differs from Cubit and Sugden's original formalization. Theirs also uses the Lewisian notion of "indication" in C2-C4. That notion is, however, defined in terms of conditional reasons to believe, and once unpacked it yields the formulation above.

2 The Sailboat Case

Our goal in this paper is to provide support to the view that Lewisian common knowledge constitutes a plausible account of what it means for a proposition or a state of affairs to be public. We do so by analyzing the "Sailboat case" devised by Lederman [22]. Before explaining in more detail the structure of our argument, let us introduce the case.

> SAILBOAT: Roman and Columba are ideal reasoners playing in a game show.[3] Each contestant has a single button on a console in front of him or her. They have an unobstructed view of each other's faces, and of an area in the middle of the stage, where the hosts will place a sailboat. First, the hosts will bring out a toy sailboat (the 'test') with a 100 cm mast. They will then replace it with a sailboat chosen randomly from an array of sailboats of various sizes. If the mast of the new sailboat is taller than the test and both players press their respective buttons, they receive $ 1,000 each. If the mast is not taller than the test and both press, or if only one person presses their button, the person or people who pressed must pay the show $ 100. Today, the mast of the chosen boat is 300 cm tall. [22, p.1075]

Lederman assumes that the reader will judge the following to be intuitively true in this example, which we also accept for the sake of argument in this paper.

> It is public among Roman and Columba that the mast is larger than 100 cm.

We write $[> 100]$ for "the mast is larger than 100 cm." Here Lederman uses a pre-theoretical notion of publicity, which stays unanalyzed in the paper. Our point here is precisely that this notion can be fleshed out in terms of Lewisian common knowledge.

Lederman argues that, even though this proposition is public, $[> 100]$ is not common knowledge in the iterative sense. The argument rests on the assumption that Roman and Columbia's knowledge is inexact, that is, it has a margin of error m, and that this is public between them. In his words:

> [Interpresonal Ignorance] For all r, if it looks to be r cm tall to one of the agents, then for all that agent knows, it looks to be $[r \pm m]$ cm tall to the other. [22, p.1081]

In this quote, "for all the agent knows" is capturing the notion of "considering it possible", or "not being ruled out by the agent's information" frequently represented by the "diamond" operator in epistemic logic. Interpersonal ignorance thus says that for all n, if it looks to an agent that the mast is n cm tall, then that agent considers it possible that the mast looks $n - m$ tall to the other. That fact is also considered to be public. The argument that they do not have common knowledge then runs like many of the classical examples or "paradoxes" of common knowledge, most prominently the Byzantine Generals case. It builds a chain of "Roman considers it possible that Columba considers it possible that..."[4]

3 Philosophical Argument

We now present a first version of our argument. We claim that even granting $[> 100]$ is not common knowledge in the iterative sense, Lewisian common knowledge can recapture the pre-theoretical intuition

[3] For our argument the fact that Roman and Columba are competing in a game show is not important.

[4] With this in hand Lederman goes on to argue for the admittedly more thought provoking claim that, for all practical purposes, iterative common knowledge does not exists, because cases of inexact knowledge are overwhelmingly common. This is much stronger than what is suggested by the Sailboat case alone, and our argument in this paper is independent of whether that stronger claim is correct.

that $[> 100]$ is public. In other words, a failure of iterative common knowledge of $[> 100]$ is consistent with Lewisian common knowledge of that proposition.

Three remarks are in order before proceeding. First, we start with an informal version of the argument to clarify its philosophical commitments. Second, despite us talking in terms of consistency, we do not take issue with Lederman's claim that theories associating publicity and iterative common knowledge for ideal reasoners are inconsistent with the fact that $[> 100]$ is public in the Sailboat case [22, p.1075]. Our argument can, in fact, be read as adding support to that claim. Relatedly, it is worth re-emphasizing that we follow Lederman and assume that Roman or Columba are ideal reasoners. So the argument is not based on them only being able to entertain finitely many levels of knowledge or on them not forming the beliefs that they have reasons to form. The argument rests rather on the simple observation that having a reason to believe a proposition A is consistent with considering $\neg A$ possible, i.e. not completely ruling out that A might be false.

Assumption: existence of a reflexive common indicator A: To argue that $[> 100]$ is Lewisian common knowledge in the Sailboat example we need to identify a "basis", i.e. the reflexive common indicator A that grounds the relevant higher-order reasons to believe. We simply take for granted that because Roman and Columba "have an unobstructed view of each other's faces and an area in the middle of the stage, where the hosts will place a sailboat" they are in a "shared environment" that constitutes the relevant basis A for Lewisian common knowledge. We furthermore assume that A can be approximately described by a proposition, so that statements of the form "$R_i(A)$" are meaningful. This means, in other words, that we assume that conditions C1 and C2 hold for this appropriately described A. Finally, we assume that this A entails certain other propositions, crucially with regard to the margin of errors and $[> 100]$.

Assumption: shared standards of reasoning: We also assume that Roman and Columba have "shared standards of reasoning." They are ideal reasoners, and this is public between them. Standards of reasoning go, however, beyond "mere" logical omniscience, since they also capture evidential relations and common sense, defeasible inferences that either of them can make. This should be modeled more precisely using some defeasible or nonmonotonic theories for the "implies" and "if... then" statements in C4, but we leave that issue aside in this paper. See, however, [31] for one way to model defeasibility in Lewisian common knowledge, using probabilistic credences.

Publicity of the margins: Let us write $[n-m, n+m]$ for the fact that the mast is at least $n-m$ cm, at most $n+m$ cm tall, with n being any number and m the given margin of error. We write $[n]$ for the fact that the mast is exactly n cm tall. Publicity of the margins would then correspond to saying that A is a reflexive common indicator for the following:

$$\forall n \forall i \text{ if } [n] \text{ then } R_i([n-m, n+m]) \qquad \text{(Margins)}$$

For this we need to argue that C1-C4 hold for (Margins), given the basis A that we assume exists. We have already mentioned that C1, C2, and C4 hold by assumption about A. The fact that the margins are public then simply boils down to saying that C3 holds with B being (Margins). In other words, if Roman and Columba have a reason to believe that A holds, then they have a reason to believe (Margins). We find it a plausible assumption.

By "Lewis' Theorem" we then obtain that there is iterated, higher-order reason to believe (Margins), which we write $r^n(Margins)$.

In particular, for every finite level k, we have that Roman and Columba have level-k reason to believe that (Margins) holds. For instance we get that:

$$R_R(\forall n \forall i \in \{R,C\} \text{ if } [n] \text{ then } R_i([n-m,n+m]))$$

$$R_C R_R(\forall n \forall i \in \{R,C\} \text{ if } [n] \text{ then } R_i([n-m,n+m]))$$

$$R_R R_C R_R(\forall n \forall i \in \{R,C\} \text{ if } [n] \text{ then } R_i([n-m,n+m]))$$

... and so on, where $R_R(...)$ should be read as "Roman has a reason to believe that...", and similarly for R_C and Columba. We furthermore assume that reasons to believe are sufficiently closed under logical consequence[5] for Roman and Columba to have the corresponding (higher-order) reasons to believe every instantiation of these conditionals with specific values of n and agents R or C. We assume, for instance, that the following also holds:

$$R_R(\text{ if } [300] \text{ then } R_C([300-m,300+m]))$$

$$R_C R_R(\text{ if } [300-m] \text{ then } R_C([300-2m,300]))$$

$$R_R R_C R_R(\text{ if } [300-2m] \text{ then } R_C([300-3m,300-m]))$$

... and so on.

Publicity of $[>100]$: We also assume that the basis A for Lewisian common knowledge is such that C3 holds for B being $[>100]$. So we also have $r^n([>100])$. Depending on the size of the margin, there must thus be a number k of iterations of $R_C R_R ...$, which we write $r_k ...$, and a number l such that $300-(l+m)<100$ and for which the following two statements hold.

$$r_k(\text{ if } [300-l] \text{ then } R_C([300-(l+m),(300-l)+m])$$

$$r_k(R_C[>100])$$

We now informally claim that these two statements are consistent. The first statement in the scope of r_k is a conditional, and only its consequent contradicts $R_C([>100])$. As long as this consequent does *not* detach, there is no inconsistency. It would detach if we had $r_k([300-l])$, assuming that the R_i operators satisfy a version of the K axiom, and taking into account the type of defeasible conditional in the scope of r_k. We content that Roman and Columba do not have this type of unconditional iterated reason to believe, except for 300cm, the actual size of the mast. The epistemic-plausibility model that we construct below can be seen as a way to flesh out this part of the informal argument.

Consistency of publicity with failure of iterative common knowledge: Up to now our argument has been aimed at showing that even though the margins are public, it can be Lewisian common knowledge that the mast is greater than 100 cm. We now want to argue that this Lewisian common knowledge is consistent with Roman and Columba *not* having iterative common knowledge of that same fact.

We grant for the sake of the argument that Roman and Columba consider it possible that the mast is $300-m$. They have, after all, a reason to believe that $[300-m,300+m]$. As Lederman argues,

[5] In the philosophical argument we require closure under logical consequence only to go from the general form of $r^n(Margins)$ to its instantiations. In the formal argument, however, use the standard semantics for the modal operators, which means that both agents are modeled as logically omniscient.

because the margins of errors are public, both consider it possible that the other has a reason to believe that $[300-2m, 300]$, which in turn entails that they both consider it possible that the other considers it possible that the other has a reason to believe that $[300-3m, 300-m]$, and so on. This iteration does not stop when the lower bound of the interval is below 100, so it is not common knowledge between Roman and Columba, in the iterative sense, that the mast is taller than 100 cm.

This failure of iterative common knowledge of $[>100]$ is consistent with Lewisian common knowledge of that proposition for the simple reason that not ruling out something, say $\neg C$, is consistent with having a reason to believe that C is, in fact, the case. Recall that we follow Lederman and assume that Roman and Columba both consider it possible that the mast is $300-m$ cm tall. This does not entail that either of them has a reason to believe $[300-m]$, or has a reason *not* to disbelieve $[300-m]$. Reasons to believe come in various strengths, and even a conclusive reason to believe might not ground absolute certainty. Roman and Columba might be quite confident in their assessment of the mast, but they are also aware of the margin of error. This means that even though the reason for $[300]$ might be the strongest piece of evidence that each has, that reason might not be strong enough to completely rule out $[300-m]$. In other words, having a reason to believe $[300]$ is perfectly consistent with not ruling out $[300-m]$.

This point generalizes to all $[n]$, which allows the iteration of "Roman considers it possible that Columba considers it possible that Roman..." to continue past the lower bound of 100 cm without threatening Lewisian common knowledge of $[>100]$. As we saw in the previous section, because the margins are public, Roman and Columba have a conditional reason to believe that *if* the mast *was* $300-m$ tall, then the other *would* have a reason to believe $[300-2m, 300]$. And, of course, since they consider $[300-m]$ possible, they do consider it possible that the other has a reason to believe $[300-2m, 300]$ and, by the same argument as in the previous paragraph, this is consistent with the other considering it possible that the mast is $300-2m$ cm tall. But this, so to speak, "possible reason to believe", does not translate into an unconditional, second-order reason to believe that the other has a reason to believe $[300-2m, 300]$, and *a fortiori* that the other has a reason to believe that the mast is $300-2m$ cm tall. Each has, after all, a whole array of such possible reasons to believe, one for each number n in the interval $[300-m, 300+m]$. Consequently, while there is a number of iterations of "Roman considers it possible that Columba considers it possible that..." for which the mast is less than 100 cm tall, because the consequents of the relevant conditional reasons to believe do not detach, neither at the first nor at any higher-order levels, this failure of iterated common knowledge does not clash with the Lewisian common knowledge of $[>100]$. Again, "considering it possible" is different, and crucially much weaker than having a reason to believe.[6]

4 An epistemic-plausibility model to flesh out the philosophical argument

We now argue that the intuitions underlying our philosophical argument can be fleshed out using a variant of epistemic-plausibility models [3, 35]. We should emphasize at the outset that we interpret these models in a slightly non-standard way, for two reasons. First, these models and the language describing them have been developed to study the relation between knowledge and conditional beliefs, and not necessarily the reasons supporting the latter. Second, the standard semantics of conditional beliefs in epistemic-plausibility models is not well suited to handle cases where the condition is known to be false. To address this, we augment epistemic plausibility models with selection functions, a standard tool for analyzing counterfactual beliefs in belief revision [1] and game theory [34, 19].

[6] This point can also be made using the distinction between pro tanto and conclusive reasons [9]. We omit it here for reasons of space, but we elaborate on it in the full version of the paper.

Before we proceed, a further disclaimer is in order: We do *not* claim that the specific variant of epistemic-plausibility models we use provides, in general, a good account of reasons to believe or of Lewisian common knowledge, nor that it should be seen as an alternative to existing formalizations of the latter [12, 33, 31, 36]. In the full paper, we expand on the reasons why. Our claim is that, given the specific features and idealizations of the Sailboat case, the construction below allows us to flesh out the philosophical consistency argument in the previous section.

Language: We work with a propositional modal language whose propositional variables are of the form $[n]$, for all $n \in \mathbb{N}$, together with a designated proposition $\top$ that will denote propositional tautologies. We use p to designate elements of $Prop = \{[n] : n \in \mathbb{N}\} \cup \{\top\}$. Our set of agents A is simply $\{\rho, \kappa\}$, for Roman and Columba. The propositional language is then extended with two modalities, $R_i(\varphi|p)$ and $K_i\varphi$, to be read "agent i has a reason to believe φ, given p" and "agent i knows φ." For technical reasons to be explained below, we restrict the syntax so that only propositional variables can appear as conditions in formulas of the form $R_i(\varphi|p)$:

$$p \mid \varphi \wedge \psi \mid \neg\varphi \mid R_{i\in A}(\varphi|p) \mid K_{i\in A}\varphi$$

Formulas of the form $R_i(\varphi|p)$ express conditional reasons to believe. We have a subjective interpretation of these conditional reasons, such as "*i* has a reason to believe that, given p, φ is true", as opposed to "given p, i has a reason to believe φ". Unconditional reasons to believe $R_i(\varphi)$ are defined as $R_i(\varphi|\top)$. These formulas have a dual, which we will write $\langle R_i\rangle(\varphi|p)$ and define as $\neg R_i(\neg\varphi|p)$, and $\langle K_i\rangle\varphi$, defined as usual. We do not use the dual of having reasons to believe in our argument below, whereas $\langle K_i\rangle\varphi$ is the object-language expression of "considering it possible" modality we introduced early on. With this in hand, we are able to define the higher-order notions that we will need below:

- $r^0(\varphi|p) = \bigwedge_{i\in A} R_i(\varphi|p)$. Given $r^k(\varphi|p)$, we define $r^{k+1}(\varphi|p)$ as $\bigwedge_{i\in A} R_i(r^k(\varphi|p)|p)$, and $r^n(\varphi|p)$ as $r^k(\varphi|p)$, for all $k \in \mathbb{N}$.
- Similarly, $\langle K_i^0\rangle\varphi = \langle K_i\rangle\varphi$, and $\langle K_i^{k+1}\rangle\varphi$ as $\langle K_i\rangle\langle K_j^k\rangle\varphi$ for $i \neq j$, and $\langle K_i^n\rangle\varphi$ as $\langle K_i^k\rangle\varphi$ for all n.[7]

Epistemic-Plausibility Frames, Models, and Selection Functions: An epistemic-plausibility frame $\mathscr{F}$, given our set of agents A is a tuple $\langle W, \{\leq_i\}_{i\in A}\rangle$ where W is a set of states and each $\leq_i$ is a preorder, i.e., reflexive and transitive relation on W. The strict relations $w <_i w'$ are defined as $w \leq_i w'$ but not $w' \leq_i w$. The *epistemic accessibility* relation $\approx_i$ is defined as: $w \approx_i w'$ iff either $w \leq_i w'$ or $w' \leq_i w$. We write $[w]_i$ for $\{w' : w \approx_i w'\}$. Given our set of propositional letters $Prop = \{\top\} \cup \{[n] : n \in \mathbb{N}\}$, a *epistemic-plausibility model* $\mathscr{M}$ is an epistemic-plausibility frame $\mathscr{F}$ together with a valuation $V : Prop \to \mathscr{P}(W)$ such that $V(\top) = W$. Given an epistemic-plausibility model $\mathscr{M}$, a *selection function* for $\mathscr{M}$ is any $f : (Prop \times W) \to W$.[8]

Truth conditions: Let $\mathscr{M}$ be an epistemic-plausibility model and f a selection function for it. Then we set:

- $\mathscr{M}, w \models p$ iff $w \in V(p)$.

[7] For the construction below we need only this limited, diamond version of higher-order knowledge, but we take the iterative common knowledge to be defined in the usual way.

[8] For the construction below we only need to define conditional reason to believe, and therefore selection functions, for the case where the condition is a propositional variable of our language. Hence the syntactic restriction on $R_i(\varphi|p)$ and the definition of the selection function taking only arguments in *Prop* instead of the full language $\mathscr{L}$.

- $\mathscr{M}, w \models R_i(\varphi|p)$ iff $max_{\leq_i}(||p|| \cap [f(p,w)]_i) \subseteq ||\varphi||$.
- $\mathscr{M}, w \models K_i\varphi$ iff $[w]_i \subseteq ||\varphi||$

where $||\varphi|| = \{w : \mathscr{M}, w \models \varphi\}$ and for any $X \subseteq W$, $max_{\leq_i}(X) = \{w \in X :$ there is no $v \in X$ such that $w <_i v\}$.

Remarks: The truth conditions for the two modal operators are spelled out using the underlying plausibility orderings $\leq_i$. This requires some explanation.

As their name suggests, the orderings $\leq_i$ are typically interpreted in terms of doxastic plausibility: $w \leq_i v$ is then meant to model the fact that *i* considers *v* at least as plausible as *w*. Here we slightly deviate from that standard interpretation. We take $w \leq_i v$ to model the fact that *i has a reason to believe* that *v* is at least as plausible as *w*. We take these reasons to be subjective, conclusive and normative [25].

The conditional reason-to-believe operators are then interpreted using a combination of this plausibility ordering and the selection function. Leaving the latter aside for a moment, the semantics above yields that an agent has a reason to believe ψ conditional on p whenever ψ is true in all p states (compatible with *i*'s total evidence) for which *i* has a reason to believe are maximally plausible.[9]

The truth conditions for both the conditional reason to believe and the knowledge operators are relativized to the agent's informational environment $[w]_i$ at any given state. This environment is, in turn, defined using the plausibility orderings $\leq_i$: the equivalence relation $\approx_i$ generated by the connected components of $\leq_i$ [3]. Philosophically, this gives the following interpretation. All states in $[w]_i$ are those that are not excluded for the reasons to believe that *i* has at *w*. $[w]_i$ can thus be seen as capturing *i*'s total evidence at *w* [38].

Returning to the selection functions, they are meant to capture *i*'s conditional reasons to believe $R_i(\varphi|p)$ both in cases where *i*'s total evidence rules out the truth of p, and in cases where *i*'s total evidence is consistent with p. We define a concrete selection function below, but we could also define them using general constraints that characterize minimal revisions [34]. For example, one could require that if $f(\varphi,w) = w'$ then $\mathscr{M}, w' \models \varphi$, or that $f(\varphi,w) = w$ whenever $\mathscr{M}, w \models \varphi$. The selection function below satisfies the former, but not the latter.

Butterflies and the butterfly flutter: We are now ready to describe the main construction that fleshes out the philosophical argument presented in the previous section. We construct a specific epistemic plausibility model that fulfills three functions. First, it describes the actual situation of Roman and Columba, where both the fact that the mast is 300 cm tall, and thus larger than 100 cm, and the margins of errors, are public. Here we capture publicity by the fact that Roman and Columba have iterated higher-order reasons to believe these propositions.[10] Second, the model captures the fact that they both entertain the relevant chains of "Roman considers it possible that Columba considers it possible... that the mast is *k* cm tall", even for *k* smaller than 100. Finally, the model also captures the fact that Roman and Columba have the relevant reasons to believe that the first two conditions would hold, counterfactually, for any $k \neq 300$. The construction accordingly proceeds in three steps corresponding to these three functions.

[9] This interpretation is not unlike standard semantics for conditional obligations in deontic logic that "lift" reason-based orderings on states to orderings on propositions [21, 13]. Our treatment of conditional reasons to believe is similar also to treatment of counterfactual type of beliefs in the Lewis-Stalnaker tradition.

[10] Although we argue indirectly for this, the formal argument thus leaves out an explicit representation of the fact that conditions C1-C4 hold for a given reflexive common indicator. The way publicity is captured in our formal model is therefore reminiscent of the understanding of publicity in terms of common commitment to believe [37].

Let *Prop* and A be as above. Now take $k,m \in \mathbb{N}$. Intuitively, k will be the 'center' of the 'butterfly', and m the margin, with $k-m>0$. Our goal is to define inductively a *butterfly centered on* k as an epistemic-plausibility model $\mathscr{M}_k$. We start with the base case.

Step 1: Body of the butterfly. The *body* $\mathscr{M}_k^0$ of the butterfly, illustrated in Figure 1, Left, is defined as:

- $W^0 = \{w_0, w_1, w_2, w_3, w_4\}$.
- $w_0 >_\rho^0 w_1$, $w_0 >_\rho^0 w_2$, $w_0 >_\chi^0 w_3$, $w_0 >_\chi^0 w_4$
- $V^0(k) = \{w_0\}, V^0(k-m) = \{w_1, w_3\}, V^0(k+m) = \{w_2, w_4\}$, $V^0(\top) = W^0$, and $V^0([l]) = \emptyset$ for all $l \neq k, k-m, k+m$.

The world w_0 in the body of the butterfly centered on k is called its *center*. The body of the butterfly models the shared environment, i.e. the reflexive common indicator A in our philosophical argument, of Roman and Columba. As we shall see below, taking $k = 300$ ensures that while both know that the mast is at most $300+m$ and at least $300-m$ tall, they have an iterated reason to believe that it is 300 cm, that is, $r^n([300])$ holds and therefore also that the mast is larger than 100 cm.

Step 2: The k-centered butterfly. Suppose $\mathscr{M}_k^n$ is defined. We call a state w minimal in $\mathscr{M}_k^n$ whenever there is no state w' such that $w >_i w'$ for some agent i. Observe that if w is minimal in $\mathscr{M}_k^n$ then there is a state w' such that $w' >_i w$ for $i \in \{\rho, \chi\}$. It also holds that for each such minimal state w, it is never the case that there are states w' and w'' such that $w' >_\rho w$ and $w'' >_\chi w$. In other words, a minimal state w is connected to other states w' by the relation of a unique agent $i \in \{\rho, \chi\}$. For that reason, we can label minimal states with that unique agent $i \in \{\rho, \chi\}$, and so call them i-terminating. $\mathscr{M}_k^{n+1}$ is then constructed as follows. For each i-terminating state w in $\mathscr{M}_k^n$, we add two additional states v and v' to W^n, and extend the relations $\geq_i^n$ and the valuation V^n as follows.

- Add all pairs $w >_j^{n+1} v$ and $w >_j^{n+1} v'$, with $j \neq i$, to the relevant $\geq_j^n$.
- Extend the valuation V^n with: $V^{n+1}([\bar{V}^n(w)+m)]) = \{v\}$, $V^{n+1}([\bar{V}^n(w)-m])) = \{v'\}$, $V^n(\top) = W^{n+1}$.

In this definition $\bar{V}^n(w)$ is the inverse of the valuation function V^n, giving us the number $k \in \mathbb{N}$ that is assigned to w in $\mathscr{M}_k^n$. If $\bar{V}^n(w) - m < 0$, we only add v. $\mathscr{M}_k$, the butterfly centered on k, is then defined as $\bigcup_{n\in\mathbb{N}} \mathscr{M}_k^n$. See Figure 1, Right. The induction steps thus extend the butterfly with two "wings", whose

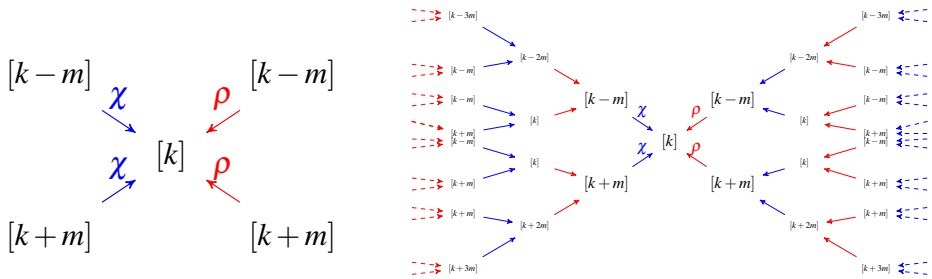


Figure 1: A k-centered butterfly (Right), and its body (Left). The red and blue arrows represent Roman's and Columba's plausibility orderings, respectively.

function is to capture the higher-order uncertainty of Roman and Columba. Indeed, as we shall also see below, moving along each of the paths starting at k generates a chain of "Roman/Columba considers it possible that Columba/Roman considers it possible... that the mast is l cm", even for l smaller than 100.

Step 3: The butterfly flutter. The last step is simply to put together the k-centered butterflies, for all $k \in \mathbb{N}$, to build a flutter. These copies will be used to define the conditional beliefs. A *butterfly flutter* $\mathscr{S}$ is the disjoint union $\biguplus_{k\in\mathbb{N}} \mathscr{M}_k$ of the k-centered butterflies for all $k \in \mathbb{N}$, together with a selection function f.

We can now define a selection function in detail as follows: For all w and $[k] \in Prop$, $f([k],w)$ is the center of the k-centered butterfly in $\mathscr{S}$. For $\top$, we set $f(\top,w)$ to be the center of the k-centered butterfly for the unique k such that $w \in V([k])$. Observe that by this construction all conditional and unconditional belief statements are evaluated at the center of a butterfly. Introducing a selection function as such allows us to evaluate within epistemic plausibility models reasons to believe a proposition A that is consistent with considering $\neg A$ possible.

Key facts: Take the butterfly flutter defined above, and w be the center of the butterfly centered on 300, and let m be the margin of error. The following facts capture the key points of our philosophical argument. Their proofs are elementary, so for reason of space we only sketch them here.

1. Both agents know that the mast is $300 \pm m$ cm tall: $\mathscr{S}, w \models K_\rho([300-m] \vee [300] \vee [300+m])$, and similarly for χ. Argument: clear by construction.
2. Both agents have an unconditional reason to believe that the mast is 300 cm tall: $\mathscr{S}, w \models R_\rho([300])$, and similarly for χ. Argument: $[300]$ is true at w, by the definition of f, we have $f(\top,w) = w$, and w is maximal in $[w]_\rho$.
3. Roman and Columba have iterated higher-order reasons to believe that the mast is $300 \pm m$ cm tall: $\mathscr{S}, w \models r^n_{\{\rho,\chi\}}([300-m] \vee [300] \vee [300+m])$. This also means $\mathscr{S}, w \models r^n_{\{\rho,\chi\}}([> 100])$. Argument: We have already seen that $f(\top,w) = w$ and that w is maximal in $[w]_\rho$ and $[w]_\chi$. So at every iteration of higher-order reasons to believe the only state considered is w itself.
4. Roman and Columba have the corresponding counterfactual, iterated higher-order reasons to believe that the mast is k cm tall for any $k \in \mathbb{N}$: For $i \in \{\rho,\chi\}$, we have $\mathscr{S}, w \models r^n(R_j([k])|[k])$. Argument: This is ensured by modifying for k the argument above for 300, together with the fact that $f(k,w)$ is the center of the relevant butterfly centered on k in $\mathscr{S}$.
5. $[> 100]$ is not common knowledge in the iterated sense, i.e., for all l such that $300 - (l.m) > 0$, there is r such that $\mathscr{S}, w \models \langle K_i \rangle \langle K^r_j \rangle [300 - (l.m)]$. Argument: recall that $v \approx_i v'$ is defined as $v \geq_i v'$ or $v' \geq_i v$. That the fact holds can then be verified by following the uppermost paths to the left and the right of the center in Figure 1, assuming that $k = 300$.

Facts 1 to 5 together flesh out the philosophical argument that we developed in Section 3. They show that iterative higher-order reasons to believe that $[> 100]$ is consistent, now in the precise, logical sense of the term, with, first, a failure of iterative common knowledge of that same proposition and, second, the relevant iterative, higher-order counterfactual reasons to believe that we described in Section 3. The fact that these counterfactual reasons to believe are consistent with $[> 100]$ at w shows that they do not detach.

5 Conclusion

We have argued both philosophically and formally that understanding publicity in terms of Lewisian common knowledge allows one to give a satisfactory analysis of the Sailboat example. The crux of the philosophical argument is the fact that having a reason to believe in a proposition A is consistent with considering it possible that A is false. This allows Lewisian common knowledge of $[> 100]$ to hold

even when iterative common knowledge of that proposition fails. The formal construction in Section 4 makes this intuition concrete by modeling, respectively, reasons to believe using plausibility orderings and knowledge using their connected components.

We should emphasize that these arguments should not be read as criticism of, or proposal to "replace" the iterative or fixed-point definitions of common knowledge with its Lewisian counterpart. These definitions have been extremely helpful in a number of areas, in particular for understanding the type of "back-and-forth" reasoning underlying game-theoretical solution concepts [32, 29]. What the argument suggests is that Lewisian common knowledge is perhaps better suited than iterative common knowledge to capture precisely the pre-theoretical notion of publicity, and that this should be taken into account when analyzing concrete cases that involve the latter. Of course, there remain numerous other cases where this view should be put to test: e.g. the classical Byzantine generals example, the Halpern-Moses problem, the consecutive numbers example, or philosophical accounts of collective agency that rely on iterative common knowledge. Other logical framework could be used to represent those cases, for instance the ones developed in [2] or in [14, 6] We leave those for future work, but conjecture that Lewisian common knowledge would prove illuminating in those cases too.

Acknowledgements

The authors would like to thank the anonymous reviewers of TARK for helpful comments and suggestions, as well as the participants of the CELIA project meeting 2025. Financial support of the DFG-GACR project CELIA (RO 4548/13-1) is gratefully acknowledged.

References

[1] Carlos E Alchourrón, Peter Gärdenfors & David Makinson (1985): *On the logic of theory change: Partial meet contraction and revision functions*. The journal of symbolic logic 50(2), pp. 510–530, doi:10.2307/2274239.

[2] Alexandru Baltag & Johan van Benthem (2018): *Some Thoughts on the Logic of Imprecise Observation*. Tsinghua University Journal of Philosophy, Issue 2018.

[3] Alexandru Baltag & Sonja Smets (2016): *A Qualitative Theory of Dynamic Interactive Belief Revision*. Readings in Formal Epistemology: Sourcebook, pp. 813–858, doi:10.1007/978-3-319-20451-2-39.

[4] Jon Barwise (1988): *Three views of common knowledge*. In: Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge, pp. 365–379, doi:10.1007/978-3-319-20451-2-37.

[5] Jon Barwise (1989): *The Situation in Logic*. CSLI lecture notes.

[6] Marta Bílková, Hans van Ditmarsch, Roman Kuznets & Rojo Randrianomentsoa (2024): *Bisimulation for Impure Simplicial Complexes*. arXiv preprint arXiv:2406.16785, doi:10.48550/arXiv.2406.16785.

[7] Olle Blomberg (2016): *Common knowledge and reductionism about shared agency*. Australasian Journal of Philosophy 94(2), pp. 315–326, doi:10.1080/00048402.2015.1055581.

[8] Michael E Bratman (2013): *Shared agency: A planning theory of acting together*. Oxford University Press.

[9] John Broome (2005): *Does rationality give us reasons?* Philosophical Issues 15, pp. 321–337, doi:10.1111/j.1533-6077.2005.00069.x.

[10] Michael Suk-Young Chwe (2013): *Rational ritual: Culture, coordination, and common knowledge*. Princeton University Press, doi:10.23943/princeton/9780691158280.001.0001.

[11] Herbert H. Clark & Catherine R. Marshall (1981): *Definite Knowledge and Mutual Knowledge*. In Aravind K. Joshi, Bonnie L. Webber & Ivan A. Sag, editors: *Elements of Discourse Understanding*, Cambridge University Press, pp. 10–63.

[12] Robin P Cubitt & Robert Sugden (2003): *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*. *Economics & Philosophy* 19(2), pp. 175–210, doi:10.1017/S0266267103001123.

[13] Franz Dietrich & Christian List (2013): *A reason-based theory of rational choice*. *Nous* 47(1), pp. 104–134, doi:10.1111/j.1468-0068.2011.00840.x.

[14] Hans van Ditmarsch (2021): *Wanted dead or alive: Epistemic logic for impure simplicial complexes*. In: *Logic, Language, Information, and Computation: 27th International Workshop, WoLLIC 2021, Virtual Event, October 5–8, 2021, Proceedings 27*, Springer, pp. 31–46, doi:10.1007/978-3-030-88853-4_3.

[15] John Geanakoplos (1994): *Common knowledge*. *Handbook of game theory with economic applications* 2, pp. 1437–1496, doi:10.1016/S1574-0005(05)80072-4.

[16] Margaret Gilbert (1992): *On social facts*. Princeton University Press, doi:10.1515/9780691214627.

[17] Margaret Gilbert (2007): *Mutual recognition, common knowledge, and joint attention*. *Hommage à Wlodek: Philosophical papers dedicated to Wlodek Rabinowicz*, pp. 1–21.

[18] Yannai A Gonczarowski & Yoram Moses (2023): *Common Knowledge, Regained*. *arXiv preprint arXiv:2311.04374*, doi:10.48550/arXiv.2311.04374.

[19] Joseph Y Halpern (2001): *Substantive rationality and backward induction*. *Games and Economic Behavior* 37(2), pp. 425–435, doi:10.1006/game.2000.0838.

[20] Joseph Y Halpern & Yoram Moses (1990): *Knowledge and common knowledge in a distributed environment*. *Journal of the ACM (JACM)* 37(3), pp. 549–587, doi:10.1145/79147.79161.

[21] Dick de Jongh & Fenrong Liu (2009): *Preference, priorities and belief*. In: *Preference change: Approaches from philosophy, economics and psychology*, Springer, pp. 85–107, doi:10.1007/978-90-481-2593-7_4.

[22] Harvey Lederman (2018): *Uncommon knowledge*. *Mind* 127(508), pp. 1069–1105, doi:10.1093/mind/fzw072.

[23] David Lewis (2008): *Convention: A philosophical study*. John Wiley & Sons.

[24] John McCarthy, Masahiko Sato, Takeshi Hayashi & Shigeru Igarashi (1978): *On the model theory of knowledge*. Technical Report.

[25] David McNaughton & Piers Rawling (2018): *Motivating reasons and normative reasons*. *The Oxford handbook of reasons and normativity*, pp. 171–196, doi:10.1093/oxfordhb/9780199657889.013.0008.

[26] Paul Milgrom (1981): *An axiomatic characterization of common knowledge*. *Econometrica: Journal of the Econometric Society*, pp. 219–222, doi:10.2307/1911137.

[27] Stephen Morris (2014): *Coordination, timing and common knowledge*. *Research in Economics* 68(4), pp. 306–314, doi:10.1016/j.rie.2014.04.004.

[28] Martin J Osborne & Ariel Rubinstein (1994): *A course in game theory*. MIT press.

[29] Eric Pacuit & Olivier Roy (2017): *Epistemic Foundations of Game Theory*. In Edward N. Zalta, editor: *The Stanford Encyclopedia of Philosophy*, Summer 2017 edition, Metaphysics Research Lab, Stanford University.

[30] Rohit Parikh (2005): *Logical omniscience and common knowledge: WHAT do we know and what do WE know?* In: *TARK*, 5, Citeseer, pp. 62–77, doi:10.1145/1089933.1089943.

[31] Cedric Paternotte (2011): *Being realistic about common knowledge: a Lewisian approach*. *Synthese* 183(2), pp. 249–276, doi:10.1007/s11229-010-9770-y.

[32] Andrés Perea (2012): *Epistemic game theory: reasoning and choice*. Cambridge University Press, doi:10.1017/CBO9780511844072.

[33] Giacomo Sillari (2008): *Common knowledge and convention*. Topoi 27, pp. 29–39, doi:10.1007/s11245-008-9030-7.

[34] Robert Stalnaker (1996): *Knowledge, belief and counterfactual reasoning in games*. Economics & Philosophy 12(2), pp. 133–163, doi:10.1017/S0266267100004132.

[35] Johan van Benthem (2011): *Logical dynamics of information and interaction*. Cambridge University Press, doi:10.1017/CBO9780511974533.

[36] Huub Vromen (2023): *Reasoning with reasons: Lewis on common knowledge*. Economics & Philosophy, pp. 1–22, doi:10.1017/S0266267123000238.

[37] J Robert G Williams (2023): *Publicity and Common Commitment to Believe*. Erkenntnis 88(3), pp. 1059–1080, doi:10.1007/s10670-021-00393-x.

[38] Timothy Williamson (1997): *Knowledge as evidence*. Mind 106(424), pp. 717–741, doi:10.1093/mind/106.424.717.

Graded Distributed Belief

Emiliano Lorini
IRIT, CNRS, Toulouse University
Toulouse, France
Emiliano.Lorini@irit.fr

Dmitry Rozplokhas
TU Wien
Vienna, Austria
dmitry@logic.at

We introduce a new logic of graded distributed belief that allows us to express the fact that a group of agents distributively believe that a certain fact φ holds with at least strength k. We interpret our logic by means of computationally grounded semantics relying on the concept of belief base. The strength of the group's distributed belief is directly computed from the group's belief base after having merged its members' individual belief bases. We illustrate our logic with an intuitive example, formalizing the notion of epistemic disagreement. We also provide a sound and complete Hilbert-style axiomatization, decidability result obtained via filtration, and a tableaux-based decision procedure that allows us to state PSPACE-completeness for our logic.

1 Introduction

The idea of using belief bases as formal semantics for multi-agent epistemic logic was first introduced in [26] and further developed in [27, 28]. This approach aligns with the sentential (or syntactic) perspective on knowledge representation [21, 13, 33, 20], which holds that an agent's body of knowledge should be represented as a set of sentences in a formal language. The key novelty of belief base semantics, compared to traditional epistemic logic semantics based on multi-relational Kripke models [31, 12], lies in two main aspects. First, a possible world (or state) in a model is not treated as a primitive entity but is instead composed of the agents' belief bases and a valuation of propositional atoms. Second, the agents' accessibility relations are not explicitly part of the model but are determined *a posteriori* from their belief bases. Specifically, in this semantics, an agent at state S considers state S' possible (or epistemically accessible) if and only if S' satisfies all the formulas in the agent's belief base at S. This decomposition of a state into more fundamental elements is shared by various approaches in symbolic model checking and computationally grounded semantics for epistemic logic. These include frameworks based on interpreted systems, where a global state is decomposed into individual agents' local states [10, 25], as well as those that rely on the primitive notion of observability or visibility [17, 6, 14, 4]. At the language level, the belief base approach distinguishes explicit (or actual) belief from implicit (or potential) belief. The distinction between explicit and implicit belief has been widely discussed in the literature [23, 9]. In the belief base approach, this distinction is based on the concept of *deducibility*: explicit beliefs are those directly stored in an agent's belief base, while implicit beliefs consist of any information that can be logically inferred from those explicit beliefs.

In two subsequent works, the belief base approach has been shown to successfully represent notions of distributed and common belief [15, 29], as well as graded belief [30]. On the one hand, belief base semantics allows for a natural distinction between explicit and implicit distributed belief. While the explicit distributed belief of a group is given by the merging of the belief bases of its members, the implicit distributed belief corresponds to what can be deduced from the (collective) belief base resulting from this merging. On the other hand, the approach allows us to define a natural notion of the degree (or strength) of an agent's implicit belief that φ, understood as the maximum number of pieces of information

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 203–215, doi:10.4204/EPTCS.437.19

© E. Lorini & D. Rozplokhas
This work is licensed under the
Creative Commons Attribution License.

that can be removed from the agent's belief base without preventing the agent from deducing φ from their explicit beliefs.

In this paper, we present a generalization of the belief base semantics for epistemic logic we introduced in [26, 28]. In the original semantics, agents' belief bases were simply sets of formulas built from a language including propositional facts and explicit beliefs: an agent's belief base could include both information about the world and information about other agents' belief bases. Our generalization moves from plain (ungraded) belief bases to graded belief bases by using a multiset representation. In a graded belief base, each piece of information is associated with a natural number representing the strength of the agent's explicit belief, with 0 associated to a formula α meaning that the agent has no explicit belief that α. We use this more general semantics to define a novel notion of *graded distributed belief*, as a piece of information φ that a group can deduce from their collective belief base with a given strength k. Given a graded belief base for each agent in a group, we compute the group's graded distributed belief in two steps. First, we merge the graded belief bases of the group's members to obtain a collective graded belief base. The degree that the group assigns to a formula α corresponds to the sum of the degrees that each member assigns to α. Second, we compute the group's degree of distributed belief in a certain fact φ as the amount of information that can be removed from the group's belief base without preventing it from deducing φ. We also show how our framework can be used to define a quantitative notion of of epistemic disagreement within a group, based on the amount of information that must be removed from the group's collective belief base to restore consistency. This notion of disagreement bears similarities to the measure of inconsistency defined in [19], namely, the minimal number of formulas that need to be removed from a belief base to restore consistency.

The paper is organized as follows. In Section 2, we first present the general framework: the graded belief base semantics, the modal language for representing implicit graded distributed belief, its semantic interpretation and an example illustrating our framework as well as the notion of epistemic disagreement within a group that can be defined in our framework. Following [28], in Section 3, we introduce an alternative Kripke-style semantics for our modal language, which serves as a technical tool for investigating the proof-theoretic aspects of our framework. Section 4 presents a Hilbert-style axiomatics for our logic of graded distributed belief. Then, in Section 5 we present a decision procedure based on tableaux, which allows us to establish PSPACE-completeness for our logic.

Before turning to the core of the paper, we briefly discuss some related work. Although the notion of graded distributed belief and the graded belief base semantics used to interpret it, as introduced in this paper, are new, the notion of plain distributed belief has been widely investigated in epistemic logic [12, 36, 1, 32, 7, 24, 2, 11]. Moreover, the idea of having graded belief modalities for individual agents was explored in previous work [22], in line with work in ranking theory [34]. Other approaches employ graded modalities for individual agents, where the degree of belief is determined either by the number of worlds in which the believed formula holds true [18, 16, 5], or by the amount of evidence supporting it [3]. The notion of graded belief base is also used in possibility theory [8]. In the present paper, we generalize it to the multi-agent setting and to nested beliefs.

2 Framework

In this section, we present our graded belief base semantics and show how to use it to compute graded doxastic accessibility relations for agents and groups. Then, we introduce a modal language of graded distributed belief and interpret it using the semantics. We illustrate our language and semantics with the help of a concrete example.

2.1 Notation

In this paper we will work with graded sets (or multisets), where grade (or cardinality) of each element is either a natural number (including zero) or infinity (denoted ω). To avoid confusion we use notation $\mathbb{N}_0$ and $\mathbb{N}_1$ for natural numbers with and without zero respectively, and $\mathbb{N}_0^\omega$ and $\mathbb{N}_1^\omega$ for their extensions with element ω. We represent a graded set over set X by a function $f\colon X \to \mathbb{N}_0^\omega$, and define the *support* of this graded set as $Supp(f) = \{x \in X \mid f(x) > 0\}$. We denote the set of all multisets over X by $\mathscr{M}(X)$.

For $X \subseteq \mathbb{N}_0^\omega$ we will use notation $\min^* X = \min(X \cup \{0\})$ and $\max^* X = \max(X \cup \{\omega\})$ to avoid dealing with the case of empty set. We will also consider potentially infinite sums of grades. Since grades are natural numbers (or inifinity) such sums have natural well-behaved definition: we define it as sum of non-zero summands if there are finitely many such summands and none of them is ω, and as ω otherwise. Finally, we will use the notion of *partitions*, i.e. functions dividing the given grade $k \in \mathbb{N}_0$ among agents in group J: $\mathscr{P}art(J,k) = \{\delta : J \to \mathbb{N}_0 \mid \sum_{i \in J} \delta(i) = k\}$.

2.2 Semantics

We are going to present a belief base semantics for epistemic attitudes of agents that generalizes the belief base semantics introduced in [26, 28] to multisets. Multisets are used to represent agents' explicit beliefs with their strengths, weight or epistemic importance. Unlike the standard Kripke semantics for epistemic logic in which the notions of epistemic alternative and plausibility of a world (or state) are given as primitive, in this semantics they are defined from the primitive concept of graded belief base.

Assume a countably infinite set of atomic propositions Atm and a finite set of agents $Agt = \{1, \ldots, n\}$. The set of non-empty groups is denoted by $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$. We define the language $\mathscr{L}_0(Atm, Agt)$ for representing agents' graded explicit beliefs by the following grammar:

$$\alpha \quad ::= \quad p \mid \neg\alpha \mid \alpha \wedge \alpha \mid \triangle_i^k \alpha,$$

where p ranges over Atm, i ranges over Agt and k ranges over $\mathbb{N}_1^\omega$. The formula $\triangle_i^k \alpha$ is read "agent i explicitly believes that α with at least degree k". For notational convenience, we abbreviate $\triangle_i \alpha \stackrel{def}{=} \triangle_i^1 \alpha$. The formula $\triangle_i \alpha$ is simply read "agent i explicitly believes that α".

For notational convenience we write $\mathscr{L}_0$ instead of $\mathscr{L}_0(Atm, Agt)$, when the context is unambiguous.

Definition 1. *A state is a tuple $S = (\mathscr{B}_1, \ldots, \mathscr{B}_n, V)$ where for every $i \in Agt$, $\mathscr{B}_i \in \mathscr{M}(\mathscr{L}_0)$ is agent i's graded belief base, and $V \subseteq Atm$ is the actual environment. The set of all states is denoted by* **S**.

$\mathscr{B}_i$ assings a weight $\mathscr{B}_i(\alpha)$ to each formula from $\alpha \in \mathscr{L}_0$, capturing the strength of agent i's explicit belief that α.

The language $\mathscr{L}_0$ is interpreted with respect to states, as follows.

Definition 2. *Let $S = (\mathscr{B}_1, \ldots, \mathscr{B}_n, V) \in$ **S**. Then:*

$$\begin{aligned}
S \models p &\iff p \in V, \\
S \models \neg\alpha &\iff S \not\models \alpha, \\
S \models \alpha_1 \wedge \alpha_2 &\iff S \models \alpha_1 \text{ and } S \models \alpha_2, \\
S \models \triangle_i^k \alpha &\iff \mathscr{B}_i(\alpha) \geq k.
\end{aligned}$$

Observe in particular the following interpretation of the graded explicit belief operator: agent i explicitly believes that α with at least degree k if and only if the information α has an importance for the agent at least equal to k.

From the agents' graded belief bases $\mathscr{B}_1,\ldots,\mathscr{B}_n$ it is natural to compute the collective graded belief base $\mathscr{B}_J \in \mathscr{M}(\mathscr{L}_0)$ of a group J: the degree of explicit belief of the group is equal to the sum of the degrees of beliefs of the group's members.

Definition 3. *Let $S = (\mathscr{B}_1,\ldots,\mathscr{B}_n,V) \in \mathbf{S}$ and $J \in 2^{Agt*}$. Then, $\mathscr{B}_J(\alpha) = \sum_{i\in J} \mathscr{B}_i(\alpha)$ for every $\alpha \in \mathscr{L}_0$.*

This definition extends the notion of "pooling" belief bases from [15] to the weighted case, assuming agent independence and epistemic egalitarianism. Note that we do not rely on more involved forms of belief base aggregation aimed at preserving consistency, since in our framework individual belief bases may already be inconsistent.

The following definition introduces the notion of graded doxastic alternative for a group.

Definition 4. *Let $J \in 2^{Agt*}$ and let $k \in \mathbb{N}_0$. Then, $\mathscr{R}^k_J$ is the binary relation on the set $\mathbf{S}$ such that, for all $S = (\mathscr{B}_1,\ldots,\mathscr{B}_n,V), S' = (\mathscr{B}'_1,\ldots,\mathscr{B}'_n,V') \in \mathbf{S}$:*

$$S\mathscr{R}^k_J S' \text{ if and only if } \sum_{\substack{\alpha\in\mathscr{L}_0 \\ S'\not\models\alpha}} \mathscr{B}_J(\alpha) \leq k.$$

$S\mathscr{R}^k_J S'$ means that, from the point of view of the group J at state S, state S' is at most k-implausible, where the degree of implausibility of a state for a group is equal to the weighted sum of the group's explicit beliefs that are not satisfied at the state. This means that the degree of implausibility of a state for a group depends on i) how much information that the group has in its belief base is not satisfied at the state, and ii) how important is that information for the group.

Notice that $S\mathscr{R}^k_J S'$ can also be interpreted as the fact that state S' is considered possible for the group J after removing from its collective belief base a body of information of importance at most equal to k. Indeed, $\sum_{\substack{\alpha\in\mathscr{L}_0 \\ S'\not\models\alpha}} \mathscr{B}_J(\alpha)$ can also be conceived as the total amount of importance for the group J at state S of the information that is not satisfied at state S'.

A graded doxastic accessibility relation $\mathscr{R}^k_J$ induces a plausibility ordering over states, as in [34, 22]. For notational convenience, we write $\mathscr{R}_J$ instead of $\mathscr{R}^0_J$. Clearly, $S\mathscr{R}_J S'$ if and only iff $\forall\alpha \in \mathscr{L}_0$, if $\mathscr{B}_J(\alpha) > 0$ then $S' \models \alpha$. In words, a state is 0-implausible from the point of view of a group if it satisfies all information in the group's belief base.

Before concluding this section, we define the concept of a model as a state supplemented with a set of states, called *context*. The latter includes all states compatible with the the agents' common ground [35], i.e., the body of information that the agents commonly believe to be the case.

Definition 5. *A multi-agent graded belief model (MAGBM) is a pair (S,U), where $S \in \mathbf{S}$ and $U \subseteq \mathbf{S}$. The class of models is denoted by* $\mathbf{M}$.

2.3 Language

We consider a modal language $\mathscr{L}(Atm,Agt)$ that extends the language $\mathscr{L}_0(Atm,Agt)$ given above with graded distributed belief modalities. It is is defined by the following grammar:

$$\varphi \quad ::= \quad \alpha \mid \neg\varphi \mid \varphi\wedge\varphi \mid \Box^k_J\varphi,$$

where α ranges over $\mathscr{L}_0(Atm,Agt)$, J ranges over 2^{Agt*} and k ranges over $\mathbb{N}_0$. For notational convenience we write and $\mathscr{L}$ instead of $\mathscr{L}(Atm,Agt)$, when the context is unambiguous. The other Boolean constructions $\top$, $\bot$, $\rightarrow$ and $\leftrightarrow$ are defined in the standard way.

We interpret the modal language $\mathscr{L}$ relative to a model by means of the graded accessibility relations $\mathscr{R}^k_J$.

Definition 6. *Let* $(S,U) \in \mathbf{M}$. *Then:*

$$\begin{array}{rcl}(S,U) \models \alpha & \Longleftrightarrow & S \models \alpha,\\ (S,U) \models \neg\varphi & \Longleftrightarrow & S \not\models \varphi,\\ (S,U) \models \varphi_1 \wedge \varphi_2 & \Longleftrightarrow & S \models \varphi_1 \textit{ and } S \models \varphi_2,\\ (S,U) \models \Box^k_J \varphi & \Longleftrightarrow & \forall S' \in U, \textit{ if } S \mathscr{R}^k_J S' \textit{ then } (S',U) \models \varphi.\end{array}$$

The modal formula $\Box^k_J\varphi$ is read "group J would implicitly believe that φ, for every removal from its belief base of a body of information of importance at most equal to k". The value k can also be conceived as the extent to which group J distributively believes that φ. Indeed, the higher the importance of the information that can be removed from the group's belief base without affecting what the group can infer, the stronger the inference and so the group's resulting distributed belief. Thus, $\Box^k_J\varphi$ can also be read "group J has an implicit distributed belief that φ of degree (or strength) at least k". The abbreviation $\Diamond^k_J\varphi \stackrel{\text{def}}{=} \neg\Box^k_J\neg\varphi$ defines the concept of distributed belief compatibility. The formula $\Diamond^k_J\varphi$ has to be read "φ would be compatible with group J's explicit beliefs, for some removal from its collective belief base of a body of information of importance at most equal to k".

2.4 Conceptual Analysis

We are going to show how our language and semantics can be leveraged to formally represent graded distributed belief as well as degree of epistemic disagreement within a group of agents. The latter notion is formally defined by the following abbreviation: $Disagree(J,k) \stackrel{def}{=} \Box^{k-1}_J \bot$ for $k \geq 1$. $Disagree(J,k)$ means that within the group J there is an epistemic disagreement of at least strength k.

Example 1. *Ann, Bob, Cath and John are the four members of a research project evaluation committee. Their task is to decide whether a submitted project for funding can be included in the list of "fundable" projects or not. They are all convinced with at least strength* $k_0 > 0$ *that a project should be included in the list (in) if and only if its idea is innovative (id) and, at the same time, the project's consortium is of high scientific standard (hi). This hypothesis is captured by the following abbreviation:*

$$\alpha_1 \stackrel{def}{=} \bigwedge_{i\in\{Ann,Bob,Cath,John\}} \triangle^{k_0}_i \big(in \leftrightarrow (id \wedge hi)\big).$$

However, they have diverging opinions and, in some cases, have not yet formed an opinion regarding these qualities of the project. In particular, Ann explicitly believes with degree $k_1 > 0$ *that the project's idea is innovative, and Cath believes the opposite with degree* $k_3 > 0$*, Bob explicitly believes with degree* $k_2 > 0$ *that the project's consortium is of high scientific standard, and John explicitly believes the opposite with degree* $k_4 > 0$*. This hypothesis is captured by the following abbreviation:*

$$\alpha_2 \stackrel{def}{=} \triangle^{k_1}_{Ann} id \wedge \triangle^{k_2}_{Bob} hi \wedge \triangle^{k_3}_{Cath} \neg id \wedge \triangle^{k_4}_{John} \neg hi.$$

It is routine to verify that group $\{Ann,Bob\}$ *implicitly believes with degree* $(\min\{2k_0,k_1,k_2\}-1)$ *that the project should be included in the list, while group* $\{Cath,John\}$ *believes the opposite with degree* $(\min\{2k_0,k_3+k_4\}-1)$*:*

$$\models (\alpha_1 \wedge \alpha_2) \;\rightarrow\; (\Box^{\min\{2k_0,k_1,k_2\}-1}_{\{Ann,Bob\}} in \;\wedge\; \Box^{\min\{2k_0,k_3+k_4\}-1}_{\{Cath,John\}} \neg in).$$

Moreover, when the explicit information is restricted to $\alpha_1 \wedge \alpha_2$ there is no disagreement within these groups (i.e. there exist models where $\alpha_1 \wedge \alpha_2$ is true and $Disagree(J,1)$ is false), while all four agents together have an epistemic disagreement of at least strength $(\min\{k_1,k_3\}+\min\{k_2,k_4\})$:

$$\begin{array}{lll} \not\models (\alpha_1 \wedge \alpha_2) & \rightarrow & (Disagree(\{Ann,Bob\},1) \vee Disagree(\{Cath,John\},1)) \\ \models (\alpha_1 \wedge \alpha_2) & \rightarrow & Disagree(\{Ann,Bob,Cath,John\},\min\{k_1,k_3\}+\min\{k_2,k_4\}). \end{array}$$

3 Alternative semantics

To explore the proposed logic we will follow the general approach of [26] and introduce two alternative equivalent semantical characterizations: *notional* and *quasi-notional graded doxastic models*.

First, we redefine models in terms closer to Kripke semantics for modal logics by introducing the notion of distance between states for a given group of agents as the sum of degrees in the first world of all beliefs not satisfied in the second world for all agents in the group. Note that this notion of distance is not (in general) symmetric.

Definition 7. *A* notional graded doxastic model (NGDM) *is a tuple $M = (W, \mathscr{D}, \rho, \mathscr{V})$ where W is a set of worlds, $\mathscr{D}\colon Agt \times W \to \mathscr{M}(\mathscr{L}_0)$ is a doxastic function, $\rho\colon 2^{Agt*} \times W \times W \to \mathbb{N}_0^{\omega}$ is a distance function, and $\mathscr{V}\colon \mathbb{P} \to 2^W$ is a valuation, such that:*

$$\rho(J,w,u) = \sum_{\substack{\alpha \in \mathscr{L}_0 \\ (M,u) \not\models \alpha}} \sum_{i \in J} \mathscr{D}(i,w)(\alpha) \qquad \text{(NGDM-DOX)}$$

with satisfaction relation defined as follows:

$$\begin{array}{lll} (M,w) \models p & \Leftrightarrow & w \in \mathscr{V}(p) \\ (M,w) \models \neg\varphi & \Leftrightarrow & (M,w) \not\models \varphi \\ (M,w) \models \varphi_1 \wedge \varphi_2 & \Leftrightarrow & (M,w) \models \varphi_1 \text{ and } (M,w) \models \varphi_2 \\ (M,w) \models \triangle_i^k \alpha & \Leftrightarrow & \mathscr{D}(i,w)(\alpha) \geq k \\ (M,w) \models \Box_J^k \varphi & \Leftrightarrow & \forall u \in W : \rho(J,w,u) \leq k \Rightarrow (M,u) \models \varphi \end{array}$$

M is called finite *when W is finite and $Supp(\mathscr{D}(i,w))$ is finite for every i and w.*

Notice that condition (NGDM-DOX) fully defines distance function ρ via doxastic function $\mathscr{D}$ and the resulting distance function is *additive on groups*:

$$\rho(J,w,u) = \sum_{i \in J} \rho(\{i\},w,u) \qquad \text{(NGDM-}\rho\text{-ADD)}$$

However, condition (NGDM-DOX) itself is not axiomatizable (see Remark 1 below), therefore we also introduce the notion of *quasi-notional graded doxastic models (QNGDMs)* where this condition is weakened to become axiomatizable. First, instead of equality, we require that the distance is no smaller than the sum of degrees of the unsatisfied beliefs (condition (QNGDM-DOX)). Additionally, we require a weakened version of (NGDM-ρ-ADD) (condition (QNGDM-ρ-ADD)): the finite distance for every group J can be partitioned into summands $\delta(i)$ for every agent $i \in J$ such that $\delta(i)$ is at least the distance for any agent $i \in J$, and moreover the distance for any sub-group does not exceed the sum of $\delta(i)$ for the agents involved.

Definition 8. *A* quasi-notional graded doxastic model (QNGDM) *is a tuple $M = (W, \mathscr{D}, \rho, \mathscr{V})$ where W, ρ, $\mathscr{D}$, $\mathscr{V}$ are as in Def. 7 except that (NGDM-DOX) is replaced by the two following weaker conditions*

for every $J \in 2^{Agt*}$ *and* $w,u \in W$ *if* $\rho(J,w,u) \neq \omega$*:*

$$\rho(J,w,u) \geq \sum_{\substack{\alpha \in \mathscr{L}_0 \\ (M,u) \not\models \alpha}} \sum_{i \in J} \mathscr{D}(i,w)(\alpha) \qquad \text{(QNGDM-DOX)}$$

$$\exists \delta \in \mathscr{P}art(\rho(J,w,u),J) \text{ such that for any non-empty } J' \subset J, \sum_{i \in J'} \delta(i) \geq \rho(J',w,u) \quad \text{(QNGDM-}\rho\text{-ADD)}$$

We will now show equivalence of the three semantics following the strategy of [28]: we will first show that semantics of QNGDMs satisfies the finite model property (Lem. 1), then we will show how every satisfying *finite* QNGDM can be transformed into a satisfying NGDM (Lem. 2), satisfying NGDM — into a satisfying MAGBM (Lem. 3), and a satisfying MAGBM into a satisfying QNGDM (Lem. 4), closing the circle.

To establish Finite Model Property for QNGDMs we adapt the standard filtration technique.

Lemma 1. *If* M *is QNGDM satisfying* $\varphi \in \mathscr{L}$ *then there exists a finite QNGDM* M' *satisfying* φ*.*

Proof. (Sketch) We consider an equivalence relation $\equiv_\varphi$ on worlds of M, relating worlds with the same evaluation on all subformulas of φ. Then we transform $M = (W,\rho,\mathscr{D},\mathscr{V})$ into finite a QNGDM $(W',\rho',\mathscr{D}',\mathscr{V}')$ as follows: $W' = {}^{W}\!/_{\equiv_\varphi}$; $\rho'(J,U,V) \stackrel{def}{=} \min\{\rho(J,u,v) \mid u \in U, v \in V\}$; $\mathscr{D}'(i,U)(\alpha) \stackrel{def}{=} \max^*\{k \mid \triangle_i^k \alpha$ *is a subformula of* φ *and* $\mathscr{D}(i,u)(\alpha) \geq k$ *for all* $u \in U\}$; $\mathscr{V}'(p) \stackrel{def}{=} \{U \mid U \subseteq \mathscr{V}(p)\}$. We check that M' preserves evaluation on subformulas of φ and satisfies both conditions of QNGDMs. □

Now we show how to transform a finite QNGDM into a (finite) NGDM. We will adapt the idea of two-stage model tranformation for distributed belief bases from [15] to our graded setting. At the fist stage we achieve (NGDM-ρ-ADD) in a QNDM by creating copies of each world for every possible group and redefining distances for them on the basis of partitions given by condition (QNGDM-ρ-ADD). At the second stage we achieve (NGDM-DOX) by adapting the transformation from [30] to our case: introducing a fresh characterizing atom for each world and add such atoms as beliefs with the required degree to satisfy the equality in (NGDM-DOX).

Lemma 2. *If* M *is a finite QNGDM satisfying* $\varphi \in \mathscr{L}$ *then there exists a finite NGDM* M'' *satisfying* φ*.*

Proof. (Sketch) We first change the set of worlds W in M to $W' = W \times 2^{Agt*}$, keep $\mathscr{V}$ and $\mathscr{D}$ the same for each copy, and redefine distances for copies using δ from condition (QNGDM-ρ-ADD) for J, w and u: $\rho'(J',(w,J''),(u,J)) = \sum_{i \in J'} \delta(i)$ for $J' \subseteq J$ and $\rho'(J',(w,J''),(u,J)) = \omega$ otherwise. With such definition of distances, the transformed model M' preserves the satisfaction relation for each copy, trivially satisfies (NGDM-ρ-ADD), and still satisfies (QNGDM-DOX), since distances between copies did not decrease w.r.t. original distances (by condition (QNGDM-ρ-ADD)). At the second stage, for each $w' \in W'$ we select a distinctive atom $\chi(w')$ not appearing in φ and in any $\alpha \in Supp(\mathscr{D}'(i,w'))$ in M' (which we can do since M' is finite), and change $\mathscr{V}'$ to $\mathscr{V}''$ such that $\mathscr{V}''(\chi(w')) = W' \setminus \{w'\}$ (not changing the valuation on other atoms). Then we change the degrees of these atoms: $\mathscr{D}'(i,w)(\chi(u)) = \rho(i,w,u) - \sum_{\substack{\alpha \in \mathscr{L}_0 \\ (M',u) \not\models \alpha}} \mathscr{D}(i,w)(\alpha)$, turning inequality in (QNGDM-DOX) into equality for $M'' = (W',\rho',\mathscr{D}'',\mathscr{V}'')$. □

Remark 1. *Contrapositive reading of Lemma 2 implies that any formula from* $\mathscr{L}$ *satisfied in all NGDMs is also satisfied in all QNGDMs, which is a proper superclass of NDMs. Therefore there can be no characterizing axiom* $\xi \in \mathscr{L}$ *that is true exactly in the QNGDMs satisfying* (NGDM-DOX)*. On the other hand, as we will see in the next section, there is a characterizing axioms in this sense for properties* (QNGDM-DOX) *and* (QNGDM-ρ-ADD)*.*

We can also easily transform a satisfying NGDM into a satisfying MAGBM.

Lemma 3. *If φ is satisfied by some NGDM then φ is satisfied by some MAGBM.*

Proof. (Sketch) Each world can be mapped into a state (by reconstructing belief bases from the doxastic function) and the evaluation will be preserved thanks to the condition (NGDM-DOX). □

And, finally, we can move from MAGBMs back to QNGDMs, closing the cycle.

Lemma 4. *If φ is satisfied by some MAGBM then φ is satisfied by some QNGDM.*

Proof. (Sketch) The doxastic function is defined by belief bases, and the distances can be defined via condition (NGDM-DOX), apart from the distances to the initial state S defined as ω (to reflect that, in general, S does not belong to the context U). □

Thus, we have established the equivalence of all three introduced semantics (MAGBMs, NGDMs, and QNGDMs), which will be crucial for axiomatization of the proposed logic. Moreover, the described transformations turn an arbitrary QNGDM into an NGDM of exponential size, which implies decidability of the proposed logic (since all NGDMs of bounded size can be checked for satisfaction in finite time).

4 Axiomatics

In this section, we present an axiomatization for the proposed logic **LGDDA** (Logic of Graded Distributed Doxastic Attitudes) based on the semantics of QNGDMs introduced in the previous section.

Definition 9. *Logic **LGDDA** extends the classical propositional logic by the following axioms and rules:*

$$\frac{\varphi}{\Box_J^k \varphi} \qquad (\mathbf{Nec}_{\Box_J^k})$$

$$\Box_J^k(\varphi \to \psi) \to (\Box_J^k \varphi \to \Box_J^k \psi) \qquad (\mathbf{K}_{\Box_J^k})$$

$$\triangle_i^k \alpha \to \triangle_i^{k'} \alpha \quad \text{if } k \geq k' \qquad (\mathbf{Mon}_{\triangle_i^l})$$

$$\Big(\bigwedge_{\triangle_i^{k'}\alpha \in \Omega} \triangle_i^{k'}\alpha\Big) \to \Box_J^k \bigvee_{\substack{\Omega' \subseteq \Omega \\ Sum(\Omega') \leq k}} \bigwedge_{\triangle_i^{k'}\alpha \in \Omega \setminus \Omega'} \alpha \quad \begin{array}{l} \text{if } \Omega \subseteq \{\triangle_i^{k'}\alpha \mid i \in J\} \\ \text{and } \triangle_i^{k'}\alpha, \triangle_i^{k''}\alpha \in \Omega \Rightarrow k' = k'' \\ \text{where } Sum(\Omega') = \sum_{\triangle_i^{k'}\alpha \in \Omega'} k' \end{array} \qquad (\mathbf{Int}_{\triangle_i^{k'}, \Box_J^k})$$

$$\Big(\bigwedge_{\Box_{J'}^{k'}\psi \in \Psi} \Box_{J'}^{k'}\psi\Big) \to \Box_J^k \bigvee_{\delta \in \mathscr{P}art(k,J)} \bigwedge_{\substack{\Box_{J'}^{k'}\psi \in \Psi \\ \sum_{i \in J'} \delta(i) \leq k'}} \psi \quad \text{if } \Psi \subseteq \{\Box_{J'}^{k'}\psi \mid J' \subseteq J\} \qquad (\mathbf{Int}_{\Box_{J'}^{k'}, \Box_J^k})$$

*$\varphi \in \mathscr{L}$ is said to be derivable from $\Gamma \subseteq \mathscr{L}$ in **LGDDA** (denoted $\Gamma \vdash_{\mathbf{LGDDA}} \varphi$) when there is finite $\Gamma_f \subseteq \Gamma$ such that formula $(\bigwedge_{\psi \in \Gamma_f} \psi) \to \varphi$ can be derived using axioms and rules of **LGDDA**.*

Rule ($\mathbf{Nec}_{\Box_J^k}$) and axiom ($\mathbf{K}_{\Box_J^k}$) reflect the fact that $\Box_J^k$ is a normal modality. Monotonicity axiom ($\mathbf{Mon}_{\triangle_i^l}$) reflects the fact that k in modality $\triangle_i^k$ gives a lower bound on the weight of the belief. Two final axioms determine the interaction between triangles and boxes, and between boxes with different groups and degrees respectively. Informally, axiom ($\mathbf{Int}_{\triangle_i^{k'}, \Box_J^k}$) states that group J believes with level k that their pulled beliefs are correct apart from some subset with the cummulative importance not exceeding k. Axiom ($\mathbf{Int}_{\Box_{J'}^{k'}, \Box_J^k}$) captures the fact that the distance d between states (as defined by (NGDM-DOX)) for

group J can be partitioned into distances $\delta(i)$ for $i \in J$ such that any graded belief for a subgroup J' with degree k' is preserved for group J with degree $k \geq d$, as long as k' is greater than the sum of distances $\delta(i)$ for $i \in J'$. Notice that the following two validities, defining monotonicity of boxes w.r.t. group and degrees, are instances of $(\mathbf{Int}_{\Box^{k'}_{J'},\Box^{k}_{J}})$ for the case when there is only one box on the left:

$$\Box^k_J \varphi \to \Box^{k'}_J \varphi \qquad \text{if } k \geq k' \qquad (\mathbf{Mon}^k_{\Box^k_J})$$

$$\Box^k_J \varphi \to \Box^k_{J'} \varphi \qquad \text{if } J \subseteq J' \qquad (\mathbf{Mon}^J_{\Box^k_J})$$

Remark 2. *Note that the axiomatization in Def. 9 does not simply adjust and merge axioms for graded beliefs and distributed beliefs. While* $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ *is a natural adjustment of axiom* $(\mathbf{Int}_{\Delta_i,\Box_i})$ *from [30] for the case of graded belief bases, the logic of distributed belief from [15] requires only axiom* $(\mathbf{Mon}^J_{\Box^k_J})$*, and our logic for distributed graded belief requires significantly more sophisticated (and perhaps less intuitive) axiom* $(\mathbf{Int}_{\Box^{k'}_{J'},\Box^k_J})$ *that reflects the combinatorics of partition of distances for groups and subgroups captured in condition* (QNGDM-ρ-ADD)*.*

The proposed axiomatization is sound and complete w.r.t. QNGDM model semantics. To prove completeness we construct a canonical QNGDM, using rule $(\mathbf{Nec}_{\Box^k_J})$ and axioms $(\mathbf{K}_{\Box^k_J})$ and $(\mathbf{Mon}_{\triangle^l_i})$ to establish the truth lemma, and axioms $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ and $(\mathbf{Int}_{\Box^{k'}_{J'},\Box^k_J})$ characterizing conditions (QNGDM-DOX) and (QNGDM-ρ-ADD) of QNGDMs respectively.

Theorem 1. $\Gamma \vdash_{\textbf{\textit{LGDDA}}} \varphi$ *iff* $\Gamma \vDash_{QNGDM} \varphi$ *(i.e. any pointed QDNM satisfying all formulas from* Γ *also satisfies* φ*).*

Proof. (Sketch) **Soundness:** Validity of axiom $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ in any QNGDM follows from the condition (QNGDM-DOX) (if $\rho(J,w,u) \leq k$ then the disjunct for $\Omega' = \{\triangle^{k'}_i \alpha \in \Omega \mid M,u \not\models \alpha\}$ will be satisfied in u) and validity of axiom $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ — from condition (QNGDM-ρ-ADD) (it ensures existence of partition δ that corresponds to a satisfied disjunct), validity of other axioms and preservation of validity by the rules is trivial.

Completeness: We construct a canonical model $M^C = (W^C, \mathscr{D}^C, \rho^C, \mathscr{V}^C)$, where W^C is the set of all **LGDDA**-maxiconsistent sets of formulas (i.e. maximal sets not deriving $\bot$), $\mathscr{D}^C(i,\Phi)(\alpha) = \max^*\{l\colon \triangle^l_i \alpha \in \Phi\}$; $\rho^C(J,\Phi,\Phi') = \min^*\{k\colon \Box^k_J \varphi \in \Phi \Rightarrow \varphi \in \Phi'\}$; $\mathscr{V}^C(p) = \{\Phi \in W^C\colon p \in \Phi\}$. The truth lemma stating $M^C,\Phi \models \psi \Leftrightarrow \psi \in \Phi$ is proved by structural induction on ψ, axiom $(\mathbf{Mon}_{\triangle^l_i})$ is used for the case of triangles, and the case of box follows from presence of $(\mathbf{K}_{\Box^k_J})$ and $(\mathbf{Nec}_{\Box^k_J})$ as usual. Then conditions (QNGDM-DOX) and (QNGDM-ρ-ADD) can be proved contrapositively using axioms $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ and $(\mathbf{Int}_{\Box^{k'}_{J'},\Box^k_J})$ respectively. Failure of (QNGDM-DOX) for some w, u, and J would allow us to select some finite set Ω of beliefs false in u with aggregated weight in w greater than $\rho(J,w,u)$, and applying axiom $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ we would conclude that some beliefs in Ω are true in u. Failure of (QNGDM-ρ-ADD) for some w, u, and J would imply existence of subgroup $J_\delta \subset J$ for every $\delta \in \mathscr{P}art(\rho^C(J,w,u),J)$ such that $d_\delta = \sum_{i\in J_\delta} \delta(i) < \rho^C(J_\delta,\Phi,\Phi')$, which also implies existence of $\psi_\delta \in \mathscr{L}$ such that $\Box^{d_\delta}_{J_\delta}\psi_\delta$ is satisfied in w but ψ_δ is not satisfied in u. Taking $\Omega = \{\Box^{d_\delta}_{J_\delta}\psi_\delta \mid \delta \in \mathscr{P}art(\rho^C(J,w,u),J)\}$ and applying axiom $(\mathbf{Int}_{\triangle^{k'}_i,\Box^k_J})$ to it, we can derive that some ψ_δ is satisfied in u, reaching contradiction. □

5 Tableau calculus and satisfiability checking

In this section, we present a tableau-based decision procedure for our logic and establish PSPACE-completeness of satisfiability checking (the same complexity as the basic logic of belief bases in [28]).

Definition 10. *The tableau calculus* $\mathbf{Tab}_{\mathbf{LGDDA}}$ *extends the standard tableau calculus* **LK** *for the classical logic with the following two rules:*

$$\frac{\{\triangle_i^k \alpha, \neg\triangle_i^{k+t}\alpha\} \cup X}{\{\bot\}}\ (\triangle\text{-}Mon) \qquad \frac{\{\neg\Box_J^k \varphi\} \cup X}{\{\neg\varphi\} \cup Y_1^{\Box\downarrow} \cup Y_1^{\Delta_J\downarrow} \mid \ldots \mid \{\neg\varphi\} \cup Y_N^{\Box\downarrow} \cup Y_N^{\Delta_J\downarrow}}\ (\Box\text{-}Elim)$$

$$\begin{array}{rll} \textit{where} & \{Y_1,\ldots,Y_N\} & = \{Y \subseteq X \mid \exists \delta \in \mathscr{P}art(k,J): \quad \Box_{J'}^{k'}\psi \in Y \Rightarrow \sum_{i\in J'}\delta(i) \le k' \textit{ and} \\ & & \qquad\qquad \mathscr{D}_{agg}(i, X\setminus Y) \le \delta(i)\ \forall i \in J\} \\ & \mathscr{D}_{agg}(i, X\setminus Y) & = \sum_{\alpha\in\mathscr{L}_0} \max^*\{k' \colon \triangle_i^{k'}\alpha \in X\setminus Y\} \\ & Y^{\Box\downarrow} \cup Y^{\Delta_J\downarrow} & = \{\psi \mid \Box_{J'}^{k'}\psi \in Y\} \cup \{\alpha \mid \triangle_i^{k'}\alpha \in Y, i \in J\} \end{array}$$

Each tableau rule states that the satisfiability of a set of formulas above the line (called *numerator*) implies satisfiability of at least one of the sets of formulas below the line separated by the symbol '—' (called *denominators*). We can use it to derive the non-satisfiability of some formula by applying the rules sequentially, with each branch ending with a set containing $\bot$ (thus unsatisfiable). Such tree-like derivations are called *closed tableaux*. For our logic, it is sufficient to add just one rule for each modality (in addition to standard rules for the classical connectives). The rule *($\triangle$-Mon)* captures the monotonicity of triangles w.r.t. grades, and the rule *($\Box$-Elim)* adapts the rule eliminating negative boxes to our logic: if $\{\neg\Box_J^k\varphi\} \cup X$ is satisfied in some world w of some model M then there should exist a world u with $\rho(J,w,u) \le k$ and a partition $\delta \in \mathscr{P}art(k,J)$, such that u satisfies $\neg\varphi$, all boxed formulas for subgroups with smaller distance, and some selection of triangled formulas in X, such that the cumulative degree of the rest of triangled formulas does not exceed k. Note that the rule *($\Box$-Elim)* has potentially exponentially many denominators. At the same time, in all rules (apart from the closing ones) the denominators contain only subformulas of formulas in the numerator, and the total number of connectives and modalities in each denominator decreases w.r.t. the numerator, so the length of each branch is linear w.r.t. size of the initial formula. The resulting calculus is sound and complete w.r.t. logic **LGDDA**.

Theorem 2. *A closed tableau in* $\mathbf{Tab}_{\mathbf{LGDDA}}$ *starting from* $\{\varphi\}$ *exists iff* φ *is not satisfiable in* **LGDDA**.

Proof. (Sketch) The soundness of all rules w.r.t. QNGDM semantics is straightforward. To prove completeness we show that for each set Γ underivable in $\mathbf{Tab}_{\mathbf{LGDDA}}$ (i.e. there are no closed tableaux starting from Γ) there is a pointed QNGDM model (M,w_0) satisfying all the formulas in Γ. We do it by induction on the number of connectives and modalities in Γ. First, if Γ contains formulas with double negation, conjunction or negated conjunction on the top level, they can be decomposed according to the rules and the statement follows straightforwardly from the inductive hypotheses. Otherwise Γ has only (possibly negated) atoms, $\triangle$-formulas and $\Box$-formulas. We start a model construction with one world w_0 which satisfy exactly atoms in $(\mathbb{P}\cap\Gamma)$ and with doxastic function $\mathscr{D}(i,w_0)(\alpha) = \max^*\{k' : \triangle_i^{k'}\alpha \in \Gamma\}$. It is easy to check that all formulas from $\mathscr{L}_0 \cap \Gamma$ will be satisfied in this world. To satisfy also all boxed formulas in Γ we use the inductive hypothesis for every negated box $\neg\Box_J^k\varphi$ in Γ: the application of the rule *($\Box$-Elim)* to that negated box should have at least one underivable denominator (with the corresponding partition δ) and by inductive hypothesis there is a pointed model (M',w_0'), satisfying all formulas in that denominator. Incorporating M' into M and defining distance from w_0 to w_0' on J and all its subgroups as the sum of the corresponding values of δ (and as ω for the non-subgroups) we ensure that $\Box_J^k\varphi$ is falsified in w_0 while all the non-negated boxes in Γ are satisfied. $\square$

Thus, to check satisfiability of a formula φ in **LGDDA** we can perform an exhaustive search for the closed tableaux in $\mathbf{Tab}_{\mathbf{LGDDA}}$ for $\{\varphi\}$, for which polynomial space w.r.t. size of φ is sufficient.[1]

Theorem 3. *Satisfiability checking in **LGDDA** is PSPACE-complete.*

Proof. (Sketch) PSPACE-hardness follows from PSPACE-hardenss of logic **LDA** [26], which is equivalent to a fragment of **LGDDA**. It can be solved in PSPACE by a proof search in $\mathbf{Tab}_{\mathbf{LGDDA}}$, starting from $\{\varphi\}$ and trying all possible rule applications. When trying the rule *($\Box$-Elim)* we iterate through all possible $Y \subseteq X$ and for each of them iterate through all partitions from $\mathscr{P}art(k,J)$, which both can be stored using polynomial space w.r.t. $|\varphi|$, and the depth of the proof search is also bounded polynomially. □

6 Conclusion

We have presented a proof-theoretic and complexity analysis of the notion of graded distributed belief, using a formal semantics based on graded belief bases. Following Spohn's ranking theory [34], we plan to study, in future work, a more general variant of the graded belief semantics based on ordinals instead of natural numbers. The notion of graded distributed belief we have defined is based on a counting view (Definition 4). In future work, we intend to investigate a qualitative version by replacing the counting view with a qualitative perspective based on set inclusion. Last but not least, we aim to move from a static to a dynamic setting by extending our framework with a notion of graded belief base change.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the ANR project EpiRL (grant number ANR-22-CE23-0029) and European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.

References

[1] T. Ågotnes & Y. N. Wáng (2017): *Resolving distributed knowledge*. *Artificial Intelligence* 252, pp. 1–21, doi:10.1016/J.ARTINT.2017.07.002.

[2] P. Balbiani & H. van Ditmarsch (2024): *Towards Dynamic Distributed Knowledge*. In A. Ciabattoni, D. Gabelaia & I. Sedlár, editors: *Advances in Modal Logic, AiML 2024, Prague, Czech Republic, August 19-23, 2024*, College Publications, pp. 125–146.

[3] P. Balbiani, D. Fernández-Duque, A. Herzig & E. Lorini (2019): *Stratified Evidence Logics*. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 1523–1529, doi:10.24963/IJCAI.2019/211.

[4] J. van Benthem, J. van Eijck, M. Gattinger & K. Su (2015): *Symbolic Model Checking for Dynamic Epistemic Logic*. In: *Proceedings of the 5th International Workshop on Logic, Rationality and Interaction (LORI 2015)*, *LNCS* 9394, Springer-Verlag, pp. 366–378, doi:10.1007/978-3-662-48561-3_30.

[1] Notice, that during the proof search we only need to store the grades appearing in the formulas and iterate through all partitions in $\mathscr{P}art(k,J)$ when applying the rule *($\Box$-Elim)*, which consist of at most $|Agt|$ numbers not exceeding k. Therefore, the representation of the grades inside the formulas and how we measure their size does not matter, as long as it is the same for the numbers in the iterated partitions. For simplicity, we assume that any numbers occupies only constant memory and the size of the formula is measured as the number of nodes in its tree-representation, but the same reasoning can be applied to more refined measures.

[5] K. Budzynska & M. Kacprzak (2008): *A Logic for Reasoning about Persuasion*. Fundamenta Informaticae 85, pp. 51–65, doi:10.3233/FUN-2008-851-405.

[6] T. Charrier, A. Herzig, E. Lorini, F. Maffre & F. Schwarzentruber (2016): *Building Epistemic Logic from Observations and Public Announcements*. In: Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016), AAAI Press, pp. 268–277.

[7] W. van Der Hoek, B. van Linder & J.-J. Meyer (1999): *Group knowledge is not always distributed (neither is it always implicit)*. Mathematical social sciences 38(2), pp. 215–240, doi:10.1016/S0165-4896(99)00013-X.

[8] D. Dubois, J. Lang & H. Prade (1994): *Possibilistic Logic*. In D. M. Gabbay, C. J. Hogger & J. A. Robinson, editors: Handbook of Logic in Artificial Intelligence and Logic Programming, 3, Oxford University Press, Oxford, pp. 439–513, doi:10.1093/oso/9780198537472.003.0009.

[9] R. Fagin & J. Y. Halpern (1987): *Belief, Awareness, and Limited Reasoning*. Artificial Intelligence 34(1), pp. 39–76, doi:10.1016/0004-3702(87)90003-8.

[10] R. Fagin, J. Y. Halpern, Y. Moses & M. Vardi (1995): *Reasoning about Knowledge*. MIT Press, Cambridge, doi:10.7551/mitpress/5803.001.0001.

[11] R. Galimullin & L. B. Kuijer (2024): *Varieties of Distributed Knowledge*. In Agata Ciabattoni, David Gabelaia & Igor Sedlár, editors: Advances in Modal Logic, AiML 2024, Prague, Czech Republic, August 19-23, 2024, College Publications, pp. 379–400.

[12] J. Y. Halpern & Y. 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.

[13] S. O. Hansson (1993): *Theory contraction and base contraction unified*. Journal of Symbolic Logic 58(2), pp. 602–625, doi:10.2307/2275221.

[14] A. Herzig, E. Lorini & F. Maffre (2015): *A Poor Man's Epistemic Logic Based on Propositional Assignment and Higher-Order Observation*. In: Proceedings of the 5th International Workshop on Logic, Rationality, and Interaction (LORI 2015), LNCS 9394, Springer, pp. 156–168, doi:10.1007/978-3-662-48561-3_13.

[15] A. Herzig, E. Lorini, E. Perrotin, F. Romero & F. Schwarzentruber (2020): *A Logic of Explicit and Implicit Distributed Belief*. In: Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020), Frontiers in Artificial Intelligence and Applications 325, IOS Press, pp. 753–760, doi:10.3233/FAIA200163.

[16] W. van der Hoek (1992): *Modalities for Reasoning about Knowledge and Quantities*. Ph.D. thesis, Free University of Amsterdam.

[17] W. van der Hoek, P. Iliev & M. Wooldridge (2012): *A logic of revelation and concealment*. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), IFAAMAS, pp. 1115–1122.

[18] W. van der Hoek & J.-J. Ch. Meyer (1991): *Graded modalities in epistemic logic*. Logique & Analyse 133-134, pp. 251–270.

[19] A. Hunter & S. Konieczny (2008): *Measuring Inconsistency through Minimal Inconsistent Sets*. In: Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR2008), AAAI Press, pp. 358–366.

[20] D. de Jongh & F. Liu (2009): *Preference, Priorities and Belief*. In T. Gruene-Yanoff & S. O. Hansson, editors: Preference Change, Springer, pp. 85–108, doi:10.1007/978-90-481-2593-7_4.

[21] K. Konolige (1986): *A deduction model of belief*. Morgan Kaufmann Publishers, Los Altos.

[22] N. Laverny & J. Lang (2004): *From Knowledge-Based Programs to Graded Belief-Based Programs Part I: On-Line Reasoning*. In: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), IOS Press, pp. 368–372.

[23] H. J. Levesque (1984): *A logic of implicit and explicit belief*. In: Proceedings of the Fourth AAAI Conference on Artificial Intelligence (AAAI'84), AAAI Press, pp. 198–202.

[24] J. Lindqvist, F. R. Velázquez-Quesada & T. Ågotnes (2022): *Cautious Distributed Belief*. In Carlos Areces & Diana Costa, editors: *Dynamic Logic. New Trends and Applications - 4th International Workshop, DaLí 2022, Haifa, Israel, July 31 - August 1, 2022, Revised Selected Papers*, *Lecture Notes in Computer Science* 13780, Springer, pp. 106–124, doi:10.1007/978-3-031-26622-5_7.

[25] A. Lomuscio, H. Qu & F. Raimondi (2017): *MCMAS: an open-source model checker for the verification of multi-agent systems*. *International Journal on Software Tools for Technology Transfer* 19, pp. 9–30, doi:10.1007/s10009-015-0378-x.

[26] E. Lorini (2018): *In Praise of Belief Bases: Doing Epistemic Logic Without Possible Worlds*. In S. A. McIlraith & K. Q. Weinberger, editors: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Press, pp. 1915–1922, doi:10.1609/aaai.v32i1.11560.

[27] E. Lorini (2019): *Exploiting Belief Bases for Building Rich Epistemic Structures*. In L. S. Moss, editor: *Proceedings of the Seventeenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2019)*, *EPTCS* 297, pp. 332–353, doi:10.4204/EPTCS.297.21.

[28] E. Lorini (2020): *Rethinking epistemic logic with belief bases*. *Artif. Intell.* 282, doi:10.1016/j.artint.2020.103233.

[29] E. Lorini & E. Rapion (2022): *Logical Theories of Collective Attitudes and the Belief Base Perspective*. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp. 833–841.

[30] E. Lorini & F. Schwarzentruber (2021): *A Computationally Grounded Logic of Graded Belief*. In: *Proceedings of the 17th European Conference on Logics in Artificial Intelligence (JELIA 2021)*, *LNCS* 12678, Springer, pp. 245–261, doi:10.1007/978-3-030-75775-5_17.

[31] J. J. Meyer & W. van der Hoek (1995): *Epistemic logic for AI and computer science*. Cambridge University Press, doi:10.1017/CBO9780511569852.

[32] F. Roelofsen (2007): *Distributed knowledge*. *Journal of Applied Non-Classical Logics* 17(2), pp. 255–273, doi:10.3166/jancl.17.255-273.

[33] Y. Shoham (2009): *Logical Theories of Intention and the Database Perspective*. *Journal of Philosophical Logic* 38(6), pp. 633–648, doi:10.1007/s10992-009-9116-8.

[34] 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.

[35] R. Stalnaker (2002): *Common Ground*. *Linguistics and Philosophy* 25(5-6), pp. 701–721, doi:10.1023/A:1020867916902.

[36] Y. N. Wáng & T. Ågotnes (2013): *Public announcement logic with distributed knowledge: expressivity, completeness and complexity*. *Synthese* 190(1), pp. 135–162, doi:10.1007/s11229-012-0243-3.

Modal Logic for Reasoning About Uncertainty and Confusion

Marta Bílková*
The Czech Academy of Sciences, Institute of Philosophy, Prague, Czech Republic
bilkova@cs.cas.cz

Thomas M. Ferguson
Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, USA
tferguson@gradcenter.cuny.edu

Daniil Kozhemiachenko
Aix Marseille Univ, CNRS, LIS, Marseille, France
daniil.kozhemiachenko@lis-lab.fr

We consider a modal logic that can formalise statements about uncertainty and beliefs such as *I think that my wallet is in the drawer rather than elsewhere* or *I am confused whether my appointment is on Monday or Tuesday*. To do that, we expand Gödel modal logic $\mathbf{K}\mathsf{G}$ with the involutive negation $\sim$ defined as $v(\sim\phi,w) = 1 - v(\phi,w)$. We provide semantics with the finite model property for our new logic that we call $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ and show its equivalence to the standard semantics over $[0,1]$-valued Kripke models. Namely, we show that ϕ is valid in the standard semantics of $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ iff it is valid in the new semantics. Using this new semantics, we construct a constraint tableaux calculus for $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ that allows for an explicit extraction of countermodels from complete open branches and then employ the tableaux calculus to obtain the PSPACE-completeness of the validity in $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$.

1 Introduction

When people evaluate their uncertainty or beliefs, they can do it in two ways: either quantitatively, by assigning a specific number to the value of the belief, e.g., *'the chance of rain today is 73%'* or qualitatively, by comparing degrees of two beliefs without explicitly mentioning their numerical values as in *'I think that my wallet is more likely to be in the drawer than in the bag'*. Usually, a person does not assign exact values to their beliefs and thus reasons qualitatively. Qualitative reasoning can be formalised, e.g., in Gödel modal logic $\mathbf{K}\mathsf{G}$ or, if one interprets 'rather' as 'strictly more likely', its expansion with $\triangle$ (Baaz' Delta [3]) or $\prec$ (co-implication [25, 21]) $\mathbf{K}\mathsf{biG}$ (modal bi-Gödel logic).

Comparing degrees of beliefs is possible in $\mathbf{K}\mathsf{biG}$ because the semantics of its propositional fragment — bi-Gödel logic (biG) — is characterised by the Gödel t-norm $\wedge_{\mathsf{G}}$ and its residuum $\rightarrow_{\mathsf{G}}$ and their duals $\vee_{\mathsf{G}}$ and $\prec_{\mathsf{G}}$ defined on $[0,1]$ by:

$$a \wedge_{\mathsf{G}} b = \min(a,b) \quad a \rightarrow_{\mathsf{G}} b = \begin{cases} 1 & \text{if } a \leq b \\ b & \text{otherwise} \end{cases} \quad a \vee_{\mathsf{G}} b = \max(a,b) \quad a \prec_{\mathsf{G}} b = \begin{cases} 0 & \text{if } a \leq b \\ a & \text{otherwise} \end{cases} \tag{1}$$

Propositional Gödel logic is a prominent many-valued logic whose semantics defined on $[0,1]$ is suitable for formalising relative comparisons of degrees of truth. Gödel *modal* logics expand propositional Gödel logic with modalities $\Box$ and $\Diamond$ and are usually interpreted on $[0,1]$-valued Kripke frames. There,

*The first author is supported by the project Logical Structure of Information Channels, no. 21-23610M of the Czech Science Foundation. We are also grateful to the reviewers for their suggestions that helped to improve the paper.

A. Bjorndahl (Ed.): Theoretical Aspects of Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 216–232, doi:10.4204/EPTCS.437.20

© M. Bílková, T. M. Ferguson, D. Kozhemiachenko
This work is licensed under the Creative Commons Attribution License.

propositional variables in states are evaluated with numbers from $[0,1]$, the propositional connectives correspond to Gödel semantics shown in (1), and the values of $\Box\phi$ and $\Diamond\phi$ are defined as, respectively, the infimum and supremum of the values of ϕ in the accessible states.

Note that it is natural not only to compare degrees of certainty in two different events but also to assert whether the statement is *more likely to be true* or *more likely to be false*. This can be expressed by statements such as *'my wallet is likely in the drawer'*, meaning *'I think that my wallet is in the drawer rather than elsewhere'*. I.e., the (subjective) likelihood of the wallet being in the drawer is greater than $\frac{1}{2}$. Furthermore, an agent may be *confused* about a statement p, i.e., not be able to discern p with its negation.

As biG can express only 0 and 1, the first kind of statements cannot be formalised in **K**biG. However, $\frac{1}{2}$ is expressible with the involutive negation $\sim$ defined as $v(\sim\phi,w) = 1 - v(\phi,w)$: $v(p \leftrightarrow \sim p,w) = 1$ iff $v(p,w) = \frac{1}{2}$. Moreover, since $\neg p$ — the Gödelian negation of p — is defined as $p \to_{\mathsf{G}} \mathbf{0}$, p and $\neg p$ can never have the same value, and thus, can always be distinguished. I.e., the second kind of statements is not formalisable either.

In this paper, we will present a modal logic that allows us to express such statements. Let us now provide a broader context of our work.

Gödel logics with involution The modal logic that we consider in our paper expands $\mathsf{G}_{\mathsf{inv}}$ — the propositional Gödel logic with involutive negation $\sim$ that was introduced in [20]. Adding $\sim$ greatly enhances the expressivity of the Gödel logic. Not only can $\frac{1}{2}$ be expressed, but also coimplication (pseudo-difference, interpreted as ϕ excludes χ) and the Baaz Delta (the 1-detecting operator) can be defined.

$$\phi \prec \chi \coloneqq \sim(\sim\chi \to \sim\phi) \qquad\qquad \triangle\phi \coloneqq \mathbf{1} \prec (\mathbf{1} \prec \phi)$$

We observe briefly that $\sim$ is closer than $\neg$ to the intuitive reading of 'not' in contexts involving truth degrees. Indeed, if the truth degree of p ('Peter is tall') is $\frac{1}{2}$, it is reasonable to assume that the truth degree of $\sim p$ (Peter is not tall) is also $\frac{1}{2}$. Similarly, if Peter is shorter than average (i.e., the truth degree of p is smaller than $\frac{1}{2}$), one may safely accept that the truth degree of 'not-p' should be greater than $\frac{1}{2}$. On the other hand, the truth degree of $\neg p$ would be 0 whenever the truth degree of p is positive. Thus, using $\neg p$ to stand for 'Peter is not tall' is counterintuitive as it says that 'Peter is tall is contradictory'.

Note, moreover, that $\mathsf{G}_{\mathsf{inv}}$ is *paraconsistent* in the following sense: (1) $(p \wedge \sim p) \to q$ is not valid; (2) $p, \sim p \not\models_{\mathsf{G}_{\mathsf{inv}}} q$ if entailment is defined via the preservation of order on $[0,1]$. These properties were investigated in [19, 18].

Paraconsistency of the propositional fragment is a useful property when it comes to the analysis of beliefs. Indeed, from a classical standpoint, *all contradictions are equivalent* (because they are all universally false). Hence, $\Box(p \wedge \sim p) \leftrightarrow \Box(q \wedge \sim q)$ is valid in every classical modal logic. If one interprets $\Box$ as a belief modality, this means that the agent cannot differentiate between contradictions. This, however, is counterintuitive since a person can (implicitly) believe in one contradiction but not the other or reject one contradiction with more conviction than the other. In paraconsistent logics, however, contradictions are *not* equivalent, and thus, are not indistinguishable under $\Box$.

Gödel modal logics Modal expansions of Gödel logics and their applications have been extensively studied. The $\Box$ and $\Diamond$ fragments[1] of **K**G were axiomatised in [16]. Hypersequent calculi were constructed in [23, 24] and used to obtain the PSPACE-completeness of both fragments. **K**G in the bi-modal language

[1] Note that $\Box$ and $\Diamond$ are not interdefinable in Gödel modal logic.

was axiomatised in [17, 27] (for fuzzy and crisp[2] frames, respectively). Moreover, it was shown in [14, 15] that crisp and fuzzy **K**G are PSPACE-complete.

The applications of Gödel modal logics and its expansions are well researched. In [26], the completeness of **K45** and **KD45** extensions of **K**G with respect to non-normalised and normalised possibilistic frames was established. In [1, 2], a temporal logic expanding the Gödel logic with co-implication was proposed. In addition, paraconsistent expansions of **K**G that can be used to reason about contradictory beliefs were proposed and studied in [4, 5, 6].

Gödel description logics Gödel description logics (DLs) were proposed in [8] to represent graded information in the ontologies and were further studied in [9]. Just as the classical description logics are notational variants of the classical logics with the global (universal) modality, Gödel DLs can be considered a notational variant of global Gödel modal logics. Moreover, Gödel DLs are usually equipped with the involutive negation. They differ from Łukasiewicz and Product fuzzy DLs because they are decidable and often have the same complexity as their classical counterparts even in the cases of expressive logics [12, 11, 10, 13]. Note, however, that since infinite-valued Gödel DLs lack the finite model property, their decision procedures do not produce explicit (counter-)models.

Contributions and plan of the paper In this paper, we consider $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ — an expansion of **K**G with involutive negation and construct a tableaux calculus that allows for the extraction of finite countermodels from failed proofs and can be used to establish its PSPACE-completeness. Our contribution is two-fold. First, following [14, 15], we present an alternative semantics for $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ that possesses the finite model property. In particular, we show that the sets of validities with respect to the standard and the alternative semantics coincide (both in the case of crisp and arbitrary frames). Second, using this semantics, we construct a tableaux calculus by combining the techniques from [22] and [28] that allows us to read off countermodels from complete open branches. We then use this calculus to establish PSPACE-completeness of $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$.

The remainder of the text is structured as follows. In Section 2, we present $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ and its standard semantics. We also discuss the formalisation of reasoning about uncertainty and confusion in $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$. Section 3 is dedicated to the alternative semantics with the finite model property. In Section 4, we construct a tableaux calculus for $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ and show its soundness and completeness. Then, in Section 5, we use the tableaux to obtain the PSPACE-completeness of $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$. Finally, we summarise our results and provide a plan for future work in Section 6.

2 Language and standard semantics

Let us present the language and standard semantics of $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$. We fix a countable set $\mathtt{Prop}$ of propositional variables and define the language $\mathscr{L}^{\sim}_{\Box,\Diamond}$ using the grammar below.

$$\mathscr{L}^{\sim}_{\Box,\Diamond} \ni \phi ::= p \in \mathtt{Prop} \mid {\sim}\phi \mid (\phi \wedge \phi) \mid (\phi \to \phi) \mid \Box\phi \mid \Diamond\phi$$

The next definitions introduce the notions of frames and $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ models.

Definition 1 (Frames)**.**

- A *fuzzy frame* is a tuple $\mathfrak{F} = \langle W, R\rangle$ with $W \neq \varnothing$ and $R : W \times W \to [0,1]$.

[2] A frame $\langle W, R\rangle$ is called *fuzzy* if $R : W \times W \to [0,1]$ and *crisp* if $R : W \times W \to \{0,1\}$.

- A *crisp frame* is a tuple $\mathfrak{F} = \langle W, R\rangle$ with $W \neq \varnothing$ and $R : W \times W \to \{0,1\}$.

For a frame $\mathfrak{F} = \langle W, R\rangle$ and $w \in W$, we set $R(w) = \{w' : wRw' > 0\}$.

Definition 2 (Semantics of $\mathbf{KG}_{\mathsf{inv}}$)**.** A $\mathbf{KG}_{\mathsf{inv}}$*-model* is a tuple $\mathfrak{M} = \langle W, R, v\rangle$ with $\langle W, R\rangle$ being a frame and $v : \mathtt{Prop} \times W \to [0,1]$ (a $\mathbf{KG}_{\mathsf{inv}}$-valuation) extended to the complex formulas as follows (cf. (1) for the definition of $\wedge_{\mathsf{G}}$ and $\to_{\mathsf{G}}$).

$$v(\sim\phi, w) = 1 - v(\phi, w) \qquad v(\phi \wedge \chi, w) = v(\phi, w) \wedge_{\mathsf{G}} v(\chi, w)$$
$$v(\phi \to \chi, w) = v(\phi, w) \to_{\mathsf{G}} v(\chi, w) \qquad v(\Box\phi, w) = \inf_{w' \in W} \{wRw' \to_{\mathsf{G}} v(\phi, w')\}$$
$$v(\Diamond\phi, w) = \sup_{w' \in W} \{wRw' \wedge_{\mathsf{G}} v(\phi, w')\}$$

We say that $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ is $\mathbf{KG}_{\mathsf{inv}}$*-valid on a pointed frame* $\langle \mathfrak{F}, w\rangle$ ($\mathfrak{F}, w \models_{\mathbf{KG}_{\mathsf{inv}}} \phi$) iff $v(\phi, w) = 1$ for any model $\mathfrak{M}$ on $\mathfrak{F}$. ϕ is $\mathbf{KG}_{\mathsf{inv}}$*-valid on frame* $\mathfrak{F}$ ($\mathfrak{F} \models_{\mathbf{KG}_{\mathsf{inv}}} \phi$) iff $\mathfrak{F}, w \models_{\mathbf{KG}_{\mathsf{inv}}} \phi$ for any $w \in \mathfrak{F}$. Finally, ϕ is $\mathbf{KG}_{\mathsf{inv}}$*-valid* (or simply *valid*) iff $\mathfrak{F} \models_{\mathbf{KG}_{\mathsf{inv}}} \phi$ for every $\mathfrak{F}$.

Convention 1. Given a formula ϕ, we use $\ell(\phi)$ to denote the number of occurrences of symbols in ϕ. We will write $\phi \leftrightarrow \chi$ as a shorthand for $(\phi \to \chi) \wedge (\chi \to \phi)$ and use the following defined connectives:

$$\mathbf{1} := p \to p \qquad \mathbf{0} := \sim\mathbf{1} \qquad \neg\phi := \phi \to \mathbf{0}$$
$$\phi \vee \chi := \sim(\sim\phi \wedge \sim\chi) \qquad \phi \prec \chi := \sim(\sim\chi \to \sim\phi) \qquad \triangle\phi := \mathbf{1} \prec (\mathbf{1} \prec \phi)$$

One can see from (1) and Definition 2 that $v(\phi \leftrightarrow \chi, w) = 1$ if $v(\phi, w) = v(\chi, w)$ and $\min(v(\phi, w), v(\chi, w))$, otherwise; $v(\phi \vee \chi, w) = v(\phi, w) \vee_{\mathsf{G}} v(\chi, w)$; $v(\phi \prec \chi, w) = v(\phi, w) \prec_{\mathsf{G}} v(\chi, w)$; and $v(\triangle\phi, w) = 1$ if $v(\phi, w) = 1$ and $v(\triangle\phi, w) = 0$, otherwise.

Let us now look at how we can use $\mathbf{KG}_{\mathsf{inv}}$ to formalise the contexts we considered in the previous section. First, we consider the assertion that ϕ is more likely to be true than false.

Example 1. Consider the following statement from the Introduction.

wal.**:** *I think that my wallet is in the drawer rather than elsewhere.*

We write d for 'my wallet is in the drawer' and use $\Box$ for 'I think that'. Now, we need to formalise that the degree of subjective likelihood of d is greater than $\frac{1}{2}$. That is, we should write a formula that is true at w when $v(\Box d, w) > \frac{1}{2}$. We formalise wal. as follows: $\phi_{\mathsf{wal.}} := \sim\triangle(\Box d \to \sim\Box d)$. By Definition 2, and Convention 1, one can see that $v(\phi_{\mathsf{wal.}}, w) = 1$ iff $v(\Box d \to \sim\Box d, w) < 1$. This means that $v(\Box d, w) > v(\sim\Box d, w)$, i.e., $v(\Box d, w) > 1 - v(\Box d, w)$, whence, $v(\Box d, w) > \frac{1}{2}$, as required.

Let us now see how to formalise situations involving confusion using $\mathbf{KG}_{\mathsf{inv}}$.

Example 2. Assume that Ann wants to come to Brittney's birthday party. Unfortunately, the only piece of information Ann has about the party is that it is held this weekend and lasts one evening. Thus, Ann cannot discern between the birthday party happening on Saturday (s) or Sunday ($\sim s$).

Formally, this means that in every accessible state, the values of s and $\sim s$ must be the same. Thus, we write the following formula: $\phi_{\mathsf{bd}} := \Box\triangle(s \leftrightarrow \sim s)$. It is easy to verify that $v(\phi_{\mathsf{bd}}, w) = 1$ iff $v(s, w') = \frac{1}{2} = v(\sim s, w')$ in every $w' \in R(w)$. Indeed, the value of $\triangle(s \leftrightarrow \sim s)$ always belongs to $\{0,1\}$. Thus, $v(\phi_{\mathsf{bd}}, w) = 1$ iff $v(\triangle(s \leftrightarrow \sim s), w') = 1$ in every accessible state. That is, $v(s \leftrightarrow \sim s, w') = 1$, and hence, $v(s, w) = v(\sim s, w) = 1 - v(s, w)$.

We finish the section with the following observations. First, $\Box p \leftrightarrow \sim\Diamond\sim p$ is valid on $\mathfrak{F}$ iff $\mathfrak{F}$ is crisp. Second, $\Box$ and $\Diamond$ are not interdefinable on the class of all frames.

Proposition 1. $\mathfrak{F} \models_{\mathbf{KG}_{\mathsf{inv}}} \Box p \leftrightarrow \sim\Diamond\sim p$ *iff* $\mathfrak{F}$ *is crisp.*

Proof. Observe that on crisp frames, the semantics of modalities can be simplified as follows: $v(\Box\phi,w) = \inf\{v(\phi,w') : wRw'\}$ and $v(\Diamond\phi,w) = \sup\{v(\phi,w') : wRw'\}$. From here, it is evident that $\Box p \leftrightarrow \sim\Diamond\sim p$ is valid on crisp frames. For the converse, assume that $\mathfrak{F}$ is not crisp, i.e., there are $w,w' \in \mathfrak{F}$ s.t. $wRw' = x$ for some $0 < x < 1$. Now for every $w'' \in W$, set $v(p,w'') = wRw''$ and $v(p,w'') = 1$ only if $wRw'' = 1$. It is clear that $v(\Box p,w) = 1$. On the other hand, we have that $v(\Diamond\sim p,w) > 0$ since $0 < v(\sim p,w') \leq 1-x < 1$ and $wRw' = x$. But then, $v(\sim\Diamond\sim p,w) \neq 1$, as required. □

Proposition 2. *$\Box$ and $\Diamond$ are not interdefinable on the class of all frames. That is, there is a model $\mathfrak{M}$ and $w \in \mathfrak{M}$ s.t. (1) there is no $\Diamond$-free formula χ s.t. $v(\chi,w) = v(\Diamond p,w)$; (2) there is no $\Box$-free formula ψ s.t. $v(\psi,w) = v(\Box p,w)$.*

Proof. Consider Fig. 1 and note that $v(\Box p,w) = \frac{1}{5}$ and $v(\Diamond p,w) = \frac{1}{4}$. We show by induction on τ that

$$\forall \tau \in \mathscr{L}^{\sim}_{\Box,\Diamond} : v(\tau,w') \in \{0, 1/5, 4/5, 1\} \text{ and } v(\tau,w'') \in \{0, 1/4, 3/4, 1\} \tag{2}$$

The basis cases in (2) hold by the construction of the model and the cases of propositional connectives are obtained by applying the induction hypothesis. Finally, as $R(w') = R(w'') = \varnothing$, for $\tau = \Diamond\sigma$, we have $v(\Diamond\sigma,w') = v(\Diamond\sigma,w'') = 0$; and for $\tau = \Box\sigma$, we have that $v(\Box\sigma,w') = v(\Box\sigma,w'') = 1$.

Then, we can show for every $\tau \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ that

$$\begin{aligned} v(\tau,w'') = 0 &\Rightarrow v(\tau,w') = 0 & v(\tau,w'') = 1/4 &\Rightarrow v(\tau,w') = 1/5 \\ v(\tau,w'') = 3/4 &\Rightarrow v(\tau,w') = 4/5 & v(\tau,w'') = 1 &\Rightarrow v(\tau,w') = 1 \end{aligned} \tag{3}$$

and

$$\begin{aligned} v(\tau,w') = 0 &\Rightarrow v(\tau,w'') = 0 & v(\tau,w') = 1/5 &\Rightarrow v(\tau,w'') = 1/4 \\ v(\tau,w') = 4/5 &\Rightarrow v(\tau,w'') = 3/4 & v(\tau,w') = 1 &\Rightarrow v(\tau,w'') = 1 \end{aligned} \tag{4}$$

For (3), we proceed by induction. The basis case of $\tau = p$ holds by the construction of the model. If $\tau = \Diamond\sigma$, then $v(\Diamond\sigma,w') = v(\Diamond\sigma,w'') = 0$. If $\tau = \Box\sigma$, then $v(\Box\sigma,w') = v(\Box\sigma,w'') = 1$.

Let $\tau = \rho \wedge \sigma$. The cases of $v(\tau,w'') \in \{0,1\}$ are straightforward. Assume that $v(\rho\wedge\sigma,w'') = \frac{1}{4}$, then w.l.o.g. $v(\rho,w'') = \frac{1}{4}$ and $v(\sigma,w'') \in \{\frac{1}{4},\frac{3}{4},1\}$. By the induction hypothesis, we have that $v(\rho,w') = \frac{1}{5}$ and $v(\sigma,w'') \in \{\frac{1}{5},\frac{4}{5},1\}$, whence, $v(\rho\wedge\sigma,w') = \frac{1}{5}$. The case when $v(\rho\wedge\sigma,w'') = \frac{3}{4}$ is tackled in the same way.

Consider $\tau = \rho \to \sigma$. We deal only with the most instructive case of $v(\rho\to\sigma,w'') = \frac{1}{4}$. Here, we have $v(\sigma,w'') = \frac{1}{4}$ and $v(\rho,w'') \in \{1,\frac{3}{4}\}$. By the induction hypothesis, we have $v(\sigma,w') = \frac{1}{5}$ and $v(\rho,w') \in \{\frac{4}{5},1\}$. Hence, $v(\rho\to\sigma,w') = \frac{1}{5}$.

Other propositional connectives can be dealt with similarly.

For (4), the case of $\tau = p$ holds by construction; if $\tau = \Box\sigma$, then $v(\Box\sigma,w') = v(\Box\sigma,w'') = 1$; if $\tau = \Diamond\sigma$, then $v(\Diamond\sigma,w') = v(\Diamond\sigma,w'') = 0$. The cases of propositional connectives can be dealt with in the same way as for (3).

We now show by induction on formulas χ and ψ that (i) there is no $\Box$-free formula χ s.t. $v(\chi,w) \in \{\frac{1}{5},\frac{4}{5}\}$ and (ii) there is no $\Diamond$-free formula ψ s.t. $v(\psi,w) \in \{\frac{1}{4},\frac{3}{4}\}$.

For (i), the basis case holds by the construction of the model. The cases of propositional connectives can be established by a straightforward application of the induction hypothesis. Now let $\chi = \Diamond\tau$. It is clear that $v(\Diamond\tau,w) \neq \frac{4}{5}$ since $wRw' = wRw'' = \frac{2}{3}$. We check that $v(\Diamond\tau,w) \neq \frac{1}{5}$. Assume for contradiction

$$w'' : p = 1/4 \xleftarrow{2/3} w : p = 0 \xrightarrow{2/3} w' : p = 1/5$$

Figure 1: All variables have the same values as p.

that $v(\Diamond\tau, w) = \frac{1}{5}$. Then, $v(\tau, w') = \frac{1}{5}$ from (2). But using (3), we have that $v(\tau, w'') = \frac{1}{4}$, whence $v(\Diamond\tau, w) = \frac{1}{4}$. Contradiction.

For (ii), the basis case holds by the construction of the model and the cases of propositional connectives can be proven by the application of the induction hypothesis. We show that $v(\Box\tau, w) \notin \{\frac{1}{4}, \frac{3}{4}\}$. Observe that $wRw' = wRw'' = \frac{2}{3}$, whence, $v(\Box\tau, w) \neq \frac{3}{4}$. We are going to check that $v(\Box\tau, w) \neq \frac{1}{4}$. The proof is similar to the one of (i) but we use (4). If $v(\Box\tau, w) = \frac{1}{4}$, then $v(\tau, w'') = \frac{1}{4}$ by (2). But from here, we obtain $v(\tau, w') = \frac{1}{5}$ by (4). Contradiction. □

3 Semantics for $\mathbf{KG}_{\mathsf{inv}}$ with the finite model property

It is well-known (cf., e.g., [16]) that the standard semantics of $\mathbf{KG}$ (modal Gödel logic) and thus, of $\mathbf{KG}_{\mathsf{inv}}$[3], lack the finite model property (FMP). Still, semantics over so-called F-models (F stands for 'finitary') satisfying the finite model property can be provided to any expansion of $\mathbf{KG}$ with order-based connectives (i.e., connectives expressible via infima and suprema of lattice operations) [15].

One can see, however, that $\sim$ is not an order-based connective. Thus, we cannot use the general result of [15] to provide a semantics over F-models. Still, we can adapt the approach from [14]. Namely, we will redefine the models in such a way that in a given state w, there is only a finite set of values that formulas of the form $\Box\phi$ or $\Diamond\phi$ are allowed to have. Then modal formulas will be 'witnessed' in w if there is an accessible state w' where the value of ϕ is sufficiently close to the value of $\Box\phi$ or $\Diamond\phi$.

Definition 3 (F-models)**.** An F-*model* is a tuple $\mathfrak{M} = \langle W, R, T, v\rangle$ with $\langle W, R\rangle$ being a frame and $T : W \to \mathscr{P}_{<\omega}([0,1])$[4] be s.t. for all $w \in W$, it holds that: $\{0, \frac{1}{2}, 1\} \subseteq T(w)$, and if $x \in T(w)$, then $1 - x \in T(w)$. Finally, v (*valuation*) is a map $v : \mathtt{Prop} \times W \to [0,1]$ that is extended to the complex formulas as in $\mathbf{KG}_{\mathsf{inv}}$ (Definition 2) in the cases of propositional connectives and in the modal cases, as follows:

$$v(\Box\phi, w) = \max\{x \in T(w) \mid x \leq \inf_{w' \in W}\{wRw' \to_{\mathsf{G}} v(\phi, w')\}\}$$
$$v(\Diamond\phi, w) = \min\{x \in T(w) \mid x \geq \sup_{w' \in W}\{wRw' \wedge_{\mathsf{G}} v(\phi, w')\}\}$$

We say that $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ is F-*valid* iff $v(\phi, w) = 1$ in every F-model $\mathfrak{M}$ and every $w \in \mathfrak{M}$.

In the definition above, T assigns a finite subset of $[0,1]$ to each state in the model. This way, $T(w)$ is the set of values that modal formulas can have at w. Note also that the value of $\Box\phi$ in an F-model can be interpreted as an approximation from below of its 'real' value in a standard model. Dually, the value of $\Diamond\phi$ is an approximation from above. It is important to observe that we need the $1 - x$ closure of $T(w)$ and the presence of $\frac{1}{2}$. Otherwise, we could have the following 'models'.

$$\mathfrak{N}: \ w_0 \xrightarrow{R=1/2} w_1 : p = \tfrac{1}{2} \quad T(w_0) = \{0,1\} \quad \|\| \quad \mathfrak{N}': \ w'_0 \xrightarrow{R=1} w'_1 : p = \tfrac{1}{3} \quad T(w'_0) = \{0, 1/3, 1/2, 1\}$$

[3] As Gödel negation is expressible in $\mathscr{L}^{\sim}_{\Box,\Diamond}$ (Convention 1), $\mathbf{KG}_{\mathsf{inv}}$ is a conservative extension of $\mathbf{KG}$ with $\sim$.

[4] We use $\mathscr{P}_{<\omega}([0,1])$ to denote the set of all *finite* subsets of $[0,1]$.

One can see that $v(\Box p \wedge \Diamond{\sim}p, w_0) = 1$ in $\mathfrak{N}$ which is impossible in $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-models. Furthermore, $\mathfrak{N}'$ is crisp but $v'(\Box p \leftrightarrow {\sim}\Diamond{\sim}p, w_0') = 0$. By Proposition 1, this is impossible crisp in $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-models.

Let us now state the main result of the section.

Theorem 1.

1. *ϕ is $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-valid iff it is* F*-valid on all* finite models*.*
2. *ϕ is $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-valid on* crisp frames *iff it is* F*-valid on all* finite crisp models*.*

We adapt the proof of [14, Theorem 1]. We begin with some technical notions.

Definition 4 (Generated models)**.**

- A frame $\widehat{\mathfrak{F}} = \langle \widehat{W}, \widehat{R} \rangle$ is *a subframe* of $\mathfrak{F} = \langle W, R \rangle$ if $\widehat{W} \subseteq W$ and $\widehat{R}$ is the restriction of R on $\widehat{W}$.
- $\widehat{\mathfrak{M}} = \langle \widehat{W}, \widehat{R}, \widehat{v} \rangle$ is *a submodel of* $\mathfrak{M} = \langle W, R, v \rangle$ if $\langle \widehat{W}, \widehat{R} \rangle$ is a subframe of $\langle W, R \rangle$ and $\widehat{v}$ is the restriction of v on $\widehat{W}$.
- $\widehat{\mathfrak{M}} = \langle \widehat{W}, \widehat{R}, \widehat{T}, \widehat{v} \rangle$ is *a submodel of* $\mathfrak{M} = \langle W, R, T, v \rangle$ if $\langle \widehat{W}, \widehat{R} \rangle$ is a subframe of $\langle W, R \rangle$ and $\widehat{v}$ and $\widehat{T}$ are the restrictions of v and T on $\widehat{W}$.
- $\widehat{\mathfrak{M}}$ is *generated by* $X \subseteq W$ if it is the smallest submodel containing X s.t. if $w \in \widehat{W}$ and $wRw' > 0$, then $w' \in \widehat{W}$, as well.

Convention 2*.*

- A model $\mathfrak{M} = \langle W, R, v \rangle$ (or $\mathfrak{M} = \langle W, R, T, v \rangle$) is *tree-like* if $\langle W, R \rangle$ is a directed rooted tree.
- Given a tree-like model $\mathfrak{M}$, $\hbar(\mathfrak{M})$ denotes its *height*.

First, we can establish the analogue of [14, Lemma 1]. We show that to verify the validity of a formula (with respect to $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-models and F-models), it suffices to consider only tree-like models of *finite height*. The proof is standard, so we omit it here.

Lemma 1. *Let $\mathfrak{M} = \langle W, R, v \rangle$ and $\widehat{\mathfrak{M}} = \langle \widehat{W}, \widehat{R}, \widehat{v} \rangle$ be $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-models. Let further $\mathfrak{M}^{\mathsf{F}} = \langle W^{\mathsf{F}}, R^{\mathsf{F}}, T, v^{\mathsf{F}} \rangle$ and $\widehat{\mathfrak{M}^{\mathsf{F}}} = \langle \widehat{W^{\mathsf{F}}}, \widehat{R^{\mathsf{F}}}, \widehat{T}, \widehat{v^{\mathsf{F}}} \rangle$ be F-models. Then the following statements hold.*

(a) If $\widehat{\mathfrak{M}}$ (resp., $\widehat{\mathfrak{M}^{\mathsf{F}}}$) is a generated submodel of $\mathfrak{M}$ ($\mathfrak{M}^{\mathsf{F}}$), then $\widehat{v}(\phi, w) = v(\phi, w)$ for all $w \in \widehat{W}$ (resp., $\widehat{v^{\mathsf{F}}}(\phi, w) = v^{\mathsf{F}}(\phi, w)$ for all $w \in \widehat{W^{\mathsf{F}}}$) and $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$.

(b) For all $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ and $w \in \mathfrak{M}$ ($w \in \mathfrak{M}^{\mathsf{F}}$), there is a tree-like submodel $\widehat{\mathfrak{M}}$ of $\mathfrak{M}$ ($\widehat{\mathfrak{M}^{\mathsf{F}}}$ of $\mathfrak{M}^{\mathsf{F}}$) generated by w s.t. $v(\phi, w) = \widehat{v}(\phi, w)$ ($v^{\mathsf{F}}(\phi, w) = \widehat{v^{\mathsf{F}}}(\phi, w)$) and $\hbar(\widehat{\mathfrak{M}}) \leq \ell(\phi)$.

(c) Let $g : [0,1] \to [0,1]$ be s.t. $g(0) = 0$, $g(1) = 1$, $\underline{g(1-x) = 1 - g(x)}$, and $x \leq x'$ iff $g(x) \leq g(x')$[5]. Let further $\mathfrak{N} = \langle W_{\mathfrak{N}}, R_{\mathfrak{N}}, v_{\mathfrak{N}} \rangle$ (resp., $\mathfrak{N}^{\mathsf{F}} = \langle W^{\mathsf{F}}_{\mathfrak{N}}, R^{\mathsf{F}}_{\mathfrak{N}}, T_{\mathfrak{N}}, v^{\mathsf{F}}_{\mathfrak{N}} \rangle$) be s.t. $W_{\mathfrak{N}} = W$, $wR_{\mathfrak{N}}w' = g(wRw')$, and $v_{\mathfrak{N}}(p, w) = g(v(p, w))$ ($W^{\mathsf{F}}_{\mathfrak{N}} = W^{\mathsf{F}}$, $wR^{\mathsf{F}}_{\mathfrak{N}}w' = g(wR^{\mathsf{F}}w')$, $T_{\mathfrak{N}}(w) = g(T(w))$, $v^{\mathsf{F}}_{\mathfrak{N}}(p, w) = g(v^{\mathsf{F}}(p, w))$) for every $w, w' \in W$ ($w, w' \in W^{\mathsf{F}}$) and $p \in \mathtt{Prop}$. Then it holds that $v_{\mathfrak{N}}(\phi, w) = g(v(\phi, w))$ (resp., $v^{\mathsf{F}}_{\mathfrak{N}}(\phi, w) = g(v^{\mathsf{F}}(\phi, w))$) for every $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ and $w \in W$ ($w \in W^{\mathsf{F}}$).

We remark quickly that because of ${\sim}$, we need the underlined part of (c) that was not present in the original statement. Otherwise, we could have $g(\frac{1}{2}) = \frac{2}{3}$ which would fail (c) since $v_{\mathfrak{M}}(p \leftrightarrow {\sim}p, w) = 1$ if $v_{\mathfrak{M}}(p, w) = \frac{1}{2}$ but $v_{\mathfrak{N}}(p \leftrightarrow {\sim}p, w) = \frac{1}{3}$ since $v_{\mathfrak{N}}(p, w) = \frac{2}{3}$.

The next statements are analogous to [14, Lemmas 2 and 3].

Lemma 2. *For any tree-like F-model $\mathfrak{M} = \langle W, R, T, v \rangle$ of finite height with root x_0 there is a tree-like $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-model $\widehat{\mathfrak{M}} = \langle \widehat{W}, \widehat{R}, \hat{v} \rangle$ with root $\hat{x}_0$ s.t. $v(\phi, x_0) = \hat{v}(\phi, \hat{x}_0)$ for each $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$. In addition, if $\mathfrak{M}$ is crisp, so is $\widehat{\mathfrak{M}}$.*

[5] Observe that this, in particular, means that $x > x'$ implies $g(x) > g(x')$.

Proof. The statement may be proven by induction on $\hbar(\mathfrak{M})$. The case in which $\hbar(\mathfrak{M}) = 0$ serves as basis, setting $\widehat{W} = W$, $\widehat{R} = \varnothing$, and $\widehat{v} = v$; in this case, the statement follows trivially. As induction step, suppose that $\hbar(\mathfrak{M}) = n+1$. By the induction hypothesis, for each $y \in R(x_0)$, the submodel of $\mathfrak{M}$ generated by $\{y\}$ — the model $\mathfrak{M}_y = \langle W_y, R_y, v_y\rangle$ — there is a $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ tree model $\widehat{\mathfrak{M}} = \langle \widehat{W_y}, \widehat{R}_y, \widehat{v_y}\rangle$ such that $\hat{v}_y(\phi, \hat{y}) = v_y(\phi, y) = v(\phi, y)$.

Take $T(x_0) = \{\alpha_0, \ldots, \alpha_m\}$ and for each $k \in \mathbb{N}$, pick an order embedding $h_k : [0,1] \to [0,1]$ such that $h_k(0) = 0$, $h_k(1) = 1$, and $h(1-j) = 1 - h(j)$ for all $j \in [0,1]$ such that:

$$h_k(\alpha_i) = \alpha_i \text{ for all } i \leq m \text{ and } k \in \mathbb{N}$$
$$h_k[(\alpha_i, \alpha_{i+1})] = (\alpha_i, \min(\alpha_i + {}^1\!/_k, \alpha_{i+1})) \text{ for all } i \leq m-1 \text{ and even } k \in \mathbb{N}$$
$$h_k[(\alpha_i, \alpha_{i+1})] = (\max(\alpha_i - {}^1\!/_k, \alpha_{i+1}), \alpha_{i+1}) \text{ for all } i \leq m-1 \text{ and odd } k \in \mathbb{N}$$

For each $y \in R(x_0)$ and $k \in \mathbb{N}$, there exists an F-model $\widehat{\mathfrak{M}}_y^k$ defined so that $\widehat{W}_y^k$ includes for each $\hat{x}_y \in \widehat{W}_y$ a point $\hat{x}_y^k$ with root $\hat{y}^k$ and for all $\hat{x}_y$, $\hat{z}_y$, and $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$, $\hat{x}_y^k \widehat{R}_y^k \hat{z}_y^k = h_k(\hat{x}_y \widehat{R}_y \hat{z}_y)$ and $\widehat{V}_y^k(\psi, \hat{x}_y^k) = h_k(\widehat{v}_y(\psi, \hat{x}_y))$. We can define a $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-tree-model $\widehat{\mathfrak{M}}$ where

$$\widehat{W} = \bigcup_{y \in R(x_0)} \bigcup_{k \in \mathbb{N}} \widehat{W}_y^k \cup \{\hat{x}_0\}$$

$$x \widehat{R} z = \begin{cases} h_k(x_0 R y) & \text{if } x = \hat{x}_0 \text{ and } z = \hat{y}^k \text{ for some } y \in R(x_0) \text{ and } k \in \mathbb{N} \\ x \widehat{R}_y^k z & \text{if } x, z \in \widehat{W}_y^k \text{ for some } y \in R(x_0) \text{ and } k \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{v}(p, x) = \begin{cases} v(p, x_0) & \text{if } x = x_0 \\ \widehat{v}_y^k(p, x) & \text{if } x \in \widehat{W}_y^k \text{ for some } y \in R(x_0) \text{ and } k \in \mathbb{N} \end{cases}$$

The induction hypothesis ensures that for each $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ and $\hat{x}_y^k \in \widehat{W}$ distinct from $\hat{x}_0$, $\widehat{v}_y^k(\phi, \hat{x}_y^k) = \widehat{v}(\phi, \hat{x}_y^k)$. What remains is then to prove that $\widehat{v}(\phi, \hat{x}_0) = \widehat{v}(\phi, x_0)$.

This is proven by induction on $\ell(\phi)$. In case ϕ is an atom, this follows by construction of $\widehat{v}$. The non-modal connectives (including involutive negation) follow standardly. We consider the case of $\phi = \Box\psi$.

Suppose that $\phi = \Box\psi$. Then in case that $v(\Box\psi, x_0) = 1$, for all $y \in R(x_0)$, $x_0 R y \leq v(\psi, y)$, whence by Lemma 1, $v(\psi, y) = v_y(\psi, y) = \widehat{v}_y(\psi, \hat{y})$. As for any y $x_0 R y \leq \widehat{v}_y(\psi, \hat{y})$, it follows that for any k, $\hat{x}_0 \widehat{R} \hat{y}^k \leq \widehat{v}_y^k(\psi, \hat{y}^k)$, entailing that $\widehat{v}(\Box\psi, \hat{x}_0) = 1$ as well. Suppose then that $v(\Box\psi, x_0) = \alpha_i \neq 1$. Then for all $z \in W$, $x_0 R z \to_{\mathsf{G}} v(\psi, z) \geq \alpha_i$ and by construction of the order embeddings, we get the important feature that for all $z \in \widehat{W}$, $\hat{x}_0 \widehat{R} z \to_{\mathsf{G}} \hat{v}(\psi, z) \geq \alpha_i$. We consider two cases. In case there is a $y_0 \in W$ for which $x_0 R y \to_{\mathsf{G}} v(\psi, y) = \alpha_i$, then $x_0 R y_0 > v(\psi, y_0) = \alpha_i$ and the order embeddings ensure that for all $k \in \mathbb{N}$, $\hat{x}_0 \widehat{R} \hat{y}_0^k > \hat{v}(\psi, \hat{y}_0^k) = \alpha_i$. Otherwise, there must be some $y_0 \in W$ for which $x_0 R y_0 \to_{\mathsf{G}} v(\psi, y_0) \in (\alpha_i, \alpha_{i+1})$. By construction of the order embeddings, for any $\varepsilon > 0$ there is some $k \in \mathbb{N}$ for which $h_k(x_0 R y_0 \to_{\mathsf{G}} v(\psi, y_0)) = \hat{x}_0 \widehat{R} \hat{y}_0^k \to_{\mathsf{G}} \hat{v}(\psi, \hat{y}_0^k) \in (\alpha_i, \alpha_i + \varepsilon)$. In either case, $\hat{v}(\Box\psi, \hat{x}_0) = \alpha_i = v(\Box\psi, x_0)$.

Finally, we note that if $\mathfrak{M}$ is crisp, so is each $\widehat{\mathfrak{M}}_y$ and thus so is each $\widehat{\mathfrak{M}}_y^k$, whence $\widehat{\mathfrak{M}}$ is crisp. □

Definition 5 (Fragments)**.** A *fragment* is a set $\Sigma \subseteq \mathscr{L}^{\sim}_{\Box,\Diamond}$ s.t. it is closed under taking subformulas and $\{\mathbf{0}, \mathbf{1}\} \subseteq \Sigma$. Now let $\cdot^{\mathcal{J}} : \mathscr{L}^{\sim}_{\Box,\Diamond} \to \mathscr{L}^{\sim}_{\Box,\Diamond}$ be as follows:

$$\phi^{\mathcal{J}} = \begin{cases} {\sim}\phi \text{ if there is no } \xi \in \mathscr{L}^{\sim}_{\Box,\Diamond} \text{ such that } \phi = {\sim}\xi \\ \xi \text{ if there is a } \xi \in \mathscr{L}^{\sim}_{\Box,\Diamond} \text{ such that } \phi = {\sim}\xi \end{cases}$$

We say that $\Sigma \subseteq \mathscr{L}^{\sim}_{\Box,\Diamond}$ is an *involutively closed fragment (ICF)* if it is a fragment and for every $\phi,\chi \in \Sigma$, $\circ \in \{\wedge,\rightarrow\}$, and $\heartsuit \in \{\Box,\Diamond\}$, it holds that:

- if $\phi \in \Sigma$, then $\phi^{\mathscr{J}} \in \Sigma$;
- if $\phi \circ \chi \in \Sigma$, then $\{\phi^{\mathscr{J}} \circ \chi, \phi \circ \chi^{\mathscr{J}}, \phi^{\mathscr{J}} \circ \chi^{\mathscr{J}}\} \subseteq \Sigma$;
- if $\heartsuit\phi \in \Sigma$, then $\heartsuit\phi^{\mathscr{J}} \in \Sigma$.

Given an involutively closed fragment Σ, we define:

$$\Sigma_{\Box} = \{\Box\phi \mid \Box\phi \in \Sigma\} \qquad \Sigma_{\Diamond} = \{\Diamond\phi \mid \Diamond\phi \in \Sigma\} \qquad \Sigma_{\mathtt{Lit}} = \{l \mid l \in \Sigma \text{ and } l \text{ is a literal}\}$$
$$\Sigma^{\sim}_{\Box} = \{\sim\Box\phi \mid \sim\Box\phi \in \Sigma\} \qquad \Sigma^{\sim}_{\Diamond} = \{\sim\Diamond\phi \mid \sim\Diamond\phi \in \Sigma\}$$

Lemma 3. *Let Σ be a finite ICF and $\mathfrak{M} = \langle W,R,v\rangle$ a tree-like $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-model of finite height with root x_0. Then there is a tree-like F-model $\widehat{\mathfrak{M}} = \langle \widehat{W},\widehat{R},\widehat{T},\widehat{v}\rangle$ with root x_0 s.t. $\langle \widehat{W},\widehat{R}\rangle \subseteq \langle W,R\rangle$, $|W| \leq |\Sigma|^{\hbar(\mathfrak{M})}$, $|\widehat{T}(w')| \leq |\Sigma|$, and $v(\phi,x_0) = \widehat{v}(\phi,x_0)$ for every $\phi \in \Sigma$ and $w' \in \widehat{\mathfrak{M}}$. In addition, if $\mathfrak{M}$ is crisp, so is $\widehat{\mathfrak{M}}$.*

Proof. The statement may be proven by induction on $\hbar(\mathfrak{M})$. As the basis, letting $\widehat{W} = W$, $\widehat{R} = \varnothing$, $\widehat{v} = v$, and $\widehat{T}(x_0) = \{0,\frac{1}{2},1\}$ establishes the statement immediately.

As induction step, suppose that $\hbar(\mathfrak{M}) = n+1$. For each $y \in R(x_0)$, we note that the submodel $\mathfrak{M}_y$ generated by $\{y\}$ is a $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ model with $\hbar(\mathfrak{M}_y) \leq n$ and by induction hypothesis there exists an accompanying finite F-tree-model $\widehat{\mathfrak{M}}_y$ with root y such that for all $\phi \in \Sigma$, $\widehat{v}_y(\phi,y) = v_y(\phi,y) = v(\phi,y)$. Additionally, by induction hypothesis, $|\widehat{W}_y| \leq |\Sigma|^n$ and $|\widehat{T}_y(x)| \leq |\Sigma|+1$ for each $x \in \widehat{W}_y$. We build an appropriate finite F-model by selecting a set of appropriate $y \in R(x_0)$.

For a $\Delta \subseteq \mathscr{L}^{\sim}_{\Box,\Diamond}$ let $v_x[\Delta] = \{v(\phi,x) \mid \phi \in \Delta\}$; the set $v_{x_0}[\Sigma_{\Box} \cup \Sigma^{\sim}_{\Box} \cup \Sigma_{\Diamond} \cup \Sigma^{\sim}_{\Diamond}] \cup \{0,\frac{1}{2},1\}$ is a finite set $\{\alpha_1,\ldots,\alpha_m\}$ closed under $1-x$ where if $i<j$ then $\alpha_i < \alpha_j$. For each $\Box\phi \in \Sigma_{\Box}$ such that $v(\Box\phi,x_0) = \alpha_i < 1$ and $\alpha_j = 1-\alpha_i$, fix points $y_{\Box\phi}, y_{\sim\Box\phi} \in R(x_0)$ for which $x_0Ry_{\Box\phi} \rightarrow_{\mathsf{G}} v(\phi,y) < \alpha_{i+1}$ and $x_0Ry_{\sim\Box\phi} \rightarrow_{\mathsf{G}} v(\phi,y_{\sim}) < \alpha_{j+1}$. Similarly, for each $\Diamond\phi \in \Sigma_{\Diamond}$ such that $v(\Diamond\phi,x_0) = \alpha_i > 0$ and $\alpha_j = 1-\alpha_i$, fix points $y_{\Diamond\phi}, y_{\sim\Diamond\phi} \in R(x_0)$ such that $\min(x_0Ry_{\Diamond\phi}, v(\phi,y_{\Diamond\phi})) \geq \alpha_{i-1}$ and $\min(x_0Ry_{\sim\Diamond\phi}, v(\phi,y_{\sim\Diamond\phi})) \geq \alpha_{j-1}$. Then let $Y = \{y_\psi \mid \psi \in \Sigma_{\Box} \cup \Sigma^{\sim}_{\Box} \cup \Sigma_{\Diamond} \cup \Sigma^{\sim}_{\Diamond}\}$.

Now, define an F model $\widehat{\mathfrak{M}} = \langle \widehat{W},\widehat{R},\widehat{T},\widehat{v}\rangle$ such that

$$\widehat{W} = \bigcup_{y \in W} \widehat{W}_y \cup \{x_0\}$$

$$x\widehat{R}z = \begin{cases} x_0Rz & \text{if } x = x_0 \text{ and } z \in R(x_0) \\ x\widehat{R}_y z & \text{if } x,z \in \widehat{W}_y \text{ for some } y \in Y \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{T}(x) = \begin{cases} v_{x_0}[\Sigma_{\Box} \cup \Sigma^{\sim}_{\Box} \cup \Sigma_{\Diamond} \cup \Sigma^{\sim}_{\Diamond}] \cup \{0,\frac{1}{2},1\} & \text{if } x = x_0 \\ \widehat{T}_y(x) & \text{if } x \in \widehat{W}_y \text{ for some } y \in Y \end{cases}$$

$$\widehat{v}(p,x) = \begin{cases} v(p,x_0) & \text{if } x = x_0 \\ \widehat{v}_y(p,x) & \text{if } x \in \widehat{W}_y \text{ for some } y \in Y \end{cases}$$

Importantly, note that $\widehat{R}^+[x_0] = Y$ where $Y \subseteq R(x_0)$ and that for each $y \in Y$ and $\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$, $x_0\widehat{R}y = x_0Ry$ and $\hat{v}(\phi,y) = v(\phi,y)$. Now, we prove that for all ϕ, $\hat{v}(\phi,x_0) = v(\phi,x_0)$ by induction on length of ϕ. The atomic case follows by construction of $\hat{v}$ and the non-modal connectives (including involutive negation) are standard. We thus prove the case in which $\phi = \Box\psi$.

In case $v(\Box\psi,x_0)=1$, for each $y\in R^+[x_0]$, $x_0Ry \to_{\mathsf{G}} v(\psi,y)=1$, whence for each such y, $x_0Ry \leq v(\psi,y)$. As $\widehat{R}^+[x_0]\subseteq R(x_0)$, the induction hypothesis ensures that $x_0\widehat{R}y \to_{\mathsf{G}} \hat{v}(\psi,y)=1$ for all $y\in\widehat{R}^+[x_0]$ whence, as $1\in\widehat{T}(x_0)$, $\hat{v}(\Box\psi,x_0)=1$. In case $v(\Box\psi,x_0)=\alpha_i<1$, by construction, $\alpha_i\in\widehat{T}(x)$. Then for all $y\in\widehat{W}$, $\alpha_i\leq\inf\{x_0\widehat{R}y\to_{\mathsf{G}}\hat{v}(\psi,y)\mid y\in\widehat{W}\}$. But given the choice of $y_{\Box\psi}\in\widehat{W}$, $x_0\widehat{R}y_{\Box\psi}\to_{\mathsf{G}}\hat{v}(\psi,y_{\Box\psi})<\alpha_{i+1}$. Then by construction of $\widehat{T}$, $\hat{v}(\Box\psi,x_0)=\alpha_i$. The case of $\Diamond\psi$ is similar.

Note, now, that $|\widehat{W}|\leq|Y|\cdot|\sigma|^n\leq|\Sigma|^{n+1}=|\Sigma^{\mathscr{k}(\mathfrak{M})}|$ and that as $|\widehat{T}(x_0)|\leq|\Sigma_\Box\cup\Sigma^\sim_\Box\cup\Sigma_\Diamond\cup\Sigma^\sim_\Diamond|+3$, $|\widehat{T}(x_0)|\leq|\Sigma|+1$. Finally, as $\langle\widehat{W},\widehat{R}\rangle\subseteq\langle W,R\rangle$, the new model inherits crispness. □

Theorem 1 is now immediate from Lemmas 1, 2, and 3 as for any given ϕ there exists a finite ICF containing it.

We finish the section with a short remark. Note from the proof Lemma 3 that we could not have assumed a 'global' T (i.e., that $T(w)=T(w')$ for every $w,w'\in\mathfrak{M}$). Indeed, we need to form $T(w)$'s according to the values of modal formulas in the corresponding states of the standard model. These values, however, are independent and thus, cannot be simulated by one T.

4 Tableaux

In Section 3, we used the result of [14] to obtain that $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$ has the finite model property with respect to F-models. This, however, does not give a decision algorithm. Thus, we will construct a tableaux calculus that allows us to extract countermodels from open branches and use it to define a decision procedure that takes polynomial space. In [28], tableaux for $\mathbf{K}\mathsf{G}$ are presented but they use the fact that all its connectives are order-based which is not the case for $\mathscr{L}^\sim_{\Box,\Diamond}$. To incorporate $\sim$, we follow [22] and provide a *constraint tableaux calculus*. We begin by giving formal definitions of the needed notions. We will then explain them in further detail.

Definition 6 (Structures and constraints)**.** We fix countable sets $\mathsf{WL}=\{w,w',w_0,\ldots\}$ of *state-labels* and $\mathsf{Var}=\{c,d,e,c',\ldots\}$ of variables, define the set of T-symbols $\mathsf{T}=\{t_i(w)\mid w\in\mathsf{WL},i\in\mathbb{N}\}\cup\{t^s_i(w)\mid w\in\mathsf{WL},i\in\mathbb{N}\}\cup\{\mathbb{0},\mathbb{1}\}$, and let $\triangledown\in\{\leqslant,<,\geqslant,>,=\}$.

We define sets of *labelled formulas* (LF) and *relational terms* (RT) as follows:

$$\mathsf{LF}=\{w:\phi\mid w\in\mathsf{WL},\phi\in\mathscr{L}^\sim_{\Box,\Diamond}\}\qquad\qquad \mathsf{RT}=\{wRw'\mid w,w'\in\mathsf{WL}\}$$

The set VT of *value terms* and Str of *structures* are defined as follows:

$$\mathsf{VT}\ni\mathbb{T}::=c\in\mathsf{Var}\mid\mathbf{t}\in\mathsf{T}\mid\rho\in\mathsf{RT}\mid 0\mid 1\mid 1-\mathbb{T}\qquad\qquad \mathsf{Str}=\mathsf{VT}\cup\mathsf{LF}$$

Finally, *constraints* have the form $\Sigma\triangledown\mathbb{T}$ s.t. $\Sigma\in\mathsf{Str}$, $\mathbb{T}\in\mathsf{VT}$.

Definition 7 ($\mathscr{T}(\mathbf{K}\mathsf{G}_{\mathsf{inv}})$ — tableaux for $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$)**.** A *tableau* is a downward-branching tree whose nodes are constraints. Each branch can be extended by one of the rules listed below. We apply the following conventions: vertical bars denote branching; $\blacktriangledown,\triangledown\in\{\leqslant,<,\geqslant,>\}$, if $\triangledown$ is $\leqslant$, then $\blacktriangledown$ is $\geqslant$ and vice versa (likewise for $<$); $\triangleright\in\{\geqslant,>\}$, $\triangleleft\in\{\leqslant,<\}$; c, w', $t(w)$, and $t^s(w)$ are fresh on the branch; wRu occurs on the branch.

$$\sim:\frac{w:\sim\phi\triangledown\mathbb{T}}{w:\phi\blacktriangledown 1-\mathbb{T}}\qquad\qquad \wedge_\triangleright:\frac{w:\phi\wedge\chi\triangleright\mathbb{T}}{\begin{array}{c}w:\phi\triangleright\mathbb{T}\\ w:\chi\triangleright\mathbb{T}\end{array}}\qquad\qquad \wedge_\triangleleft:\frac{w:\phi\wedge\chi\triangleleft\mathbb{T}}{w:\phi\triangleleft\mathbb{T}\mid w:\chi\triangleleft\mathbb{T}}$$

$$\rightarrow_{\triangleright}: \frac{w:\phi\rightarrow\chi\triangleright\mathbb{T}}{w:\chi\triangleright\mathbb{T}\;\Big|\;\begin{array}{c}\mathbb{T}\triangleleft 1\\ w:\chi\geqslant c\\ w:\phi\leqslant c\end{array}} \qquad \rightarrow_{\leqslant}: \frac{w:\phi\rightarrow\chi\leqslant\mathbb{T}}{1\leqslant\mathbb{T}\;\Big|\;\begin{array}{c}w:\chi\leqslant\mathbb{T}\\ w:\phi\geqslant c\\ w:\chi<c\end{array}} \qquad \rightarrow_{<}: \frac{w:\phi\rightarrow\chi<\mathbb{T}}{\begin{array}{c}w:\chi<\mathbb{T}\\ w:\phi\geqslant c\\ w:\chi<c\end{array}}$$

$$\Box_{\triangleright}: \frac{w:\Box\phi\triangleright\mathbb{T}}{\begin{array}{c}w:\Box\phi=\mathbb{1}\\ 1\triangleright\mathbb{T}\end{array}\;\Big|\;\begin{array}{c}w:\Box\phi=t(w)\\ \mathbb{T}\triangleleft t(w)\\ w':\phi>wRw'\\ w':\phi<t^s(w)\end{array}} \qquad \Box_{\leqslant}: \frac{w:\Box\phi\leqslant\mathbb{T}}{1\leqslant\mathbb{T}\;\Big|\;\begin{array}{c}\mathbb{T}\geqslant t(w)\\ w':\phi>wRw'\\ w':\phi<t^s(w)\end{array}} \qquad \Box_{<}: \frac{w:\Box\phi<\mathbb{T}}{\begin{array}{c}\mathbb{T}>t(w)\\ w':\phi>wRw'\\ w':\phi<t^s(w)\end{array}}$$

$$\Diamond_{\triangleleft}: \frac{w:\Diamond\phi\triangleleft\mathbb{T}}{\begin{array}{c}w:\Diamond\phi=\mathbb{0}\\ \mathbb{T}\triangleright 0\end{array}\;\Big|\;\begin{array}{c}w:\Diamond\phi=t^s(w)\\ t^s(w)\triangleleft\mathbb{T}\\ wRw'>t(w)\\ w':\phi>t(w)\end{array}} \qquad \Diamond_{\geqslant}: \frac{w:\Diamond\phi\geqslant\mathbb{T}}{\mathbb{T}\leqslant 0\;\Big|\;\begin{array}{c}\mathbb{T}\leqslant t^s(w)\\ wRw'>t(w)\\ w':\phi>t(w)\end{array}} \qquad \Diamond_{>}: \frac{w:\Diamond\phi>\mathbb{T}}{\begin{array}{c}\mathbb{T}<t^s(w)\\ wRw'>t(w)\\ w':\phi>t(w)\end{array}}$$

$$\Box_{=}: \frac{w:\Box\phi=\mathbb{T}}{u:\phi\geqslant\mathbb{T}\;\Big|\;\begin{array}{c}u:\phi<\mathbb{T}\\ u:\phi\leqslant wRu\end{array}} \qquad \Diamond_{=}: \frac{w:\Diamond\phi=\mathbb{T}}{wRu\leqslant\mathbb{T}\;\Big|\;\begin{array}{c}u:\phi\leqslant\mathbb{T}\\ wRu>\mathbb{T}\end{array}}$$

Let $\mathscr{B}=\{c_1,\ldots,c_n\}$ be a branch with constraints c_1, …, c_n and let further, c^t be the result of replacing every $\mathbb{T}\in\mathsf{VT}$ and $\lambda\in\mathsf{LF}$ with variables $x_{\mathbb{T}}$ and x_{λ}, respectively, and every $\mathbb{n}\in\{\mathbb{0},\mathbb{1}\}$ with $n\in\{0,1\}$. Furthermore, for every $w\in\mathsf{WL}$ occurring on $\mathscr{B}$, we set

$$\mathsf{T}(w)=\{t(w)\mid t(w)\text{ is on }\mathscr{B}\}\cup\{t^s(w)\mid t^s(w)\text{ is on }\mathscr{B}\}\cup\{\mathbb{n}\mid\exists\phi\;w:\phi=\mathbb{n}\in\mathscr{B}\}$$

and define $\mathscr{B}$ to be *closed* iff the following system of inequalities

$$\{c_1^t,\ldots,c_n^t\}\cup\{x_{t(w)}<x_{t^s(w)}\mid t(w)\text{ and }t^s(w)\text{ occur on }\mathscr{B}\}$$

does not have a solution over $[0,1]$ s.t. for every $w\in\mathsf{WL}$, the following properties hold:

$$\begin{aligned}\mathsf{T}(w)=\{t(w),t^s(w)\}&\Rightarrow\begin{bmatrix}x_{t(w)}=0\;\&\;x_{t^s(w)}=\frac{1}{2} & \text{or}\\ x_{t(w)}=\frac{1}{2}\;\&\;x_{t^s(w)}=1 & \end{bmatrix}\\ |\mathsf{T}(w)|\geq 3&\Rightarrow\begin{bmatrix}\forall\mathbf{t}\in\mathsf{T}(w)\;\exists\mathbf{t}'\in\mathsf{T}(w):x_{\mathbf{t}}=1-x_{\mathbf{t}'}\text{ and}\\ \exists\mathbf{t}_1,\mathbf{t}_2,\mathbf{t}_3:x_{\mathbf{t}_1}=0\;\&\;x_{\mathbf{t}_2}=\frac{1}{2}\;\&\;x_{\mathbf{t}_3}=1\end{bmatrix}\\ &\quad\neg\exists t(w),t^s(w),t'(w):x_{t(w)}<x_{t'(w)}<x_{t^s(w)}\end{aligned}\tag{5}$$

A branch is *open* if it is not closed. A branch $\mathscr{B}$ is *complete* when for every premise of any rule occurring on $\mathscr{B}$, its conclusion also occurs in $\mathscr{B}$. The only exceptions are branches containing constraints $w:\phi<\mathbb{T}$ and $w:\phi<1$ or $w:\phi>\mathbb{T}$ and $w:\phi>0$. In this case, the rules are applied to the constraints containing $\mathbb{T}$, not 0 and 1.

Finally, $\phi\in\mathscr{L}^{\sim}_{\Box,\Diamond}$ *has a* $\mathscr{T}(\mathbf{KG}_{\mathsf{inv}})$ *proof* if there is a tableau beginning with $w:\phi<1$ s.t. all its branches are closed.

Definition 8 (Model realising a branch)**.** Let $\mathfrak{M} = \langle W, R, T, v\rangle$ be an F-model and $\mathscr{B}$ a tableau branch. An $\mathfrak{M}$-*realisation of* $\mathscr{B}$ is a map $\mathsf{rl} : \mathsf{WL} \cup \mathsf{Str} \to W \cup [0,1]$ s.t. $\mathsf{rl}(w) \in W$ and $\mathsf{rl}(\Sigma) \in [0,1]$ for every $\Sigma \in \mathsf{Str}$ and $w \in \mathsf{WL}$ occurring on $\mathscr{B}$, and the following properties hold:

1. $\mathsf{rl}(\mathbb{0}) = 0$ and $\mathsf{rl}(\mathbb{1}) = 1$;
2. $\mathsf{rl}(w\mathsf{R}w') = \mathsf{rl}(w)R\mathsf{rl}(w')$;
3. if $\mathsf{rl}(\mathbb{T}) = x$, then $\mathsf{rl}(1 - \mathbb{T}) = 1 - x$ for every $\mathbb{T} \in \mathsf{VT}$;
4. $(\{\mathsf{rl}(\mathbf{t}) \mid \mathbf{t} \in \mathsf{T}(w)\} \cup \{0, \frac{1}{2}, 1\}) \subseteq T(\mathsf{rl}(w))$ and $\mathsf{rl}(t(w)) < \mathsf{rl}(t^s(w))$ for all $w \in \mathsf{WL}$;
5. there are no $t'(w)$, $t(w)$, and $t^s(w)$ on $\mathscr{B}$ s.t. $\mathsf{rl}(t(w)) < \mathsf{rl}(t'(w)) < \mathsf{rl}(t^s(w))$;
6. if $\mathsf{T}(w) = \{t(w), t^s(w)\}$, then $\frac{1}{2} \in \{\mathsf{rl}(t(w)), \mathsf{rl}(t^s(w))\}$;
7. if $|\mathsf{T}(w)| \geq 3$, then for each $\mathbf{t} \in \mathsf{T}(w)$ on $\mathscr{B}$, there is $\mathbf{t}' \in \mathsf{T}(w)$ on $\mathscr{B}$ s.t. $\mathsf{rl}(\mathbf{t}) = 1 - \mathsf{rl}(\mathbf{t}')$, and there are $t_1, t_2, t_3 \in \mathsf{T}(w)$ s.t. $\mathsf{rl}(t_1) = 0$, $\mathsf{rl}(t_2) = \frac{1}{2}$, and $\mathsf{rl}(t_3) = 1$.

A constraint $w : \phi \triangledown \mathbb{T}$ is *realised by* $\mathfrak{M}$ *under* rl if $v(\phi, \mathsf{rl}(w)) \triangledown \mathsf{rl}(\mathbb{T})$. A constraint $\mathbb{T} \triangledown \mathbb{T}'$ is realised by $\mathfrak{M}$ under rl if $\mathsf{rl}(\mathbb{T}) \triangledown \mathsf{rl}(\mathbb{T}')$. $\mathscr{B}$ is realised by $\mathfrak{M}$ under rl if rl realises all constraints occurring on $\mathscr{B}$.

Note that each application of a modal rule to a structure $w : \Box\phi$ or $w : \Diamond\chi$ that introduces a witnessing state w' to the branch also produces two value terms: $t(w)$ and $t^s(w)$. Each pair of such terms denotes the two consecutive elements of $T(w)$ between which $v(\phi, w')$ lies. Note also that to guarantee the soundness of such rules, we demand (cf. items 4 and 5 in Definition 8) that the intervals do not overlap in a non-trivial way. I.e., we prohibit the situations such as $\mathsf{rl}(t_0(w)) < \mathsf{rl}(t_1(w)) < \mathsf{rl}(t_1^s(w)) < \mathsf{rl}(t_0^s(w))$. On the other hand, the intervals can coincide if this does not contradict other constraints on the branch.

For $\Box$ rules, this condition guarantees that the introduced $t(w)$ is indeed the maximal member of $T(w)$ that is still *less or equal to the 'real' value of* $\Box\phi$ according to the standard semantics. Dually, for $\Diamond$ rules, the condition ensures that the introduced $t^s(w)$ is the minimal member of $T(w)$ that is *greater or equal to the value of* $\Diamond\phi$ according to the standard semantics.

Let us now look at how the tableau proofs work. In the next example, we provide a failed tableau proof of $\Box p \to {\sim}\Diamond{\sim}p$ and show how to read off countermodels from complete open branches.

Example 3. Cf. Fig. 2. The tableau has three closed branches marked with $\times$ and a complete open branch marked with ☹. Let us check the closure of branches. In the leftmost branch, we use the fact that $\mathsf{rl}(t(w)) \in [0,1]$. Steps **17** and **20** directly contradict each other. In the rightmost branch, we use that the values of $t(w)$'s should be closed under $1 - x$. Namely, we have that $t_0(w) > 0$ (otherwise, it would contradict $c \leq t_0(w)$ and $w : {\sim}\Diamond{\sim}p < c$). Hence, the values of $t(w)$'s are ordered as follows: $t_1(w)$, $t_1^s(w)$, $t_0(w)$, $t_0^s(w)$. As $\mathbb{0}$ and $\mathbb{1}$ do not occur on the branch, we have either (I) $t_1(w) = 0$, $t_1^s(w) = t_0(w) = \frac{1}{2}$, $t_0^s(w) = 1$ or (II) $t_0(w) = 0$, $t_0^s(w) = t_1(w) = \frac{1}{2}$, $t_1^s(w) = 1$. Observe, however, that $t_1^s(w) = t_0(w) = \frac{1}{2}$ in (I) would contradict **16** and **22**, and $t_0(w) = 0$ in (II) would contradict **4** and **16**.

Thus, $w' : p \geqslant t_0(w)$ and $w' : p < 1 - t_1(w)$ cannot be realised at the same time.

Let us now discuss how to construct a realising model of the open branch ☹. We note quickly that in the general case, each complete open branch can produce uncountably many models because constraints give order relations, not precise values.[6] Thus, we will provide just one example of a countermodel.

Observe that $t(w) < wRw' \leqslant v(p, w') < 1$. As there are no other constraints on $t(w)$ and $t^s(w)$, we put $\mathsf{rl}(t(w)) = 0$, $\mathsf{rl}(t^s(w)) = \frac{1}{2}$, and $\mathsf{rl}(\mathbb{1}) = 1$. We now choose any value between 0 and 1 for $v(p, w')$ and set $wRw' = v(p, w')$ as the inequality is not strict. A countermodel can be seen in Fig. 2. Using Definition 3, we see that $v(\Box p, w) = 1$ and $v({\sim}\Diamond{\sim}p, w) = \frac{1}{2}$ which falsifies $\Box p \to {\sim}\Diamond{\sim}p$ at w.

We are now ready to show the soundness and completeness of our tableaux calculus.

[6] Note that this is not a problem from the practical standpoint as systems of linear inequalities over reals often have uncountably many solutions. Yet, they are algorithmically solvable.

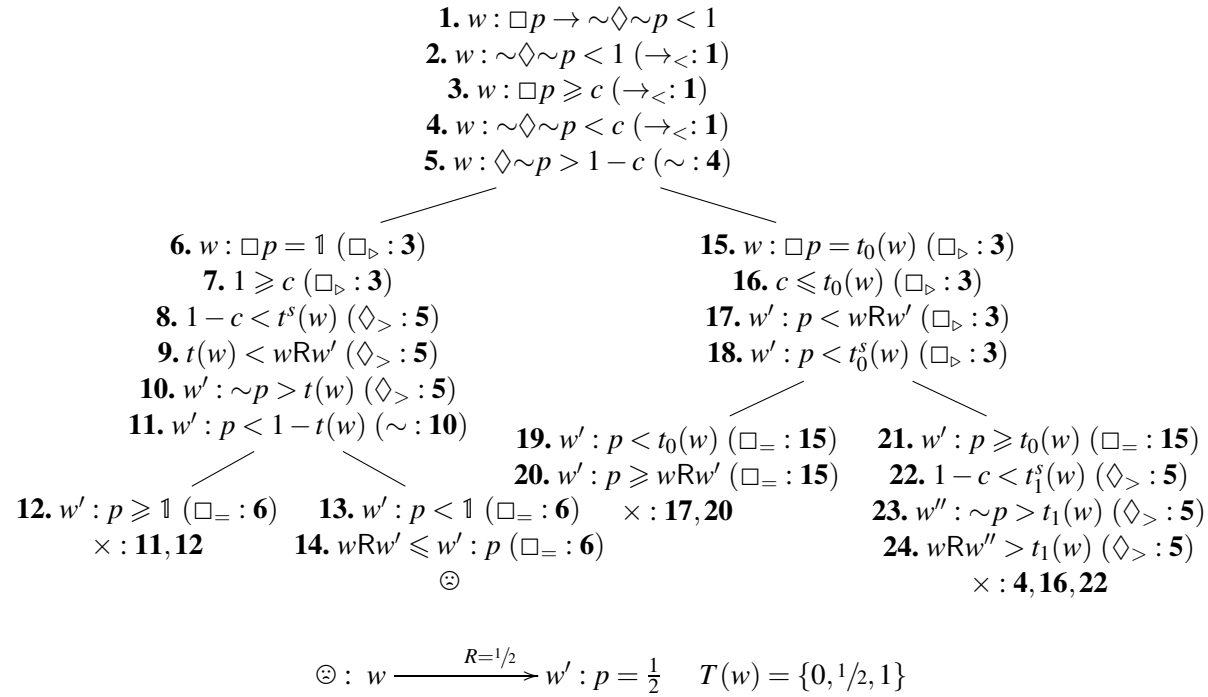


Figure 2: A failed $\mathscr{T}(\mathbf{K}\mathsf{G}_{\mathsf{inv}})$ proof of $\Box p \to \sim\Diamond\sim p$ and a model of the open branch (marked with ☹).

Theorem 2. *$\phi \in \mathscr{L}^{\sim}_{\Box,\Diamond}$ is $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-valid iff it has a $\mathscr{T}(\mathbf{K}\mathsf{G}_{\mathsf{inv}})$ proof.*

Proof. The proof is standard and follows [28, Theorems 6.19 and 6.21], so we only give a sketch thereof. For the soundness part, we show that if a tableau beginning with $w : \phi < 1$ is closed, then ϕ is $\mathbf{K}\mathsf{G}_{\mathsf{inv}}$-valid. We show that if $\mathfrak{M}$ realises the premises of a rule, then it should also realise at least one of its conclusions (note that branching rules have two conclusions). The soundness will follow since there is no F-model $\mathfrak{M}$ and realisation rl s.t. $\mathfrak{M}$ realises a closed branch $\mathscr{B}$ under rl.

We consider the case of the $\Box_\leqslant$ rule as an example. Let $\mathfrak{M} = \langle W, R, T, v\rangle$ realise $w : \Box\phi \leqslant \mathbb{T}$. If $\mathsf{rl}(\mathbb{T}) = 1$, then $1 \leqslant \mathbb{T}$ (the left conclusion) is realised. Consider the case where $\mathsf{rl}(\mathbb{T}) < 1$. Then (cf. Definition 3) $\max\{x \in T(w) \mid x \leq \inf_{w' \in W}\{\mathsf{rl}(w)Rw' \to_{\mathsf{G}} v(\phi, w')\}\} \leq \mathsf{rl}(\mathbb{T})$. This means that there is a state w' s.t. $v(\phi, w') < wRw'$ and there are $x, x' \in T(w)$ s.t. x' is the immediate successor of x, $v(\Box\phi, w) = x$, and $v(\phi, w') < x'$. Now setting $\mathsf{rl}(t(w)) = x$, $\mathsf{rl}(t^s(w)) = x'$, and $\mathsf{rl}(w') = w'$, we have that $t(w) \leqslant \mathbb{T}$ and $w' : \phi < t^s(w)$ are realised. Hence, the right conclusion is realised.

For the completeness part, we prove that for every complete open branch, there is an F-model $\mathfrak{M}$ and a realisation rl s.t. $\mathfrak{M}$ realises $\mathscr{B}$ under rl. Let $\mathscr{B}$ be a complete open branch and consider the system of inequalities $\mathscr{B}^t$ as shown in (5) and its solution as specified in Definition 7. Now define $\mathfrak{M}$ as follows: $W = \{w \mid w \text{ occurs in } \mathscr{B}\}$, $wRw' = x_{w\mathsf{R}w'}$, $T(w) = \{x_{t(w)} \mid t(w) \text{ occurs in } \mathscr{B}\} \cup \{x_{t^s(w)} \mid t^s(w) \text{ occurs in } \mathscr{B}\} \cup \{0, \frac{1}{2}, 1\}$ and $v(p, w) = x_{w:p}$ for every w, w', and $p \in \mathtt{Prop}$ occurring in $\mathscr{B}$. We also define $\mathsf{rl}(w) = w$, $\mathsf{rl}(\mathbb{T}) = x_{\mathbb{T}}$, and $\mathsf{rl}(w : \chi) = x_{w:\chi}$ for every $w \in \mathsf{WL}$, $\mathbb{T} \in \mathsf{VT}$, and $w : \chi \in \mathsf{LF}$ occurring on $\mathscr{B}$. It remains to show that all constraints are realised by $\mathfrak{M}$ under rl.

For all constraints of the form $\mathbb{T} \triangledown \mathbb{T}'$ that do not contain labelled formulas, we have $\mathsf{rl}(\mathbb{T}) \triangledown \mathsf{rl}(\mathbb{T}')$ by the construction of $\mathfrak{M}$ since $\mathscr{B}$ is a complete open branch. Let us now show that all constraints $u : \psi \triangledown \mathbb{T}$

on $\mathscr{B}$ are realised. We proceed by induction on formulas. Constraints of the form $w : p\nabla\mathbb{T}$ are realised by construction of $\mathfrak{M}$. The cases of constraints whose principal connective is $\sim$, $\wedge$, or $\rightarrow$ can be shown by simple applications of the induction hypothesis.

Let us consider the case of $w : \Box\phi \leq \mathbb{T}$ as an example and show that $v(\Box\phi, w) \leq \mathsf{rl}(\mathbb{T})$. Since $w : \Box\phi \leqslant \mathbb{T}$ is in $\mathscr{B}$ and $\mathscr{B}$ is complete and open, then it also contains (a) $1 \leqslant \mathbb{T}$ or (b) $t(w) \leqslant \mathbb{T}$, $w' : \phi < w\mathsf{R}w'$, and $w' : \phi < t^s(w)$. If (a) is the case, $\mathbb{T}$ is a value term (observe from Definition 6 that constraints can contain at most one labelled formula), whence $1 \leqslant \mathbb{T}$ is realised by the construction of $\mathfrak{M}$. Thus, $\mathsf{rl}(\mathbb{T}) = 1$ and $v(\Box\phi, w) \leq \mathsf{rl}(\mathbb{T})$, as required. If (b) is the case, then by the induction hypothesis $t(w) \leqslant \mathbb{T}$, $w' : \phi < w\mathsf{R}w'$, and $w' : \phi < t^s(w)$ are realised, whence, $v(\phi, w') < wRw'$, $x_{t(w)} \leq \mathsf{rl}(\mathbb{T})$, and $v(\phi, w') < x_{t^s(w)}$ with $\{x_{t(w)}, x_{t^s(w)}\} \subseteq T(w)$ and $x_{t^s(w)}$ being the immediate successor of $x_{t(w)}$. But in this case, it is clear that $v(\Box\phi, w) = \max\{x \in T(w) \mid x \leq \inf\limits_{w' \in W}\{wRw' \rightarrow_{\mathsf{G}} v(\phi, w')\}\} \leq x_{t(w)}$, whence, $v(\Box\phi, w) \leq \mathsf{rl}(\mathbb{T})$, as required. □

5 Complexity

Let us now use $\mathscr{T}(\mathbf{KG}_{\mathsf{inv}})$ to obtain the PSPACE-completeness of $\mathbf{KG}_{\mathsf{inv}}$. Our proof adapts the standard algorithm for **K** from [7].

Theorem 3. *Validity of* $\mathbf{KG}_{\mathsf{inv}}$ *is* PSPACE*-complete.*

Proof. PSPACE-hardness is immediate since **KG** is PSPACE-hard [15, Theorem 21] and $\mathbf{KG}_{\mathsf{inv}}$ is a conservative extension of **KG**. Let us now tackle the upper bound.

First of all, we observe that all tableaux terminate. Indeed, the branching factor of rules is at most 2, and every rule except for $\Box_{\triangleright}$ and $\Diamond_{\triangleleft}$ reduces the number of connectives in the labelled formula in the premise. If $\Box_{\triangleright}$ or $\Diamond_{\triangleleft}$ is applied, then $w : \Box\phi = t(w)$ or $w : \Diamond\phi = t^s(w)$, respectively, is introduced to which one can apply $\Box_{=}$ or $\Diamond_{=}$ which does decompose the formula. Thus, at some point, all formulas will be decomposed into atoms. We note, moreover, that given a branch $\mathscr{B}$, it takes us nondeterministic polynomial time to find a solution to $\mathscr{B}^t$ over $[0,1]$ s.t. (5) holds with respect to it.

Let us now provide a decision procedure that uses only polynomial space with respect to $\ell(\phi)$. We aim to build the model realising $w_1^0 : \phi < 1$ 'on the fly' (in which case, ϕ is not valid) or to show that it is impossible. Fig. 3 illustrates the process of building the model. As one sees, we *do not* build model branch-by-branch but rather generate all children of a given node, then pick one child, and again, generate all its children. This is because we need to apply rules $\Box_{=}$ and $\Diamond_{=}$ using generated states *all at once*. But these rules can be applied *only after* $\Box_{\triangleright}$ and $\Diamond_{\triangleleft}$.

We begin with $w_1^0 : \phi < 1$. In what follows, if a rule introduces branching to the tableau, we pick one branch, say, $\mathscr{B}$, and work with it depth-first. If the branch with which we are working is closed, we delete it and choose the next one. First, we apply propositional rules. Then we apply modal rules for constraints containing $\leqslant$ or $<$. This adds new states $w_{1,1}^1, \ldots, w_{n_1^0,1}^1$ (we will call them *children* of w_1^0), new members of $\mathsf{T}(w_1^0)$ of the form $t_i(w_1^0)$, $t_i^s(w_1^0)$, $\mathbb{0}$, and $\mathbb{1}$, relational terms $w_1^0\mathsf{R}w_{i,1}^1$, and new constraints of the form $\Box\chi = t_i(w_1^0)$ or $\Diamond\chi = t_i^s(w_1^0)$ for $i \in \{1, \ldots, n_1^0\}$. Note that $n_1^0 = \mathcal{O}(\ell(\phi))$, whence, we need only $\mathcal{O}(\ell(\phi))$ space to store the 'content' of w_1^0.

If $\mathscr{B}$ is still open, mark $w_{1,1}^1$ as 'active' and apply all rules for constraints of the form $\Box\chi = t_i(w_1^0)$ and $\Diamond\chi = t_i^s(w_1^0)$ using $w_{1,1}^1$ in the conclusion. Note that $\Box_{=}$ and $\Diamond_{=}$ are the only modal rules that can become applicable after the application of $\Box_{\triangleright}$ and $\Diamond_{\triangleleft}$. Once all $\Box_{=}$ and $\Diamond_{=}$ rules are applied, we apply the propositional rules to formulas labelled with $w_{1,1}^1$. We then apply modal rules for constraints

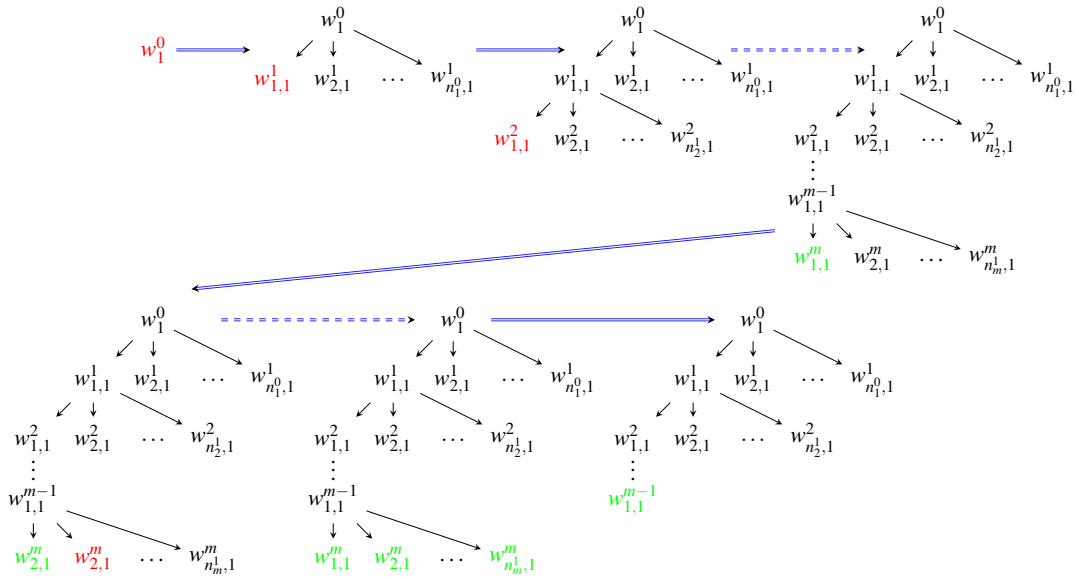


Figure 3: Building the model on the fly: red states are active; green states are safe.

with $\leqslant$ and $<$. As on the previous stage, we generate new states $w^2_{1,1}, \ldots, w^2_{n^1_2,1}$, constraints of the form $\Box\chi = t_i(w^1_{1,1})$ or $\Diamond\chi = t^s_i(w^1_{1,1})$, and members of $\mathsf{T}(w^1_{1,1})$ and relational terms corresponding to them. We repeat the procedure until we produce a state $w^m_{1,1}$ s.t. all constraints $w^m_{1,1} : \chi \nabla \mathbb{T}$ are atomic. If the branch is not closed, we mark $w^m_{1,1}$ as 'safe', delete all $w^m_{1,1} : \chi \nabla \mathbb{T}$'s and go to $w^m_{2,1}$ and repeat the procedure. Once all states $w^m_{1,1}, \ldots, w^m_{n^1_m,1}$ are marked as 'safe', we delete constraints containing them, mark $w^{m-1}_{1,1}$ as 'safe', and delete all its children. Then we proceed to $w^{m-1}_{2,1}$. The procedure is repeated until either all tableau branches are closed (which means that ϕ is valid) or w^0_1 is marked as 'safe' (and thus, we have a complete open branch, whence, ϕ is not valid).

Note that the length of the branch of the constructed model is bounded from above by the modal depth of ϕ and is thus $\mathcal{O}(\ell(\phi))$. In addition, each state of the branch of the model can have at most $\mathcal{O}(\ell(\phi))$ children because their number is bounded from above by the number of modalities on the same level of nesting. Thus, we have to store $\mathcal{O}((\ell(\phi))^2)$ states each of which contains constraints taking up to $\mathcal{O}(\ell(\phi))$ space. Thus, our decision procedure utilises $\mathcal{O}((\ell(\phi))^3)$ space. □

6 Conclusion

We presented an expansion of the Gödel modal logic with the involutive negation and showed how one can use it to formalise reasoning about uncertainty. We also constructed an alternative semantics for $\mathbf{KG}_{\mathsf{inv}}$ following [14, 15] with respect to which $\mathbf{KG}_{\mathsf{inv}}$ has the finite model property. We then used these semantics to provide the tableaux calculus $\mathscr{T}(\mathbf{KG}_{\mathsf{inv}})$ that allows us to read off finite countermodels from complete open branches. Finally, we proved PSPACE-completeness of $\mathbf{KG}_{\mathsf{inv}}$-validity using $\mathscr{T}(\mathbf{KG}_{\mathsf{inv}})$.

Our next steps are as follows. First, we want to continue the systematic study of $\mathbf{KG}_{\mathsf{inv}}$. In particular, we plan to provide a complete Hilbert-style axiomatisation of $\mathbf{KG}_{\mathsf{inv}}$ over crisp and fuzzy frames. This will help investigate the correspondence between the classes of crisp and fuzzy frames on the one side

and the $\mathscr{L}^{\sim}_{\Box,\Diamond}$-formulas that define these classes of frames, on the other. Furthermore, in [23, 24], hypersequent calculi for $\Box$- and $\Diamond$-fragments of $\mathbf{KG}$ are presented. Since these calculi are cut-free, they can serve as a syntactical decision procedure (i.e., the one that does not directly produce a countermodel) for the fragments of $\mathbf{KG}$. It thus makes sense to construct hypersequent cut-free calculi for the bi-modal $\mathbf{KG}$ and its expansions and obtain an alternative, syntactical, decision procedure for $\mathbf{KG}$.

Second, in [4, 5, 6], we considered $\mathbf{KG}^2$ — an expansion of $\mathbf{KG}$ with a so-called 'strong negation' from [30] with semantics defined on $[0,1]^{\bowtie}$ — a bi-lattice product of $[0,1]$ with itself. In particular, it was shown that a class of frames is $\mathbf{KG}^2$-definable iff it is $\mathbf{KG}$-definable by a self-dual formula (i.e., a formula equivalent to the negation of the result of replacing every connective with its dual: $\wedge$ with $\vee$, $\rightarrow$ with $\prec$, and $\Box$ with $\Diamond$). Note, however, that $\mathbf{KG}^2$ does not introduce connectives defined w.r.t. the second order on $[0,1]^{\bowtie}$. On the other hand, in [29], it is shown that the modal logic on $\{0,1\}^{\bowtie}$ in the language with all bi-lattice connectives and modalities and the bi-modal classical modal logic have the same expressivity. It would be thus instructive to consider a full bi-lattice expansion of $\mathbf{KG}^2$ with semantics on $[0,1]^{\bowtie}$ and compare its expressivity to $\mathbf{KG}$ and its expansions.

References

[1] J.P. Aguilera, M. Diéguez, D. Fernández-Duque & B. McLean (2022): *Time and Gödel: Fuzzy Temporal Reasoning in PSPACE*. In: *Logic, Language, Information, and Computation*, *Lecture notes in computer science* 13368, Springer International Publishing, Cham, pp. 18–35, doi:10.1007/978-3-031-15298-6_2.

[2] J.P. Aguilera, M. Diéguez, D. Fernández-Duque & B. McLean (2022): *A Gödel Calculus for Linear Temporal Logic*. In: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 2–11, doi:10.24963/kr.2022/1.

[3] 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.

[4] M. Bílková, S. Frittella & D. Kozhemiachenko (2022): *Paraconsistent Gödel Modal Logic*. In: *Automated Reasoning*, *Lecture notes in computer science* 13385, Springer International Publishing, Cham, pp. 429–448, doi:10.1007/978-3-031-10769-6_26.

[5] M. Bílková, S. Frittella & D. Kozhemiachenko (2023): *Crisp bi-Gödel modal logic and its paraconsistent expansion*. *Logic Journal of the IGPL*, doi:10.1093/jigpal/jzad017.

[6] M. Bílková, S. Frittella & D. Kozhemiachenko (2025): *Fuzzy bi-Gödel modal logic and its paraconsistent relatives*. *Journal of Logic and Computation* 35(2), doi:10.1093/logcom/exae011.

[7] P. Blackburn, M. de Rijke & Y. Venema (2010): *Modal logic*, 4th edition. *Cambridge tracts in theoretical computer science* 53, Cambridge University Press.

[8] F. Bobillo, M. Delgado, J. Gómez-Romero & U. Straccia (2009): *Fuzzy description logics under Gödel semantics*. *International Journal of Approximate Reasoning* 50(3), pp. 494–514, doi:10.1016/j.jcss.2017.03.012.

[9] F. Bobillo, M. Delgado, J. Gómez-Romero & U. Straccia (2012): *Joining Gödel and Zadeh fuzzy logics in fuzzy description logics*. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 20(04), pp. 475–508, doi:10.1142/S0218488512500249.

[10] S. Borgwardt (2014): *Fuzzy Description Logics with General Concept Inclusions*. Ph.D. thesis, Technische Universität Dresden, Dresden.

[11] S. Borgwardt, F. Distel & R. Peñaloza (2014): *Decidable Gödel description logics without the finitely-valued model property*. In: *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*, pp. 228–237, doi:10.5555/3031929.3031959.

[12] S. Borgwardt, F. Distel & R. Peñaloza (2014): *Gödel description logics with general models*. In: *DL 2014: Informal Proceedings of the 27th International Workshop on Description Logics, Vienna, Austria, July 17–20, 2014*, 1193, CEUR, pp. 391–403.

[13] S. Borgwardt & R. Peñaloza (2017): *Algorithms for reasoning in very expressive description logics under infinitely valued Gödel semantics*. *International Journal of Approximate Reasoning* 83, pp. 60–101, doi:10.1016/j.ijar.2016.12.014.

[14] X. Caicedo, G. Metcalfe, R. Rodríguez & J. Rogger (2013): *A finite model property for Gödel modal logics*. In L. Libkin, U. Kohlenbach & R. de Queiroz, editors: *International Workshop on Logic, Language, Information, and Computation*, *Lecture Notes in Computer Science* 8071, Springer, Berlin, Heidelberg, pp. 226–237, doi:10.1007/978-3-642-39992-3_20.

[15] X. Caicedo, G. Metcalfe, R. Rodríguez & J. Rogger (2017): *Decidability of order-based modal logics*. *Journal of Computer and System Sciences* 88, pp. 53–74, doi:10.1016/j.jcss.2017.03.012.

[16] X. Caicedo & R.O. Rodriguez (2010): *Standard Gödel modal logics*. *Studia Logica* 94(2), pp. 189–214, doi:10.1007/s11225-010-9230-1.

[17] X. Caicedo & R.O. Rodríguez (2015): *Bi-modal Gödel logic over [0,1]-valued Kripke frames*. *Journal of Logic and Computation* 25(1), pp. 37–55, doi:10.1093/logcom/exs036.

[18] M.E. Coniglio, F. Esteva, J. Gispert & L. Godo (2021): *Degree-preserving Gödel logics with an involution: Intermediate logics and (ideal) paraconsistency*. In O. Arieli & A. Zamansky, editors: *Arnon Avron on Semantics and Proof Theory of Non-Classical Logics*, *Outstanding contributions to logic* 21, Springer International Publishing, Cham, pp. 107–139, doi:10.1007/978-3-030-71258-7_6.

[19] R. Ertola, F. Esteva, T. Flaminio, L. Godo & C. Noguera (2015): *Paraconsistency properties in degree-preserving fuzzy logics*. *Soft Computing* 19(3), pp. 531–546, doi:10.1007/s00500-014-1489-0.

[20] F. Esteva, L. Godo, P. Hájek & M. Navara (2000): *Residuated fuzzy logics with an involutive negation*. *Archive for Mathematical Logic* 39(2), pp. 103–124, doi:10.1007/s001530050006.

[21] R. Goré (2000): *Dual Intuitionistic Logic Revisited*. In R. Dyckhoff, editor: *Automated Reasoning with Analytic Tableaux and Related Methods. TABLEAUX 2000*, *Lecture notes in computer science* 1847, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 252–267, doi:10.1007/10722086_21.

[22] R. Hähnle (1994): *Many-valued logic and mixed integer programming*. *Annals of Mathematics and Artificial Intelligence* 12(3-4), pp. 231–263, doi:10.1007/BF01530787.

[23] G. Metcalfe & N. Olivetti (2009): *Proof systems for a Gödel modal logic*. In: *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*, *Lecture Notes in Artificial Intelligence* 5607, Springer, pp. 265–279, doi:10.1007/978-3-642-02716-1_20.

[24] G. Metcalfe & N. Olivetti (2011): *Towards a Proof Theory of Gödel Modal Logics*. *Logical Methods in Computer Science* 7(2), doi:10.2168/LMCS-7(2:10)2011.

[25] C. Rauszer (1974): *A Formalization of The Propositional Calculus of H–B Logic*. *Studia Logica* 33, pp. 23–34, doi:10.1007/BF02120864.

[26] R.O. Rodriguez, O.F. Tuyt, F. Esteva & L. Godo (2022): *Simplified Kripke semantics for K45-like Gödel modal logics and its axiomatic extensions*. *Studia Logica* 110(4), pp. 1081–1114, doi:10.1007/s11225-022-09987-0.

[27] R.O. Rodriguez & A. Vidal (2021): *Axiomatization of Crisp Gödel Modal Logic*. *Studia Logica* 109, pp. 367–395, doi:10.1007/s11225-020-09910-5.

[28] J. Rogger (2016): *Decidability of Order-Based Modal Logics*. Ph.D. thesis, University of Bern.

[29] S.O. Speranski (2022): *Modal Bilattice Logic and its Extensions*. *Algebra and Logic* 60(6), pp. 407–424, doi:10.1007/s10469-022-09667-x.

[30] H. Wansing (2008): *Constructive negation, implication, and co-implication*. *Journal of Applied Non-Classical Logics* 18(2–3), pp. 341–364, doi:10.3166/jancl.18.341-364.

Comparing State-Representations for DEL Model Checking

Gregor Behnke, Malvin Gattinger, Haitian Wang
ILLC, University of Amsterdam
Amsterdam, The Netherlands
g.behnke@uva.nl malvin@w4eg.eu h.wang12@uva.nl

Avijeet Ghosh
Chennai Mathematical Institute
Chennai, India
avi.ghosh23@gmail.com

Model checking with the standard Kripke models used in (Dynamic) Epistemic Logic leads to scalability issues. Hence alternative representations have been developed, in particular symbolic structures based on Binary Decision Diagrams (BDDs) and succinct models based on mental programs. While symbolic structures have been shown to perform well in practice, their theoretical complexity was not known so far. On the other hand, for succinct models model checking is known to be PSPACE-complete, but no implementations are available.

We close this gap and directly relate the two representations. We show that model checking DEL on symbolic structures encoded with BDDs is also PSPACE-complete. In fact, already model checking Epistemic Logic without dynamics is PSPACE-complete on symbolic structures. We also provide direct translations between BDDs and mental programs. Both translations yield exponential outputs. For the translation from mental programs to BDDs we show that no small translation exists. For the other direction we conjecture the same.

1 Introduction

Reasoning about knowledge and its representation has become an important line of study in Artificial Intelligence, most importantly in Epistemic Planning [5]. A widely used logical framework to reason about the knowledge of multiple intelligent agents is Epistemic Logic, a version of Modal Logic [8]. Its standard semantics uses Kripke models with possible worlds to represent what agents know. Moreover, knowledge can change. For example in a card game, a player may announce its card to a subset of players privately. Then the knowledge of all players changes: some players come to know the card, and other players may still observe that such a private announcement is made. To formalise and reason about such knowledge updates, a widely used framework is Dynamic Epistemic Logic (DEL) [4, 13], of which the simplest version is Public Announcement Logic (PAL) [25, 23].

One way to make DEL and PAL practically useful is *model checking*: given a model and a formula, decide whether the formula holds. It is known that on Kripke models the computational complexity of this task for PAL is in P and for DEL is in PSPACE [1]. Model-checkers for PAL have been used in verifying protocols with security applications like the Russian card problem [14, 22]. Moreover, DEL model checkers like SMCDEL [17] can be used in combination with large language models (LLMs) in order to improve the explainability and reasoning ability compared to plain LLMs [26].

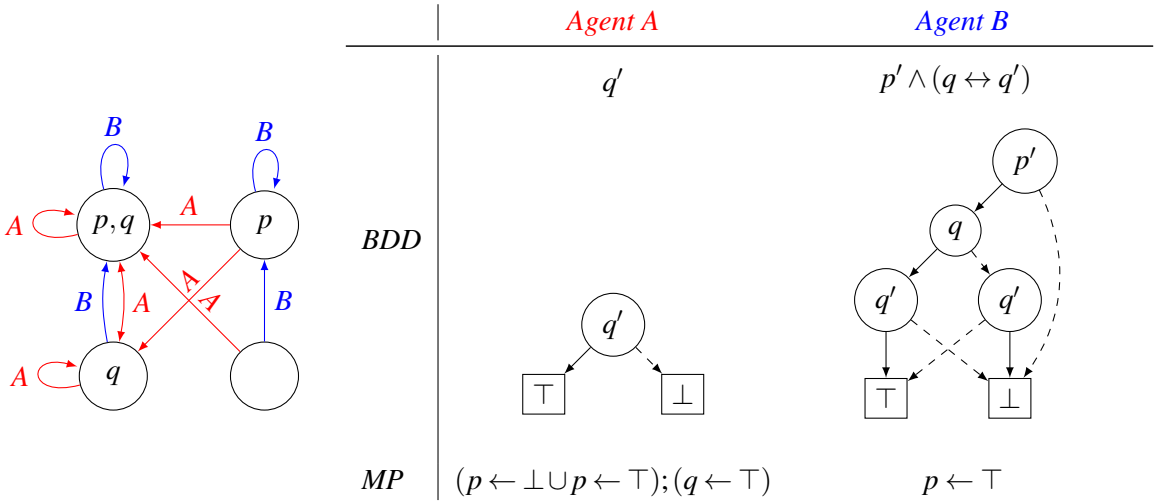
Model checking has scalability issues, because Kripke models enumerate all possible situations. For example, every possible way to distribute cards among players needs a possible world, leading to an exponentially large model. Different techniques have been developed to represent Kripke models more compactly, to make epistemic model-checking easier. Here we focus on two such works for DEL:

- Symbolic structures based on Binary Decision Diagrams (BDDs) [7, 17, 24]. This approach uses the famous data structure from [9] and is similar to techniques for temporal logics [10].
- Succinct models based on mental programs (MPs) [12], a variant of PDL [16], and similar to regular languages. In particular these programs can encode the epistemic accessibility relations.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 233–250, doi:10.4204/EPTCS.437.21

© G. Behnke, M. Gattinger, A. Ghosh & H. Wang
This work is licensed under the
Creative Commons Attribution License.

Example 1.1. *To illustrate the different representations of Kripke models, consider the model below. It has four worlds and two agents. Agent A considers all q-worlds possible and knows nothing else, whereas Agent B only considers p-worlds possible and knows the value of q. Next to the model we show how these relations are encoded by BDDs and by mental programs. We note that one relation results in a relatively small BDD, while the other is more suitable to be encoded as a short mental programs.*



Both approaches lead to a new model-checking problem, because when moving from one representation to another, the complexity of model checking can differ. Interestingly, for the succinct representation theoretical results about the model checking complexity are available, but not yet for the symbolic structures. With the present article we close this gap. We summarise the known and new results in Table 1. Note that when using symbolic structures, model checking EL, PAL and DEL all become equally hard. We discuss why this happens after we introduce knowledge structures in Definition 2.1.

Representation	Kripke models	Symbolic Structures	Succinct models
EL-S5	in P [15]	PSPACE: Theorem 2.11	-
PAL-S5	in P [13]	PSPACE: Theorem 2.11	-
DEL-S5	PSPACE [20]	PSPACE: Theorem 2.17	-
EL-K	in P [15]	PSPACE: Theorem 2.16	PSPACE [2]
PAL-K	in P [8]	PSPACE: Theorem 2.16	PSPACE [12]
DEL-K	PSPACE [1]	PSPACE: Theorem 2.16	PSPACE [12]

Table 1: Model checking complexity results. We write "PSPACE" for PSPACE-complete here.

While the EL and PAL rows in Table 1 show an easier complexity for Kripke models than for symbolic structures, the key difference lies in input size: symbolic structures can represent Kripke models exponentially more succinctly. Thus, polynomial-time algorithms for Kripke models may become exponential when measured against symbolic input. On the other hand, the DEL rows show that symbolic structures do not lead to an additional complexity blowup. The results suggest that one should not translate Kripke models to other representations, but work directly with the symbolic or succinct representation, i.e. a given system description should be formalized directly into a structure with BDDs or mental programs, to avoid ever using exponential memory that would be needed by a Kripke model.

The symbolic structures have been implemented and benchmarks show that they outperform standard Kripke models, but the succinct representations have not been implemented, with the exception of [21].

To better understand the relation between the different encodings of relations, we provide translations from BDDs to mental programs and vice versa, and study the translation complexity.

In the remainder of this section we define the languages. In section 2 we prove that symbolic model checking is PSPACE-complete, first for PAL-S5, then for DEL-K. In section 3 we recall mental programs used in succinct DEL, and in section 4 we provide translations between BDDs and mental programs. We conclude with comments about multi-pointed models and ongoing implementation work in section 5.

Definition 1.2. *Throughout this article we work with finite vocabularies V that come with an order. For any vocabulary V we define* Boolean formulas $\beta \in \mathscr{L}_B(V)$ *by* $\beta ::= \top \mid \bot \mid p \mid \neg\beta \mid \beta \wedge \beta$ *where* $p \in V$. *A* state *is a subset of the vocabulary $s \subseteq V$. That is, we identify states with the atomic propositions that are true at them. The Boolean semantics are defined as usual and denoted by* $s \vDash \beta$.

To work with both PAL and DEL, we will use a general language definition with a parameter set D.

Definition 1.3. *For any vocabulary V and any set D the language $\mathscr{L}_D(V)$ is given by* $\varphi ::= \top \mid \bot \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid [d]\varphi$ *where $p \in V$, $i \in A$ and $d \in D$. Concretely, $\mathscr{L}_{\mathsf{EL}}$ is given by $D = \varnothing$, $\mathscr{L}_{\mathsf{PAL}}$ is given by $D = \{!\varphi \mid \varphi \in \mathscr{L}_{\mathsf{PAL}}\}$, and $\mathscr{L}_{\mathsf{DEL}}$ is given by letting D be the set of events from Definition 2.3.*

The simpler logic we consider is Public Announcement Logic (PAL), where the only actions are truthful public announcements, i.e. D is the set of all formulas, used as announcements. More general actions include private announcements or factual (also called ontic) changes. In DEL these are usually modeled by letting D be the set of all pointed action models [4, 13], but here we will mostly use the transformers from [7]. Yet another option are the succinct event models from [12]. All these languages come with a mutual but well-founded recursion: elements of D are made using formulas, and formulas use elements of D. We refer to [13, Chapter 6] for details.

The following definition will be relevant to determine the input size of model checking problems.

Definition 1.4. *We define the length* $|\cdot| \colon \mathscr{L}_{\mathsf{PAL}}(V) \cup \mathscr{L}_{\mathsf{DEL}}(V) \to \mathbb{N}$ *of formulas as follows*

$$
\begin{array}{llll}
|\top| := 1 & |\neg\varphi| := |\varphi|+1 & |[!\psi]\varphi| := |\psi|+|\varphi|+1 \\
|\bot| := 1 & |\varphi_1 \wedge \varphi_2| := |\varphi_1|+|\varphi_2|+1 & |[\mathscr{X},x]\varphi| := |\mathscr{X}|+|\varphi|+1 \\
|p| := 1 & |K_i\varphi| := |\varphi|+1 &
\end{array}
$$

where $|\mathscr{X}|$ denotes the size of a transformer as in Definition 2.4 below.

2 Symbolic Structures with BDDs

We recall the definition and main features of the knowledge and belief structures from [7, 17]. In general we assume that elements of $\mathscr{L}_B$ are represented using Binary Decision Diagrams (BDDs) from [9].

Definition 2.1. *A* knowledge structure *is a tuple $\mathscr{F} = (V, \theta, O)$ where V is a finite set called* vocabulary, *$\theta \in \mathscr{L}_B(V)$ is the* state law *and $O_i \subseteq V$ for each $i \in A$ are the* observables. *A* state *of $\mathscr{F}$ is a $s \subseteq V$ such that $s \vDash \theta$. A* pointed *knowledge structure is a tuple $(\mathscr{F}, s)$ where s is a state of $\mathscr{F}$.*

Knowledge structures encode the knowledge of agents (that in Kripke models is represented by a relation) with a set of observables. Each $O_i \subseteq V$ encodes a relation $R_i \subseteq \mathscr{P}(V) \times \mathscr{P}(V)$ given by $R_i st \iff s \cap O_i = O_i \cap t$. Note that this is always an equivalence, matching S5 Kripke models. Hence the corresponding frameworks are called EL-S5, PAL-S5 and DEL-S5. The more general belief structures (for EL-K, PAL-K and DEL-K) can encode arbitrary relations using formulas over a double vocabulary. In addition to this, note that in Table 1, even though model checking for PAL and EL were much easier than DEL (P for PAL,EL whereas PSPACE for DEL) using Kripke models, when it comes to model

checking using symbolic structures, all three of them becomes equally hard (PSPACE). This is because all possible states in a Kripke model are listed distinctly when the model is represented in a general Kripke structure. Therefore, going over all of them does not cost any computational resource beyond the size of the Kripke model itself. But when represented using symbolic structures, only the set of propositions V is part of the input. Hence in order to iterate through indistinguishable possibilities of an agent (for example for checking formulas with $\hat{K}_i$ modalities), all possible (at most $2^{|V|}$) states that come out of V have to be traversed and checked. Since such $\hat{K}_i$ modalities are present in EL, PAL as well as DEL, model checking all three of them becomes equally hard.

Definition 2.2. *A* belief structure *is a tuple $\mathscr{F} = (V, \theta, \Omega)$ where V and θ are as in Definition 2.1 and for each agent i we have an* observation law $\Omega_i \in \mathscr{L}_B(V \cup V')$ *where V' denotes a fresh copy of V.*

The intuition behind Ω_i is that it is true at a pair of states iff the states are related. That is, a Boolean formula $\Omega_i \in \mathscr{L}_B(V \cup V')$ encodes the relation $R_i \subseteq \mathscr{P}(V) \times \mathscr{P}(V)$ given by $R_i st :\iff s \cup t' \vDash \Omega_i$. For example, if $\Omega_{\text{Alice}} = p \wedge q'$ then from any state where p is true Alice will consider any other state where q is true possible. In particular we will have a loop at the state $\{p,q\}$ because $\{p,q\} \cup \{p,q\}' \vDash p \wedge q'$. This encoding is also widely used for temporal model checking [10, Section 8.3].

Analogous to how symbolic structures encode Kripke models, transformers encode action models. To keep track of old propositional values before factual change, we introduce another set of fresh variables, denoted by the operation $(\cdot)^\circ$. Just as V' is a fresh copy of V, V_-° is a fresh copy of V_-.

Definition 2.3. *A* transformer *is a tuple $\mathscr{X} = (V^+, \theta^+, V_-, \theta_-, \Omega_i^+)$ where V^+ is such that $V \cap V^+ = \varnothing$, $\theta^+ \in \mathscr{L}_{\mathsf{DEL}}(V \cup V^+)$ is called the* event law*, $V_- \subseteq V$ is called the* modified subset*, $\theta_- : V_- \to \mathscr{L}_B(V \cup V^+)$ is called the* change law *and $\Omega_i^+ \in \mathscr{L}_B(V \cup V')$ for each i. An* event *is a pair $(\mathscr{X}, x)$ where $x \subseteq V^+$. To update $(\mathscr{F}, s)$ with (χ, x), let $\mathscr{F} \times \chi := (V^{new}, \theta^{new}, \Omega_i^{new})$ where $V^{new} = V \cup V^+ \cup V_-^\circ$,*

$$\theta^{new} = [V_- \mapsto V_-^\circ](\theta \wedge \|\theta^+\|_{\mathscr{F}}) \wedge \bigwedge_{q \in V^-} (q \leftrightarrow [V_- \mapsto V_-^\circ](\theta_-(q)))$$

where $[p \mapsto \psi]\varphi$ denotes a substitution, $\Omega_i^{new} = ([V_- \mapsto V_-^\circ][(V_-)' \mapsto (V_-^\circ)']\Omega_i) \wedge \Omega_i^+$, and lastly the new actual state is $s^x := (s \setminus V_-) \cup (s \cap V_-)^\circ \cup x \cup \{p \in V_- | s \cup x \vDash \theta_-(p)\}$.

Since our goal is a complexity study of model-checking algorithms, we need to be precise about the size of all structures we take as inputs. We define the size of transformers as follows.

Definition 2.4. *Given a transformer $\mathscr{X} = (V^+, \theta^+, V_-, \theta_-, \Omega_i^+)$, we define its size by mutual recursion with Definition 1.4 as $|\mathscr{X}| := |V^+| + |\theta^+| + |V_-| + \sum_{p \in V_-} |\theta_-(p)| + \sum_{i \in I} |\Omega_i^+|$.*

Also Definition 2.3 above and the following two definitions are mutually recursive.

Definition 2.5. *The semantics for $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ on knowledge and belief structures are as follows. We omit the standard Boolean cases.*

1. *For knowledge structures: $(\mathscr{F}, s) \vDash K_i\varphi$ iff for all states t of $\mathscr{F}$, if $s \cap O_i = t \cap O_i$, then $(\mathscr{F}, t) \vDash \varphi$.*
   For belief structures: $(\mathscr{F}, s) \vDash K_i\varphi$ iff for all states t of $\mathscr{F}$, if $s \cup t' \vDash \Omega_i$, then $(\mathscr{F}, t) \vDash \varphi$.
2. *$(\mathscr{F}, s) \vDash [!\psi]\varphi$ iff $(\mathscr{F}, s) \vDash \psi$ implies $(\mathscr{F}^\psi, s) \vDash \varphi$ where $\mathscr{F}^\psi := (V, \theta \wedge \|\psi\|_{\mathscr{F}}, O)$.*
3. *$(\mathscr{F}, s) \vDash [\mathscr{X}, x]\varphi$ iff $(\mathscr{F}, s) \vDash [x \sqsubseteq V^+]\theta^+$ implies $(\mathscr{F} \times \mathscr{X}, s^x) \vDash \varphi$.*

Definition 2.6. *For any structure $\mathscr{F}$ and any formula $\varphi \in \mathscr{L}_{PAL}(V) \cup \mathscr{L}_{DEL}(V)$ we define its* local Boolean translation *$\|\varphi\|_{\mathscr{F}}$ as follows.*

$$\|\top\|_{\mathscr{F}} := \top \qquad \|\bot\|_{\mathscr{F}} := \bot \qquad \|p\|_{\mathscr{F}} := p \qquad \|\neg\psi\|_{\mathscr{F}} := \neg\|\psi\|_{\mathscr{F}} \qquad \|\psi_1 \wedge \psi_2\|_{\mathscr{F}} := \|\psi_1\|_{\mathscr{F}} \wedge \|\psi_2\|_{\mathscr{F}}$$

If $\mathscr{F}$ is a knowledge structure, let $\|K_i\psi\|_{\mathscr{F}} := \forall(V \setminus O_i)(\theta \to \|\psi\|_{\mathscr{F}})$. If $\mathscr{F}$ is a belief structure, let $\|K_i\psi\|_{\mathscr{F}} := \forall V'(\theta' \to (\Omega_i \to (\|\psi\|_{\mathscr{F}})'))$. Let $\|[\psi]\xi\|_{\mathscr{F}} := \|\psi\|_{\mathscr{F}} \to \|\xi\|_{\mathscr{F}^\psi}$, where $\mathscr{F}^\psi := (V, \theta \wedge \|\psi\|_{\mathscr{F}}, O)$. Lastly, let $\|[\mathscr{X},x]\varphi\|_{\mathscr{F}} := \|[x \sqsubseteq V^+]\theta^+\|_{\mathscr{F}} \to [V_-^\circ \mapsto V_-][x \sqsubseteq V^+][V_- \mapsto \theta_-(V_-)]\|\varphi\|_{\mathscr{F}\times\mathscr{X}}$ where $F \times \mathscr{X}$ is from Definition 2.3.

The above Boolean translation is equivalent, i.e. $\mathscr{F}, s \vDash \varphi$ iff $s \vDash \|\varphi\|_{\mathscr{F}}$. The SMCDEL implementation [19] uses the translation, as it is faster than using Definition 2.5. For any Kripke model there exists an equivalent symbolic structure and vice versa, and similarly for action models and transformers [17].

There are now six different model checking problems: The three languages $\mathscr{L}_{\mathsf{EL}}$, $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ can each be interpreted on knowledge structures (S5) and on belief structures (K). We first consider the easiest case PAL-S5 and then the most general case DEL-K. Just like general Kripke models are more general than those for S5, the Ω_i can express anything that O_i can do, i.e. every knowledge structure can be seen as a belief structure, replacing O_i, with the formula $\Omega_i := \bigwedge_{p \in O_i}(p \leftrightarrow p')$. This needs roughly twice as much memory, resulting in no significant increase of the model-checking input size.

2.1 Symbolic Model Checking PAL-S5

To define the model checking problem for symbolic structures we need to say how exactly the input is given. For V and O_i this is obvious, but for θ we stress that we assume a BDD and not a formula.

Definition 2.7. *The* symbolic model checking task for PAL for S5 *is the following. Given a knowledge structure $\mathscr{F} = (V, \theta, O_i)$ where θ is a Boolean function encoded as a BDD, an actual state s and a formula $\varphi \in \mathscr{L}_{\mathsf{PAL}}(V)$ which may contain dynamic modalities, decide whether $\mathscr{F}, s \vDash \varphi$. The input size is $|V| + |\theta| + |O_i| + |\varphi|$ where $|\theta|$ is the node count of the BDD and $|\varphi|$ is the length of the formula.*

We will show that this problem is PSPACE complete. The following theorem shows that the problem is already PSPACE-hard for EL (i.e. without announcements). This means that the symbolic representation matters, because model checking EL on standard Kripke models is in P [15], but it becomes harder when moving from Kripke models to symbolic structures.

Theorem 2.8. *Model checking $\mathscr{L}_{\mathsf{EL}}$ on knowledge structures is PSPACE-hard.*

Proof. We reduce the evaluation of a Quantified Boolean Formula (QBF) to model checking $\mathscr{L}_{\mathsf{EL}}$ on a knowledge structure as follows. Before going further into QBFs, let us introduce a notation. For a set of propositions $P = \{p_1, p_2, \ldots, p_k\}$, the notation $\forall P \equiv \forall p_1 \forall p_2 \ldots \forall p_k$. Same is true for $\exists P$.

Take any QBF (wlog. in prenex form) $\psi = \forall P_1 \exists P_2 \ldots \forall P_{n-1} \exists P_n \varphi$ where φ is a Boolean formula (possibly in CNF) over $P_1 \cup \ldots \cup P_n$. Note that there are no free variables in ψ. Let $V = \bigcup_i P_i$ and let $\theta = \top$. Let the set of agents be $\{1, \ldots, n\}$ and let $O_i := V \setminus P_i$ for each i. Let $\mathscr{F} = (V, \theta, O_i)$ and let $s = \varnothing$. Let $\widehat{K}_i\varphi := \neg K_i \neg \varphi$. We now have the following equivalences:

$$
\begin{array}{lcll}
\psi \text{ is QBF-true} & \iff & \vDash \forall P_1 \exists P_2 \ldots \forall P_{n-1} \exists P_n \varphi & \text{QBF truth definition} \\
& \iff & s \vDash \forall P_1 \exists P_2 \ldots \forall P_{n-1} \exists P_n \varphi & \text{no free variables} \\
& \iff & s \vDash \forall (V \setminus O_1) \exists (V \setminus O_2) \ldots \forall (V \setminus O_{n-1}) \exists (V \setminus O_n) \varphi & \text{by definition of } O_i \\
& \iff & \mathscr{F}, s \vDash K_1 \widehat{K}_2 \ldots K_{n-1} \widehat{K}_n \varphi & \text{by Definition 2.5}
\end{array}
$$

The last formula is of the same length as the given QBF. Hence we have a polynomial reduction from QBF-truth to model checking knowledge structures. □

In short, it is hard to do model-checking in polynomial time on symbolic structures since we need to go over every possible valuation on V while evaluating $K_i\psi$, unlike in Kripke models where these valuations are enumerated explicitly in the model and thus part of the input size.

Given that $\mathscr{L}_{\mathsf{EL}}$ is a fragment of $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ we have the following corollary.

Corollary 2.9. *Model checking $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ on knowledge structures is PSPACE-hard.*

Having shown hardness, we now turn to membership. For simplicity we focus on PAL in this section. We will define an algorithm and then show its correctness and its memory usage. Note that to stay in PSPACE we cannot use the common PAL reduction axioms (see [13, Def 4.53]) as they would blow up the size of formulas by making copies of the announced formulas. For example, consider $[!K_ip]K_ip$. The reduction axiom would give us the following equivalences:

$$[![!K_ip]K_ip]K_ip \equiv ([!K_ip]K_ip) \to K_i[![!K_ip]K_ip]p \equiv ((K_ip) \to K_i[!K_ip]p) \to K_i[!((K_ip) \to K_i[!K_ip]p)]p$$

Here the last formula which would then be model checked is huge in size with respect to the initial $[![!K_ip]K_ip]K_ip$. In general, rewriting $[![![!\ldots[![!K_ip]]K_ip]K_ip]\ldots]K_ip$ leads to a formula with 2^n copies of K_ip and thus needing exponential space in terms of the original formula length.

Our Algorithm 1 check takes as inputs a knowledge structure $\mathscr{F}$, a PAL formula φ and a list of formulas L. For the model checking problem L is initially empty and populated by the algorithm itself.

Algorithm 1: check

Input: knowledge structure $\mathscr{F} = (V, \theta, O)$, list of formulas $L = [\ell_0, \ldots, \ell_k]$, state $s \subseteq V$ of $\mathscr{F} \times \ell_0 \times \cdots \times \ell_k$, formula φ

Precondition: s is a state of $\mathscr{F}$ (i.e. $s \vDash \theta$) and L can be announced on $(\mathscr{F}, s)$.

Output: true or false

```
1  switch φ do
2      case p do return p ∈ s ;
3      case ¬φ do return NOT check(F, L, s, φ) ;
4      case φ1 ∧ φ2 do return (check(F, L, s, φ1) AND check(F, L, s, φ2)) ;
5      case K_i φ1 do
6          foreach t ⊆ V do
7              if t ⊨ θ AND t ∩ O_i = s ∩ O_i then
8                  stillExists := true // did t ''survive'' the announcements in L?
9                  foreach j ∈ [0, ..., k] do
10                     if NOT check(F, [ℓ_0, ..., ℓ_{j-1}], t, ℓ_j) then stillExists := false ;
11                 if stillExists AND NOT check(F, [ℓ_0, ... ℓ_k], t, φ1) then
12                     return false
13         return true
14     case [!ψ]φ1 do
15         if check(F, L, s, ψ) then return check(F, L++[ψ], s, φ1) else return true ;
```

The algorithm proceeds by recursion on the formula φ. There are two challenges we need to take care of so that the algorithm takes only polynomial space:

- Firstly, we cannot compute the BDDs of announcements as in Definition 2.6, because short formulas may have BDDs of exponential size. Hence we track announcements in the list L.
- When evaluating formulas of the form $K_i\varphi$, we must not compute the exponentially large set of all states. But we can iterate over them instead, without ever storing the full set.

To make both solutions compatible with each other, when evaluating $K_i\varphi$ we check in Line 9 whether a state still exists after the sequence of announcements in L, using recursive calls. This technique is similar

to the model checking algorithm for succinct DEL in [1]. It is also similar to the context-dependent semantics in [27], where the announcements are stored as a list of formulas and the model update is not triggered until $K_i\varphi$ formulas are evaluated. We now formally state that check is correct.

Lemma 2.10. *Given a pointed knowledge structure $\mathscr{F}$, a state s of $\mathscr{F}$, a list of* PAL*-S5 formulas $L = [\ell_0, \ldots, \ell_k]$ and a* PAL*-S5 formula φ, we have $\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash \varphi$ iff* check$(\mathscr{F}, L, s, \varphi)$ *returns* true*.*

Proof sketch. By induction on the size of inputs φ, $\mathscr{F}$ and L. The difficult case is for $K_i\varphi_1$. It iterates over all valuations (line 6) checking whether they agree on what i observes (line 7) and recursively verifies φ_1 as per Definition 2.5. The survival of the valuation by the sequence of announcements L is checked in line 9. For a detailed proof, see the appendix.

Next comes our main result for this subsection.

Theorem 2.11. *Model checking $\mathscr{L}_{\mathsf{PAL}}, \mathscr{L}_{\mathsf{EL}}$ on knowledge structures is PSPACE-complete.*

Proof. Hardness follows from Corollary 2.9. By Lemma 2.10 it remains to show that algorithm 1 only takes space polynomial in the size of the input. Membership for $\mathscr{L}_{\mathsf{EL}}$ follows as a special case of $\mathscr{L}_{\mathsf{PAL}}$. At any instance of a call to check$(\mathscr{F}, L, s, \varphi)$, exactly one switch case matches. Every recursive call taken in check$(\mathscr{F}, L, s, \varphi)$ is of size no more than the size of the current input. Moreover, in the call stack, there are at most linear many recursive calls left to evaluate at any given instance. In each call, the space used is polynomial. The most intensive is the case for $K_i\varphi_1$. We iterate over every valuation $t \subseteq V$, which needs memory in the size of $|V|$. Note that we never store the whole list of states nor results. □

We note that the model checking algorithm performing well *in practice* (and used in SMCDEL [19]) is *not* in PSPACE. To see this, take a Boolean formula for which the BDD has to be large, for example consider the vocabulary $V = \{p_1, \ldots p_{2n}\}$ and the formula $\varphi := (p_1 \wedge p_{n+1}) \vee \ldots \vee (p_n \wedge p_{n+n})$ which has length in $\mathscr{O}(n)$. Suppose then we model check $[!\varphi]p$ on a knowledge structure $\mathscr{F}$ with state law $\theta = \top$ using the SMCDEL algorithm. This means we update the state law (stored as a BDD) from $\top$ to $\theta \wedge \|\varphi\|_{\mathscr{F}}$. As discussed in [9, p. 681] the BDD of this formula needs 2^{n+1} many nodes. Hence the resulting knowledge structure needs exponential space and this is not possible in PSPACE.

In summary, there is a trade-off between time and space here: Algorithm 1 will need exponential time to evaluate K_i, whereas the BDD-based method may need exponential space for $[!\varphi]$.

2.2 Symbolic Model Checking DEL-K

We now generalize the setting from the previous section in two aspects. First, we switch from S5 to K by using belief structures instead of knowledge structures. Second, we move from public announcements to general events. We already discussed that DEL can be seen as an extension of PAL in the introduction. We now illustrate how the move from S5 to K, and the additional feature of factual change are dealt with symbolically. The difference between S5 and K is captured by $O_i \subseteq V$ and $\Omega_i \in \mathscr{L}_B(V \cup V^+)$: O_i always encodes an equivalence relation for a hard notion of knowledge, while Ω_i can encode arbitrary relations to describe belief. Factual change is captured by V_- and θ_- in the transformers from Definition 2.3, describing which propositions are changed and how [17, Section 2.8].

Theorem 2.12. *Model checking $\mathscr{L}_{\mathsf{EL}}$ on belief structures is PSPACE-hard.*

Proof. Similar to the proof of Theorem 2.8, but instead of O_i let Ω_i be the BDD of $\bigwedge_{p \in P_i}(p \leftrightarrow p')$. This BDD has size linear in $|V|$. Hence we reduce QBF truth to model checking $\mathscr{L}_{\mathsf{EL}}$ on belief structures. □

Again, given that $\mathscr{L}_{\mathsf{EL}}$ is a fragment of $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ we have the following corollary.

Corollary 2.13. *Model checking $\mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ on belief structures is PSPACE-hard.*

Algorithm 2: checkDELK

Input : Knowledge structure $\mathscr{F} = (V, \theta, \Omega)$, list of events $L = [(\mathscr{X}_0, x_0), \ldots, (\mathscr{X}_k, x_k)]$, state $s \subseteq V$, formula $\varphi \in \mathscr{L}_{DEL}(V)$
Precondition: s is a state of $\mathscr{F}$ (i.e. $s \vDash \theta$) and L can be executed on $(\mathscr{F}, s)$.
Output: true or false

1 **switch** φ **do**
2 **case** p **do**
3 **if** $L = []$ **then**
4 **return** $p \in s$
   `// Compute the final state`
5 **foreach** $j \in [0, \ldots, k]$ **do**
6 $s := (s \setminus V_{j,-}) \cup \{q \in V_{j,-} \mid \mathsf{checkDELK}(\mathscr{F}, [(\mathscr{X}, x_0), \ldots, (\mathscr{X}, x_j)], s, [x_j^+ \sqsubseteq V_j^+]\theta_j^-(q))\}$
7 **return** $p \in s$
8 **case** $\neg\varphi$ **do**
9 **return** NOT $\mathsf{checkDELK}(\mathscr{F}, L, s, \varphi)$
10 **case** $\varphi_1 \wedge \varphi_2$ **do**
11 **return** ($\mathsf{checkDELK}(\mathscr{F}, L, s, \varphi_1)$ AND $\mathsf{checkDELK}(\mathscr{F}, L, s, \varphi_2)$)
12 **case** $K_i\psi$ **do**
13 **foreach** $t \subseteq V$, $t_0^+ \subseteq V_0^+, \ldots, t_k^+ \subseteq V_k^+$ **do**
14 **if** $t \vDash \theta$ **then**
15 **if** $s \cup t' \vDash \Omega_i$ *AND* $x_1^+ \cup t_1^{+\prime} \vDash \Omega_{0,i}^+$ *AND* $\ldots$, $x_k^+ \cup t_k^{+\prime} \vDash \Omega_{k,i}^+$ **then**
16 `accessible` := true `// did` t `"survive" the actions in` L`?`
17 **foreach** $j \in [0, \ldots, k]$ **do**
18 **if** *not* $\mathsf{checkDELK}(\mathscr{F}, [(\mathscr{X}_0, t_0^+), \ldots, (\mathscr{X}_{j-1}, t_{j-1}^+)], t, [t_j^+ \sqsubseteq V_j^+]\theta_j^+)$ **then**
19 `accessible` := false
20 **if** *accessible* **then**
21 **if** *not* $\mathsf{checkDELK}(\mathscr{F}, [(\mathscr{X}_0, t_0^+), \ldots, (\mathscr{X}_k, t_k^+)], t, \psi)$ **then**
22 **return** false
23 **return** true
24 **case** $[\mathscr{X}, x]\psi$ **do**
25 **if** $\mathsf{checkDELK}(\mathscr{F}, L, s, [x \sqsubseteq V^+]\theta^+)$ `// checking the precondition`
26 **then**
27 **return** ($\mathsf{checkDELK}(\mathscr{F}, L{+}{+}[(\mathscr{X}, x)], s, \psi)$)
28 **return** true

Similar to the previous section about PAL-S5, we prove membership by defining an algorithm. Also here the challenge is that we cannot compute the potentially too large BDD of the new state law after updating. Fortunately, also a similar solution works: the list L will now not just contain a list of public announcements, but a list of transformers. Then to check $\mathscr{F}, s \vDash [\mathscr{X}, x]\varphi$ we add that transformer to L.

Later, when we check $K_i\varphi$, we iterate "on the fly" over the states t that may be the result of the sequence of transformers, including the additional atoms introduced and taking into account the factual change.

Lemma 2.14. *Given a pointed knowledge structure $\mathscr{F} = \langle V, \theta, O_i \rangle, s \subseteq V$ and a* DEL *formula φ, we have $\mathscr{F}, s \vDash \varphi$ iff* checkDELK$(\mathscr{F}, [], s, \varphi)$ *returns* true.

Proof sketch. Similar to Lemma 2.10, see appendix for the full proof.

Theorem 2.15. checkDELK$(\mathscr{F}, [], s, \varphi)$ *takes at most polynomial space with respect to size of input.*

Proof. An argument analogous to that for Theorem 2.11 shows that algorithm 2 runs using space at most polynomial in size of the input. □

Together we now get our main result for DEL-K and as a result PAL-K and EL-K as well.

Theorem 2.16. *Model checking $\mathscr{L}_{\mathsf{EL}}, \mathscr{L}_{\mathsf{PAL}}$ and $\mathscr{L}_{\mathsf{DEL}}$ on belief structures is PSPACE-complete.*

Proof. Using algorithm 2, which is correct by Lemma 2.14 it is correct and uses polynomial space by Theorem 2.15, we have membership. By Corollary 2.13, we have hardness. □

Moreover, since DEL-S5 is a special case of DEL-K we have the following.

Theorem 2.17. *Model checking $\mathscr{L}_{\mathsf{DEL}}$ on knowledge structures is PSPACE-complete.*

Proof sketch. Just like in Theorem 2.12, for membership use the equivalent $\Omega_i = \bigwedge_{p \in O_i} p \leftrightarrow p'$ (for the belief structure) and $\Omega_i^+ = \bigwedge_{p^+ \in O_i^+} p^+ \leftrightarrow p^{+\prime}$ (for any transformer). The resulting belief structure is at most twice the size of the given knowledge structures. The hardness result is from Theorem 2.8.

3 Succinct Models with Mental Programs

An alternative to the symbolic structures used in the previous section is the framework of *Succinct* DEL as presented by [12]. It is based on a version of (Propositional Dynamic Logic) PDL, also called Dynamic Logic of Propositional Assignment (DLPA) [3].

Definition 3.1. *The language of* mental programs *over a vocabulary V is defined by*

$$\pi ::= p \leftarrow \top \mid p \leftarrow \bot \mid \beta? \mid \pi \cup \pi \mid \pi;\pi \mid \pi \cap \pi$$

where $p \in V$ and β is a Boolean formula over V. Let Π_V denote the set of all mental programs over V. The length of a mental program is defined as follows.

$$\begin{array}{llllllllll} |p \leftarrow \top| & := & 1 & |\beta?| & := & |\beta| & |\pi_1;\pi_2| & := & |\pi_1| + |\pi_2| \\ |p \leftarrow \bot| & := & 1 & |\pi_1 \cup \pi_2| & := & |\pi_1| + |\pi_2| & |\pi_1 \cap \pi_2| & := & |\pi_1| + |\pi_2| \end{array}$$

In the rest of the article we are mostly interested in how a single relation over the set of states $\mathscr{P}(V)$ can be encoded. Hence we omit further details about succinct DEL and refer to [12] for how Kripke models can be encoded using one mental program for each agent, and how actions and events can be encoded using mental programs as well.

Definition 3.2. *Two states $s,t \subseteq V$ are related by a mental program π (written as $s \xrightarrow{\pi} t$) as follows:*

$$\begin{array}{ll} s \xrightarrow{p \leftarrow \top} t & :\Longleftrightarrow\ t = s \cup \{p\} \\ s \xrightarrow{p \leftarrow \bot} t & :\Longleftrightarrow\ t = s \setminus \{p\} \\ s \xrightarrow{\beta?} t & :\Longleftrightarrow\ s = t \text{ and } s \vDash \beta \end{array} \qquad \begin{array}{ll} s \xrightarrow{\pi_1 \cup \pi_2} t & :\Longleftrightarrow\ s \xrightarrow{\pi_1} t \text{ or } s \xrightarrow{\pi_2} t \\ s \xrightarrow{\pi_1;\pi_2} t & :\Longleftrightarrow\ \exists u \subseteq V :\ s \xrightarrow{\pi_1} u \text{ and } u \xrightarrow{\pi_2} t \\ s \xrightarrow{\pi_1 \cap \pi_2} t & :\Longleftrightarrow\ s \xrightarrow{\pi_1} t \text{ and } s \xrightarrow{\pi_2} t \end{array}$$

We define $R_\pi := \{(s,t) \in V \times V \mid s \xrightarrow{\pi} t\}$.

Example 3.3. *From the state $\{p,q\}$ we can reach the state $\{p\}$ using the mental program $q \leftarrow \bot$. From the state $\{p,q\}$ we can reach the states $\{p\}$ and $\{q\}$ using the mental program $(p \leftarrow \bot) \cup (q \leftarrow \bot)$.*

Other versions of mental programs [12] also include the general assignment $p \leftarrow \beta$ and inverse π^{-1}. However, these operators do not add expressivity, as shown in [18]. Also the following shows that mental programs as defined above are complete in the sense that they can encode all relations.

Definition 3.4. *For any $x \subseteq y \subseteq V$, let $\mathsf{of}(x,y) := \bigwedge_{p\in x} p \wedge \bigwedge_{p\in y\setminus x} \neg p$. For any set $x \subseteq V$ we define $\mathsf{change}(x) := ;_{p\in x}((p \leftarrow \top) \cup (p \leftarrow \bot))$, and for any state $s \subseteq V$ we define $\mathsf{goto}(s,V) := (;_{p\in s}(p \leftarrow \top))\ ;(;_{p\in V\setminus x}(p \leftarrow \bot))$.*

We illustrate Definition 3.4 with some examples. First, $x = \{p\}$ and $y = \{p,q,r\}$ then we have $\mathsf{of}(x,y) = p \wedge \neg q \wedge \neg r$. Second, if $x = \{q,r\}$ then $\mathsf{change}(x) = (q \leftarrow \top \cup q \leftarrow \bot);(r \leftarrow \top \cup r \leftarrow \bot)$. Third, if $V = \{p,q,r\}$ then $\mathsf{goto}(\{p\},V) = p \leftarrow \top; q \leftarrow \bot; r \leftarrow \bot$.

Lemma 3.5. *For any relation $R \subseteq V \times V$ there is a mental program π such that $R_\pi = R$.*

Proof. Let $\pi := \bigcup_{(x,y)\in R}(\mathsf{of}(x,V)?;\mathsf{goto}(y))$ and apply Definition 3.2. □

In [12] it is shown how Kripke models can be encoded with mental programs and that model checking DEL on such *succinct* models is in PSPACE. This is shown not by giving a PSPACE algorithm, but an alternating Turing machine algorithm and then using the fact that PSPACE = APTIME [11]. This means the algorithm given cannot easily be translated to an actual implementation. The authors suggest that "a model checking procedure for our succinct language may use BDD techniques of [6]" [12, p. 130].

4 Comparison and Translations

To compare BDDs and mental programs, we present a list of examples in Table 2. Each row stands for a class of relations or an operation on relations, and shows how it can be represented or executed, in the different representations. Notably, observing more propositions, and thus going from the total relation (in row 2) to the identity (in row 5), leads to a shorter mental program, but to a longer Boolean formula.

	Mental program π	Boolean function Ω	Observable O
empty relation	$?\bot$	$\bot$	n/a
total relation	$\mathsf{change}(V)$	$\top$	$\varnothing$
only observe p	$\mathsf{change}(V \setminus \{p\})$	$p \leftrightarrow p'$	$\{p\}$
only observe $\{p,q\}$	$\mathsf{change}(V \setminus \{p,q\})$	$(p \leftrightarrow p') \wedge (q \leftrightarrow q')$	$\{p,q\}$
identity relation	$?\top$	$\bigwedge_{p\in V}(p \leftrightarrow p')$	V
single edge $s \to t$	$?\mathsf{of}(s,V);\mathsf{change}(V);\mathsf{of}(t,V)$	$\mathsf{of}(s,V) \wedge \mathsf{of}(t',V')$	n/a
complement	n/a	Given β, use $\neg\beta$	n/a
inverse	See translation in [18].	Given β, swap p with p' etc.	no change
composition	Given π_1 and π_2, use $\pi_1;\pi_2$	See Def. 4.1	n/a
intersection	Given π_1 and π_2, use $\pi_1 \cap \pi_2$	Given β_1 and β_2, use $\beta_1 \wedge \beta_2$	union

Table 2: Examples of relations and operations.

The "n/a" entries in Table 2 mean that this operation cannot be defined (in general) with the given representation. For example, given a mental program we cannot easily define one for the complement (Row 7). All mentioned operations can be done with Boolean functions, but with observational variables (last column) three operations are impossible because they yield non-equivalence relations.

4.1 Translating Mental Programs to BDDs

Given a mental program we now want to encode the same relation over subsets of V using a Boolean function. We give the following definition using Boolean formulas on the right side, but it also provides a mapping of mental programs to BDDs by reading all connectives on the right side as BDD operations. Recall that Π_V is the set of mental programs over V and $\mathscr{L}_B(V \cup V')$ is the set of Boolean formulas over the double vocabulary where the prime makes a fresh copy of each atomic proposition letter.

Definition 4.1. *We now define a function* $\mathsf{tr} : \Pi_V \to \mathscr{L}_B(V \cup V')$.

$$\begin{array}{ll} \mathsf{tr}(p \leftarrow \top) := p' \wedge \bigwedge_{p \neq q \in V}(q \leftrightarrow q') & \mathsf{tr}(\pi_1 \cup \pi_2) := \mathsf{tr}(\pi_1) \vee \mathsf{tr}(\pi_2) \\ \mathsf{tr}(p \leftarrow \bot) := \neg p' \wedge \bigwedge_{p \neq q \in V}(q \leftrightarrow q') & \mathsf{tr}(\pi_1 ; \pi_2) := [V'' \mapsto V'](\exists V'(\mathsf{tr}(\pi_1) \wedge \mathsf{tr}(\pi_2)')) \\ \mathsf{tr}(\beta ?) := \beta \wedge \bigwedge_{p \in V}(p \leftrightarrow p') & \mathsf{tr}(\pi_1 \cap \pi_2) := \mathsf{tr}(\pi_1) \wedge \mathsf{tr}(\pi_2) \end{array}$$

The left cases of Definition 4.1 are easy, they mostly consist of the "do not change anything else" conjunctions. Among the right three cases, the one standing out is the composition $\pi_1 ; \pi_2$. Here we write $[\cdot \mapsto \cdot]\varphi$ for simultaneous substitution of atoms in φ. For example, $[\{p', q'\} \mapsto \{p, q\}](p' \wedge q'') = (p \wedge q'')$. Strictly speaking in $[A \mapsto B]\varphi$ both A and B are ordered lists and we use the implicit bijection between them as the substitution function [17, Def. 1.0.3]. We note that $\mathsf{tr}(\pi_2)'$ is $[V \mapsto V'][V' \mapsto V'']\mathsf{tr}(\pi_2)$, i.e. this changes the formula from the vocabulary $V \cup V'$ to the vocabulary $V' \cup V''$. The Boolean quantification $\exists V'$ then eliminates all single-primed variables and lastly the outermost substitution ensures that the resulting formula is over the vocabulary $V \cup V'$ as we want. We stress that $\exists$ is only expensive when done syntactically. As a BDD operation it in fact deletes variables and nodes from a BDD [9].

Theorem 4.2. *The translation from Def. 4.1 is correct: for any* $s, t \subseteq V$, *we have* $s \xrightarrow{\pi} t$ *iff* $(s \cup t') \vDash \mathsf{tr}(\pi)$.

Proof sketch. By induction on the structure of π and applying Definition 3.2. See appendix for proof.

Note that already translating $p \leftarrow \top$ needs $\mathscr{O}(|V|)$ many nodes in the BDD. This illustrates the high length of formulas resulting from this translation. However, it is not a problem specific to the particular translation given above, but applies to any correct translation, as the following theorem states.

Theorem 4.3. *For any translation* tr' *from mental programs to BDDs there exists a mental program* π *such that* $\mathsf{tr}'(\pi)$ *has size exponential in* $|\pi|$.

Proof sketch. Take any vocabulary $V = \{p_1, \ldots p_{2n}\}$ with this ordering fixed. Let $\pi := \beta ?$ where β is the Boolean formula $(p_1 \wedge p_{n+1}) \vee \ldots \vee (p_n \wedge p_{n+n})$ from page 239. We show that the BDD corresponding to the mental program π has at least 2^{n+1} many nodes. For the details, see the appendix.

4.2 Translating BDDs to Mental Programs

We write $(t_0 \dashleftarrow (p_i) \rightarrow t_1)$ to denote a node in a BDD that is labelled with variable p and has an else-edge (dotted) pointing to node t_0 and a then-edge (solid) pointing to node t_1. Note that p_i may come from V or from V', and in the latter case we write p'_i. For the leaves, we just write $\top$ and $\bot$.

Definition 4.4. *We define a function* $\tau : BDD_{(V \cup V')} \times [V \cup V'] \to \Pi_V$ *where* $[V \cup V']$ *is the set of all lists with elements from* $V \cup V'$. *We distinguish different cases for the given BDD* Ω.

$$\tau(\bot, L) := ?\bot \qquad \tau(\top, []) := ?\top \qquad \tau(\top, [p_k, .., p_n]) := ((p_k \leftarrow \bot) \cup (p_k \leftarrow \top)) ; \tau(\top, [p_{k+1}, .., p_n])$$

$$\tau(\Omega = (t_0 \dashleftarrow (p_i) \rightarrow t_1), [p_k, .., p_n]) := \begin{cases} (?\neg p_k ; \tau(t_0, [p_k, .., p_n])) \cup (?p_k ; \tau(t_1, [p_k, .., p_n])) & \text{if } i = k \\ ((p_k \leftarrow \bot) \cup (p_k \leftarrow \top)) ; \tau(\Omega, [p_{k+1}, .., p_n]) & \text{otherwise} \end{cases}$$

$$\tau(\Omega = (t_0 \dashleftarrow (p'_i) \rightarrow t_1), [p_k, .., p_n]) := \begin{cases} (p_k \leftarrow \bot ; \tau(t_0, [p_{k+1}, .., p_n])) \\ \cup (p_k \leftarrow \top ; \tau(t_1, [p_{k+1}, .., p_n])) & \text{if } i = k \\ ((p_k \leftarrow \bot) \cup (p_k \leftarrow \top)) ; \tau(\Omega, [p_{k+1}, .., p_n]) & \text{otherwise} \end{cases}$$

Lastly, define $\tau_0 : BDD_{V \cup V'} \to \Pi_V$ *by* $\tau_0(\Omega) := \tau(\Omega, [p_0, \ldots, p_n])$.

Both "otherwise" cases above are about variables $i > k$ not mentioned by BDD. If the BDD does not mention a unprimed variable, this simply means the relation does not depend on that variable in the starting state, so the mental program also does not have to mention it. On the other hand, not mentioning a primed variable in the BDD means we must allow the mental program to change that variable arbitrarily.

Theorem 4.5. *The translation τ always terminates.*

Proof. We observe that at each recursive step either the size of the BDD Ω or the length of the list L strictly decreases, while the other size stays the same. □

Theorem 4.6. *The translation τ_o from Definition 4.4 is correct. That is, given a vocabulary V, for any $s,t \subseteq V$, we have $s \xrightarrow{\tau_0(\Omega)} t$ iff $(s \cup t') \vDash \Omega$.*

Proof sketch. For any BDD Ω, let Ω^* be its unraveling to a tree. Note that we have $\tau(\Omega) = \tau(\Omega^*)$. Hence for the proof we assume wlog. that β is a tree and proceed by induction over the tree structure. For all details, see the appendix.

We now consider the length of the mental programs resulting from the translation. Unfortunately, given a BDD for a vocabulary V with $|V| = n$ the output of τ will always have at least length $\mathcal{O}(2^n)$. The result may be simplified using equivalences such as $p \leftarrow \top; p? \equiv p \leftarrow \top$ and $p \leftarrow \bot; p? \equiv \bot?$. But even including such simplifications we believe that there is no better translation in the following sense.

Conjecture 4.7. *For any translation τ' from BDDs to mental programs that is correct in the sense of Theorem 4.6, there exists a BDD β such that $\tau'(\beta)$ has a length exponential in the number of nodes of β.*

As an example, take the BDD encoding the relation $R = \{(s,s) \mid |s| = \lceil \frac{|V|}{2} \rceil\}$, i.e. the identity restricted to the states where exactly half of the vocabulary is true. The BDD encoding R is a $|V| \cdot |V|$ grid. We can find a mental program encoding this relation using (the proof of) Lemma 3.5, but the size of the mental program is $2^{|V|/2} \cdot |V|$. It is not clear whether there exists a shorter mental program that encodes R. Essentially the mental program needs to count, and at the same time preserve the valuation. But as there are no additional variables the only way to count without forgetting is to try all valuations. Thus we believe that any mental program π encoding R must have size exponential in the size of the *BDD*.

5 Conclusions and Future Work

In this paper, we showed that the symbolic model-checking tasks for EL, PAL and DEL on both knowledge structures and belief structures are all PSPACE-complete. In addition we compared how the same relations can be encoded in belief structures with BDDs on one hand, and in succinct models with mental programs (a regular language like syntax) on the other hand. That is, we provided translations from mental programs to BDDs, and back. We have also proven that any such translation from mental programs to BDDs will lead to an exponential blowup in size, and we conjecture the same for the other direction from BDDs to mental programs.

An aspect we did not consider here but that is relevant for epistemic planning are multi-pointed models and actions. Concerning implementations, a PAL model checker using mental programs was implemented in [21]. In parallel to the theoretical work presented here we have improved the performance of this code and implemented the translations from section 4. We plan to benchmark all methods and eventually merge them into SMCDEL in the future.

References

[1] Guillaume Aucher & François Schwarzentruber (2013): *On the Complexity of Dynamic Epistemic Logic*. In Burkhard C. Schipper, editor: *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2013), Chennai, India, January 7-9, 2013*, pp. 19–28. Available at `http://www.tark.org/proceedings/tark_jan7_13/p19-aucher.pdf`.

[2] Philippe Balbiani, Andreas Herzig, François Schwarzentruber & Nicolas Troquard (2014): *DL-PA and DCL-PC: model checking and satisfiability problem are indeed in PSPACE*. *CoRR* abs/1411.7825. arXiv:1411.7825.

[3] Philippe Balbiani, Andreas Herzig & Nicolas Troquard (2013): *Dynamic Logic of Propositional Assignments: A Well-Behaved Variant of PDL*. In: *28th Annual ACM/IEEE Symposium on Logic in Computer Science*, pp. 143–152, doi:10.1109/LICS.2013.20.

[4] Alexandru Baltag, Lawrence S. Moss & Slawomir Solecki (1998): *The logic of public announcements, common knowledge, and private suspicions*. In Itzhak Gilboa, editor: *Theoretical Aspects of Rationality and Knowledge (TARK 1998)*, TARK 1998, pp. 43–56. Available at `https://dl.acm.org/citation.cfm?id=645876.671885`.

[5] Vaishak Belle, Thomas Bolander, Andreas Herzig & Bernhard Nebel (2023): *Epistemic planning: Perspectives on the special issue*. *Artif. Intell.* 316, p. 103842, doi:10.1016/J.ARTINT.2022.103842.

[6] Johan van Benthem, Jan van Eijck, Malvin Gattinger & Kaile Su (2015): *Symbolic Model Checking for Dynamic Epistemic Logic*. In Wiebe van der Hoek, Wesley H. Holliday & Wen-fang Wang, editors: *Proceedings of the 5th International Workshop on Logic, Rationality, and Interaction (LORI 2015)*, pp. 366–378, doi:10.1007/978-3-662-48561-3_30.

[7] Johan van Benthem, Jan van Eijck, Malvin Gattinger & Kaile Su (2018): *Symbolic Model Checking for Dynamic Epistemic Logic – S5 and Beyond*. *Journal of Logic and Computation (JLC)* 28(2), pp. 367–402, doi:10.1093/logcom/exx038.

[8] Patrick Blackburn, Maarten de Rijke & Yde Venema (2001): *Modal Logic*. *Cambridge Tracts in Theoretical Computer Science* 53, Cambridge University Press, doi:10.1017/CBO9781107050884.

[9] Randal E. Bryant (1986): *Graph-Based Algorithms for Boolean Function Manipulation*. *IEEE Transactions on Computers* C-35(8), pp. 677–691, doi:10.1109/TC.1986.1676819.

[10] Sagar Chaki & Arie Gurfinkel (2018): *BDD-Based Symbolic Model Checking*. In Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith & Roderick Bloem, editors: *Handbook of Model Checking*, Springer Nature, pp. 219–245, doi:10.1007/978-3-319-10575-8_8.

[11] Ashok K. Chandra, Dexter C. Kozen & Larry J. Stockmeyer (1981): *Alternation*. *Journal of the ACM* 28(1), pp. 114—-133, doi:10.1145/322234.322243.

[12] Tristan Charrier & François Schwarzentruber (2017): *A Succinct Language for Dynamic Epistemic Logic*. In S. Das, E. Durfee, K. Larson & M. Winikoff, editors: *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, AAMAS '17, International Foundation for Autonomous Agents and Multiagent Systems, pp. 123–131. Available at `https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p123.pdf`.

[13] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. Springer, doi:10.1007/978-1-4020-5839-4.

[14] Hans van Ditmarsch, Wiebe van der Hoek, Ron van der Meyden & Ji Ruan (2006): *Model Checking Russian Cards*. *Electronic Notes in Theoretical Computer Science* 149(2), pp. 105–123, doi:10.1016/j.entcs.2005.07.029.

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

[16] Michael J. Fischer & Richard E. Ladner (1979): *Propositional dynamic logic of regular programs*. *Journal of Computer and System Sciences* 18(2), pp. 194–211, doi:10.1016/0022-0000(79)90046-1.

[17] Malvin Gattinger (2018): *New Directions in Model Checking Dynamic Epistemic Logic*. Ph.D. thesis, University of Amsterdam. Available at `https://malv.in/phdthesis`.

[18] Malvin Gattinger (2020): *Towards Symbolic and Succinct Perspective Shifts*, doi:10.5281/zenodo.4767546. Presentation at the EpiP workshop at ICAPS 2020.

[19] Malvin Gattinger, Stefano Volpe, Bas Laarakker & Daniel Miedema (2024): *SMCDEL — A symbolic model checker for Dynamic Epistemic Logic*, doi:10.5281/zenodo.11061438. Available at `https://github.com/jrclogic/SMCDEL`. Version 1.3.0.

[20] Ronald de Haan & Iris van de Pol (2021): *On the Computational Complexity of Model Checking for Dynamic Epistemic Logic with S5 Models*. Journal of Applied Logics 8(3), pp. 621–658. Available at `http://www.collegepublications.co.uk/downloads/ifcolog00045.pdf#page=10`.

[21] Maickel Hartlief (2020): *Making Model Checking Scalable: Implementing Succinct Kripke Models for Public Announcement Logic*. Available at `https://fse.studenttheses.ub.rug.nl/23607/`. BSc thesis, Groningen. Code avaliable at `https://github.com/Maickel-Hartlief/SucExpModelCheckers`.

[22] Esteban Landerreche & David Fernández-Duque (2017): *A case study in almost-perfect security for unconditionally secure communication*. Designs, Codes and Cryptography 83(1), pp. 145–168, doi:10.1007/s10623-016-0210-y.

[23] 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.

[24] Daniel Miedema & Malvin Gattinger (2023): *Exploiting Asymmetry in Logic Puzzles: Using ZDDs for Symbolic Model Checking Dynamic Epistemic Logic*. In Rineke Verbrugge, editor: *Theoretical Aspects of Rationality and Knowledge 2023 (TARK 2023)*, pp. 407—-420, doi:10.4204/EPTCS.379.32.

[25] Jan Plaza (2007): *Logics of public communications*. Synthese 158(2), pp. 165–179, doi:10.1007/S11229-007-9168-7.

[26] Weizhi Tang & Vaishak Belle (2024): *ToM-LM: Delegating Theory of Mind Reasoning to External Symbolic Executors in Large Language Models*. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jimenez-Ruiz, Roberto Confalonieri, Pranava Madhyastha & Benedikt Wagner, editors: *Neural-Symbolic Learning and Reasoning*, Springer Nature Switzerland, pp. 245–257, doi:10.1007/978-3-031-71170-1_20.

[27] Yanjing Wang & Qinxiang Cao (2013): *On axiomatizations of public announcement logic*. Synthese 190, pp. 103–134, doi:10.1007/s11229-012-0233-5.

A Proofs

Lemma 2.10. *Given a pointed knowledge structure $\mathscr{F}$, a state s of $\mathscr{F}$, a list of* PAL-*S5 formulas* $L = [\ell_0, \ldots, \ell_k]$ *and a* PAL-*S5 formula* φ*, we have* $\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash \varphi$ *iff* $\mathsf{check}(\mathscr{F}, L, s, \varphi)$ *returns* true.

Proof. By induction on the input size and case distinction on φ. We only show the interesting cases for knowledge and for announcements.

Case $\varphi = K_i\varphi_1$. We have:

$$\begin{aligned}\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash K_i\varphi_1 &\text{ iff } \forall t \subseteq V : s \cap O_i = t \cap O_i \text{ and } t \text{ survives in } \mathscr{F}^{\ell_0 \ldots \ell_k} \text{ implies } \mathscr{F}^{\ell_0 \ldots \ell_k}, t \vDash \varphi_1 \\ &\text{iff } \forall t \subseteq V : s \cap O_i = t \cap O_i \text{ and } \bigwedge_{i=0}^{k} \mathscr{F}^{\ell_0 \ldots \ell_{i-1}}, t \vDash \ell_i \text{ implies } \mathscr{F}^{\ell_0 \ldots \ell_k}, t \vDash \varphi_1\end{aligned}$$

Note that, assuming for a $t \subseteq V$, $s \cap O_i = t \cap O_i$, $\bigwedge_{i=0}^{k} \mathscr{F}^{\ell_0 \ldots \ell_{i-1}}, t \vDash \ell_i$ if and only if `stillExists`$=$ true after the execution of the loop in line 9. This is so by using IH as for any $i \in [1, |L| - 1]$, $|\mathscr{F}| + |\ell_0| + \ldots + \ell_{i-1} + \ell_i < |\mathscr{F}| + \sum_{j=0}^{k} |\ell_j| + |\varphi_1| + 1$. Line 7 checks whether $s \cap O_i = t \cap O_i$. Hence

$$\begin{aligned}&\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash K_i\varphi_1 \\ &\text{iff } \forall t \subseteq V : \text{line 7 and } \texttt{stillExists} \text{ is true after loop 9 implies } \mathscr{F}^{\ell_0 \ldots \ell_k}, t \vDash \varphi_1 \\ &\text{iff } \forall t \subseteq V : \text{line 7 and } \texttt{stillExists} \text{ is true after loop 9 implies } \mathsf{check}(\mathscr{F}[\ell_0 \ldots \ell_k], t, \varphi_1) \neq \mathsf{false}\end{aligned}$$

Case $\varphi = [!\psi]\varphi_1$. We have:

$$\begin{aligned}\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash [!\psi]\varphi_1 &\text{ iff } \mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash \psi \text{ implies } \mathscr{F}^{\ell_0 \ldots \ell_k \psi}, s \vDash \varphi_1 \\ &\text{iff } \mathsf{check}(\mathscr{F}, L, s, \psi) = \mathsf{true} \text{ implies } \mathsf{check}(\mathscr{F}, L{+}{+}[\psi], s, \varphi_1) = \mathsf{true}, \text{by IH}\end{aligned}$$

Note that we make two recursive calls here. We can apply the IH to both because the input size strictly decreases: $|\mathscr{F}| + \sum_{i=0}^{k} |\ell_i| + |\psi| < |\mathscr{F}| + \sum_{i=0}^{k} |\ell_i| + |\psi| + |\varphi_1| + 1$ and $|\mathscr{F}| + \sum_{i=0}^{k} |\ell_i| + |\psi| + |\varphi_1| < |\mathscr{F}| + \sum_{i=0}^{k} |\ell_i| + |\psi| + |\varphi_1| + 1$.

The $\mathsf{check}(\mathscr{F}, L, s, \psi)$ is evaluated in line 15. If it is false, then $\mathsf{check}(\mathscr{F}, [\ell_0, \ldots, \ell_k], s, [!\psi]\varphi_1) = \mathsf{true}$ by line 15. Hence, $\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash [!\psi]\varphi_1$ iff $\mathsf{check}(\mathscr{F}, [\ell_0, \ldots, \ell_1], s, [!\psi]\varphi_1) = \mathsf{true}$. □

Lemma 2.14. *Given a pointed knowledge structure $\mathscr{F} = \langle V, \theta, O_i \rangle, s \subseteq V$ and a* DEL *formula φ, we have $\mathscr{F}, s \vDash \varphi$ iff* $\mathsf{checkDELK}(\mathscr{F}, [], s, \varphi)$ *returns* true.

Proof. We show the more general claim that for any list $L = [(\mathscr{X}_0, x_0), \ldots, (\mathscr{X}_k, x_k)]$ of events we have $(\mathscr{F}, s) \otimes (\mathscr{X}_0, x_0) \otimes \ldots \otimes (\mathscr{X}_k, x_k) \vDash \varphi$ iff $\mathsf{checkDELK}(\mathscr{F}, L, s, \varphi)$ returns true. The proof is by induction on φ, and we omit the easy cases for negation and conjunction.

Case $\varphi = p$. We have the following equivalences:

$$\begin{aligned}&\mathscr{F}^{\ell_0 \ldots \ell_k}, s \vDash p \\ &\text{iff } p \in s^k = (s^{k-1} \setminus V_{k,-}) \cup (s^{k-1} \cap V_{k,-})^{\circ} \cup x^k \cup \{p \in V_{k,-} \mid s^{k-1} \cup x_k \vDash \theta_{k,-}(p)\}, \text{by Definition 2.3} \\ &\text{iff } p \in s^{k,\prime} = (s^{k-1} \setminus V_{k,-}) \cup \{q \in V_{k,-} \mid \mathsf{checkDELK}(\mathscr{F}, [], s, [x_j \sqsubseteq V_k^{+}]\theta_{k,-}(q))\}, \text{because } p \in V \\ &\text{iff } \mathsf{checkDELK}(\mathscr{F}, L, s, p) = \mathsf{true}, \text{by definition of checkDELK}\end{aligned}$$

Note the difference between s^k and $s^{k,'}$ here. $s^{k,'}$ removes the copy variables and the event state (of which variables are from the new event vocabulary) while the last part both represent the modified propositions whose precondition was true in the previous state. However, by assumption $p \in V$, therefore we can simply remove these two parts.

Case $\varphi = K_i\psi$. We have the following equivalences:

$$
\begin{aligned}
&\mathscr{F}^{\ell_0\dots\ell_k}, s \vDash K_i\psi \\
&\text{iff for all } t \text{ in the final model } s^{final} \cup t' \vDash \Omega_i^{final} \text{ implies } \mathscr{F}^{final}, t \vDash \psi \\
&\text{iff } \forall t \subseteq V \text{ s.t. } t \vDash \theta, t_0^+ \subseteq V_0^+, \dots, t_k^+ \subseteq V_k^+ : \\
&\quad s \cup t' \vDash \Omega_i, x_1^+ \cup t_1^+ \vDash \Omega_{0,i}^+, .., x_k^+ \cup t_k^+ \vDash \Omega_{k_i} \text{ and } t \text{ survives implies } \mathscr{F}^{\ell_0\dots\ell_k}, t \vDash \psi \\
&\text{iff } \forall t \subseteq V \text{ s.t. } t \vDash \theta, t_0^+ \subseteq V_0^+, \dots, t_k^+ \subseteq V_k^+ : \\
&\quad s \cup t' \vDash \Omega_i, x_1^+ \cup t_1^+ \vDash \Omega_{0,i}^+, .., x_k^+ \cup t_k^+ \vDash \Omega_{k_i} \text{ and } \bigwedge_{i=0}^{k} \mathscr{F}^{\ell_0\dots\ell_{i-1}}, t \vDash [t_i^+ \sqsubseteq V_i^+]\theta_i^+ \text{ implies } \mathscr{F}^{\ell_0\dots\ell_k}, t \vDash \psi \\
&\text{iff } \mathsf{checkDELK}(\mathscr{F}, L, s, K_i\psi) = \mathsf{true}
\end{aligned}
$$

Case $\varphi = [\mathscr{X}, x]\psi$. We have the following equivalences:

$$
\begin{aligned}
&\mathscr{F}^{\ell_0\dots\ell_k}, s \vDash [\mathscr{X}, x]\psi \\
&\text{iff if } \mathscr{F}^{\ell_0\dots\ell_k}, s \vDash [x \sqsubseteq V^+]\theta^+ \text{ then } (\mathscr{F}^{\ell_0\dots\ell_k} \times \mathscr{X}, s^x) \vDash \psi \\
&\text{iff if } \mathsf{checkDELK}(\mathscr{F}, L, s, [x \sqsubseteq V^+]\theta^+) = \mathsf{true} \text{ then } \mathsf{check}(\mathscr{F}, L{+}{+}[(\mathscr{X}, x)], s, \psi) = \mathsf{true}, \text{by IH } \square
\end{aligned}
$$

Theorem 4.2. *The translation from Def. 4.1 is correct: for any $s,t \subseteq V$, we have $s \xrightarrow{\pi} t$ iff $(s \cup t') \vDash \mathsf{tr}(\pi)$.*

Proof. By induction on the structure of π and applying Definition 3.2.

- $s \xrightarrow{p\leftarrow\top} t \iff s \cup t' \vDash p' \wedge \bigwedge_{p \neq q \in V}(q \leftrightarrow q')$, since $p \in t$ and other variables remain the same.
- $s \xrightarrow{p\leftarrow\bot} t \iff s \cup t' \vDash \neg p' \wedge \bigwedge_{p \neq q \in V}(q \leftrightarrow q')$, since $p \notin t$ and other variables remain the same.
- $s \xrightarrow{\beta?} t \iff s = t \wedge s \vDash \beta \iff s \cup t' \vDash \beta \wedge \bigwedge_{p \in V}(p \leftrightarrow p')$, since β must be true in $s = t$.
- $s \xrightarrow{p \cup q} t \iff s \xrightarrow{p} t$ or $s \xrightarrow{q} t \iff s \cup t' \vDash tr(p)$ or $s \cup t' \vDash tr(q) \iff s \cup t' \vDash tr(p) \vee tr(q)$.
- $$
  \begin{aligned}
  s \xrightarrow{\pi_1;\pi_2} t \quad &\iff \quad \exists u \subseteq V : s \xrightarrow{\pi_1} u \text{ and } u \xrightarrow{\pi_2} t \\
  &\iff \quad \exists u \subseteq V : s \cup u' \vDash \mathsf{tr}(\pi_1) \text{ and } u \cup t' \vDash \mathsf{tr}(\pi_2) \\
  &\iff \quad \exists u \subseteq V : s \cup u' \vDash \mathsf{tr}(\pi_1) \text{ and } u' \cup t'' \vDash [V \mapsto V', V' \mapsto V'']\mathsf{tr}(\pi_2) \\
  &\iff \quad \exists u \subseteq V : s \cup u' \cup t'' \vDash \mathsf{tr}(\pi_1) \wedge [V \mapsto V', V' \mapsto V'']\mathsf{tr}(\pi_2) \\
  &\iff \quad s \cup t'' \vDash \exists V'(\mathsf{tr}(\pi_1) \wedge [V \mapsto V', V' \mapsto V'']\mathsf{tr}(\pi_2)) \\
  &\iff \quad s \cup t' \vDash [V'' \mapsto V'](\exists V'(\mathsf{tr}(\pi_1) \wedge [V \mapsto V'][V' \mapsto V'']\mathsf{tr}(\pi_2)))
  \end{aligned}
  $$
- $s \xrightarrow{p \cap q} t \iff s \xrightarrow{p} t$ and $s \xrightarrow{q} t \iff s \cup t' \vDash tr(p)$ and $s \cup t' \vDash tr(q) \iff s \cup t' \vDash tr(p) \wedge tr(q) \quad \square$

Theorem 4.3. *For any translation tr' from mental programs to BDDs there exists a mental program π such that $\mathsf{tr}'(\pi)$ has size exponential in $|\pi|$.*

Proof. Take a vocabulary $V = \{p_1, \dots p_{2n}\}$ with this ordering fixed. Consider again the formula $\beta := (p_1 \wedge p_{n+1}) \vee \dots \vee (p_n \wedge p_{n+n})$ from page 239. Let $\pi := \beta?$. Then π also has length in $\mathcal{O}(n)$. From Definition 16, $\mathsf{tr}(\pi) = \beta \wedge \bigwedge_{p \in V}(p \leftrightarrow p')$. We show that the BDDs corresponding to this boolean formula has at least 2^{n+1} many nodes. First, as discussed in [9, p. 681], the BDD that corresponds to the β formula has at least 2^{n+1} many nodes.

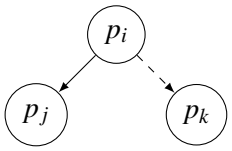


Figure 1: Node p_i in β's BDD.

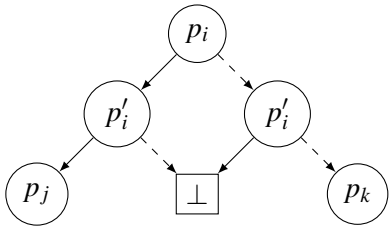


Figure 2: Node p_i in $\beta \wedge \bigwedge_{p \in V}(p \leftrightarrow p')$'s BDD.

Now we examine the size of the BDDs that corresponds to $\beta \wedge \bigwedge_{p \in V}(p \leftrightarrow p')$. Since none of the copy variables p'_i appear in β, it will not cause any subtrees to be merged. Instead, as shown in Figure 2, for each node p_i that appears in $\beta's$ BDD in Figure 1, two more nodes of p'_i are added with the solid and dashed edges encoding the $(p_i \leftrightarrow p'_i)$ sub-formula in π. Therefore, the BDD for π has at least 2^{n+1} many nodes. □

Theorem 4.6. *The translation τ_o from Definition 4.4 is correct. That is, given a vocabulary V, for any $s,t \subseteq V$, we have $s \xrightarrow{\tau_0(\Omega)} t$ iff $(s \cup t') \vDash \Omega$.*

Proof. For any BDD Ω, let Ω^* be its unraveling to a tree. Note that we have $\tau(\Omega) = \tau(\Omega^*)$. Hence for the proof we assume wlog. that β is a tree.

We proceed by induction over the structure of BDDs. In the case when the BDD is a single node $\bot$, the mental program is $?\bot$, which represents the empty relations. When the BDD is a single node $\top$, the mental program is $((p_0 \leftarrow \bot) \cup (p_0 \leftarrow \top)); \dots; ((p_n \leftarrow \bot) \cup (p_n \leftarrow \top)); ?\top$, which represents the complete relations. The translation is correct for both of these base cases.

The trickier case is when the BDD has at least three nodes (including $\top$ and $\bot$). We proceed by an inner induction on the length of the vocabulary. The base case is when the vocabulary is empty and the only BDDs that can be constructed are just $\bot$ and $\top$nodes. As shown above, the translation is correct for both cases.

For the induction step we are adding one more variable p_0 at the beginning of the vocabulary V. We have the inductive hypothesis that for all t_i that can be constructed with the vocabulary $V = [p_1, .., p_k]$, we have $s \xrightarrow{\tau(t_0,V)} t \iff s \cup t' \vDash t_0$.

We prove that the translation remains correct for the extended vocabulary $V_1 := \{p_0\} \cup V$ by showing that both t_0 under V and $\tau(t_0, V)$ encode the same relations.

Given the new vocabulary, the unraveled BDDs that can be constructed all have the shape as shown in Figure 3, where the t_is are subtrees which can be seen as the complete trees constructed in the original vocabulary V. We can construct the corresponding relations in Figure 4. Specifically, after extending the vocabulary, we make a copy of the states in the original vocabulary V and add p_0 to each of these copies to indicate that p_0 is true in these states. The states in which p_0 is false are put on the bottom level in the

figure while the states in which p_0 is true are on the top level. We construct the corresponding relations as follows:

1. Construct the relations corresponding to t_1 in the original vocabulary V on the bottom level.
2. Construct the relations corresponding to t_2 in V on the bottom level and then shift the source states of the relations from the bottom level to the corresponding states (i.e. those that differ only by p_0.) on the top level. In Figure 4 these relations are depicted as the dotted top-to-bottom arrows.
3. Construct the relations corresponding to t_3 in V on the bottom level and then shift the target states of the relations from the bottom level to the corresponding states on the top level. In Figure 4 these relations are depicted as the dotted bottom-to-top arrows.
4. Construct the relations corresponding to t_4 in V on the bottom level and then shift both the source and target states of the relations from the bottom level to the corresponding states on the top level.

Each branch t_i in Figure 3 corresponds to part of the relation in Figure 4.

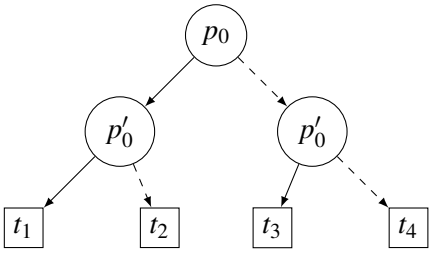


Figure 3: The BDD after p_0 is added.

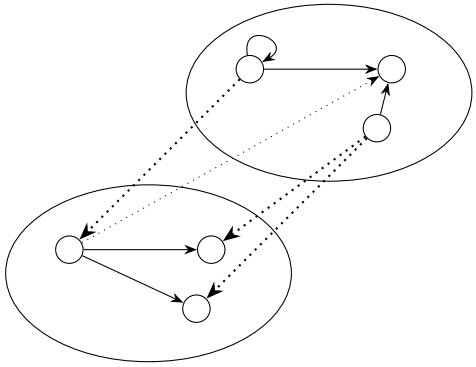


Figure 4: The corresponding Kripke model.

From inductive hypothesis we know that the translations are correct in the original vocabulary, which means that π_i and t_i encode the same relations in the original vocabulary V. Assume $\tau(t_i,V) = \pi_i$, then $\tau(t_0,V_1) = ?\neg p_0;((p_0 \leftarrow \bot;\pi_1)\cup(p_0 \leftarrow \top;\pi_2))\cup ?p_0;((p_0 \leftarrow \bot;\pi_3)\cup(p_0 \leftarrow \top;\pi_4)) = (?\neg p_0;\pi_1)\cup(?\neg p_0;(p_0 \leftarrow \top);\pi_2)\cup(?p_0;(p_0 \leftarrow \bot);\pi_3)\cup(?p_0;\pi_4)$. Similar to the reasoning in the previous part, we can see that this represents the same relation in Figure 4, e.g. the relations corresponding to π_2 go from bottom to top. Therefore, the BDD t_0 and the mental program $\tau(t_0,V_1)$ encode the same relation in the extended vocabulary $\{p_0\}\cup V$. This concludes the proof that the translation is correct. □

Choosing What Game to Play without Selecting Equilibria: Inferring Safe (Pareto) Improvements in Binary Constraint Structures

Caspar Oesterheld Vincent Conitzer
Computer Science Department
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

We consider a setting in which a principal gets to choose which game from some given set is played by a group of agents. The principal would like to choose a game that favors one of the players, the social preferences of the players, or the principal's own preferences. Unfortunately, given the potential multiplicity of equilibria, it is conceptually unclear how to tell which of even any two games is better. Oesterheld et al. [14] propose that we use assumptions about *outcome correspondence* – i.e., about how the outcomes of different games relate – to allow comparisons in some cases. For example, it seems reasonable to assume that isomorphic games are played isomorphically. From such assumptions we can sometimes deduce that the outcome of one game Γ^s is guaranteed to be better than the outcome of another game Γ, even if we do not have beliefs about how each of Γ and Γ^s will be played individually. Following Oesterheld et al., we then call Γ^s a *safe improvement* on Γ.

In this paper, we study how to derive safe improvement relations. We first show that if we are given a set of games and arbitrary assumptions about outcome correspondence between these games, deriving safe improvement relations is co-NP-complete. We then study the (in)completeness of a natural set of inference rules for outcome correspondence. We show that in general the inference rules are incomplete. However, we also show that under natural, generally applicable assumptions about outcome correspondence the rules are complete.

1 Introduction

Imagine you have to choose between two (say, normal-form) games, Γ_1 and Γ_2 for Alice and Bob to play tomorrow. You have some stake in the outcome of the game, e.g., because you are Alice's and/or Bob's friend. Which of Γ_1 and Γ_2 should you choose? In general, assessments of which game to let Alice and Bob play are conceptually difficult, because each of the games may have multiple solutions (e.g., multiple Nash equilibria, multiple rationalizable outcomes). Which game to prefer depends on how this multiplicity is resolved. For example, imagine that you can decide whether Alice and Bob play a game of Chicken, or a trivial game in which they both receive \$5. Then (depending on your preferences over the outcomes) the decision hinges on which – if any – equilibrium of Chicken you expect Alice and Bob to play.

We build on the safe Pareto improvement framework of Oesterheld et al. [14] to address this problem. Roughly, we consider some (potentially infinite) set of games. We make some (qualitative, non-probabilistic) assumptions about how the outcomes of pairs of these games, as played by a set of *agents* (Alice and Bob in the above example), relate. For example, we might assume that if two games are isomorphic, then Alice and Bob would play them isomorphically. We call such statements about the relationships of games *outcome correspondences* (OCs). From these outcome correspondences we can infer new outcome correspondences. For instance, if we know OCs between Γ_a and Γ_b and between

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 251–270, doi:10.4204/EPTCS.437.22

© C. Oesterheld and V. Conitzer

Γ_b and Γ_c, then we can infer an OC between Γ_a and Γ_c (Lemma 1.1 below). In some cases, we will be able to infer outcome correspondences that show that the outcome of one game Γ^s is always better than the outcome of another game Γ, even without resolving the multiplicity of solutions in the underlying games. We will then say that Γ^s is a *safe improvement* on Γ. If in particular we have that the outcome of Γ^s is better than the outcome Γ under a *Pareto* (partial) order over outcomes, we call Γ^s a *safe Pareto improvement*.

For example, consider the three versions of Chicken in Figure 1. Each has exactly two pure and one mixed equilibrium. We make assumptions about outcome correspondence as follows, illustrated by lines in the figure. First, in Γ_a, the pure strategy C' is strictly dominated by C. Further, notice that Γ_b is obtained from Γ_a by removing C'. It therefore seems plausible that Γ_a and Γ_b will be played in the same way. That is, if Alice and Bob play (D,D) in Γ_a, then they will also play (D,D) in Γ_b, and so on. Second, we can see that Γ_b and Γ_c are isomorphic (with payoffs transformed by $x \mapsto 2x+4$ for both players). It therefore seems plausible that Γ_b and Γ_c will be played isomorphically, i.e., if Alice and Bob play (C,C) in Γ_b, then they play (E,E) in Γ_c, and so on.

Let's try to decide whether to let Alice and Bob play Γ_a or Γ_c. Imagine that we aim to optimize for Alice and/or Bob's preferences. Now notice that from the given outcome correspondences we can infer via the aforementioned transitivity rule (Lemma 1.1 below) an outcome correspondence between Γ_a and Γ_c: if Alice and Bob play (D,D) in Γ_a, then they also play (D,D) in Γ_b and thus (F,F) in Γ_c; if Alice and Bob play (D,C) in Γ_a, then they play (F,E) in Γ_c; and so on. By considering each outcome individually, we can see that the outcome of Γ_c is guaranteed to be strictly better for both Alice and Bob. Thus, Γ_c is a safe Pareto improvement on Γ_a; when given the choice, we should let Alice and Bob play Γ_c instead of Γ_a. Importantly, to reach this conclusion, we do not need to assume anything about how Alice and Bob select an equilibrium in Γ_a, Γ_b, and Γ_c.

The concept of safe (Pareto) improvements is widely applicable to strategic settings. Oesterheld et al. [14] consider the case of multiple players instructing their respective agents by shaping their utility functions [cf. 2; 4, Sect. 4.2]. Sauerberg et al. [16] show how the concept can be used to assess *ex post* verifiable commitments. However, the broad conceptualization of safe (Pareto) improvements of the present paper points to many new types of applications. For example, mechanism design is the study of how to shape a game played by some set of players when optimizing for social preferences (efficiency) or the designers preferences, e.g., profit (as in optimal mechanisms). In some settings – specifically settings where mechanisms induce a game with multiple Nash equilibria – the concept of safe (Pareto) improvements can be used to assess and compare different possible mechanisms. As such, the concept of safe improvements complements concepts like best-Nash (cf. weak implementation in implementation theory; price of stability) and worst-Nash (cf. full/strong implementation in implementation theory; price of anarchy) that are widely used for comparing different mechanisms. (See also Monderer et al. [12], who, very roughly, use worst-undominated-strategies for comparing different mechanisms.)

Contributions: In this paper, we introduce a new, general framework for reasoning about safe (Pareto) improvements (S(P)Is) (Section 3). In doing so, we generalize the setting of Oesterheld et al. [14]: We define SIs relative to arbitrary preferences over outcomes rather than specifically Pareto preferences. Furthermore, we define SIs on arbitrary binary constraint structures similar to those featured in constraint satisfaction problems (CSP), rather than specifically normal-form games with specific outcome correspondence assumptions. We address the following question: Given some set of games and given some assumptions about outcome correspondence between these games, can we infer that one game is an SI on another?

In Section 4, we analyze the computational complexity of deciding this problem if we are given a finite set of games and explicitly represented outcome correspondence assumptions between these games. We

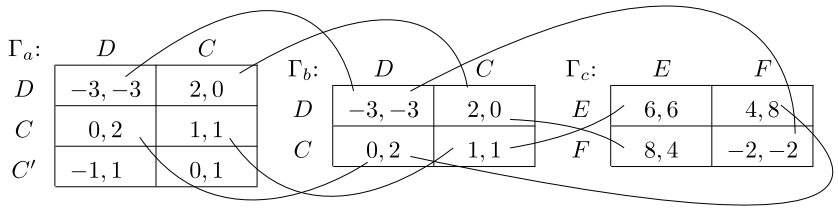


Figure 1: An example of three normal-form games with plausible outcome correspondences between them illustrated by lines: If playing Γ_b would result in (D,D), then playing Γ_c would result in (F,F), etc.

prove that this problem is co-NP-complete, even under various restrictions on the outcome correspondence assumptions. Roughly, we view the assumptions as defining a binary constraint structure *a la* the binary constraint satisfaction problem (CSP). (Binary) CSP is NP-complete. However, in the context of SIs, we are not interested in finding a satisfying assignment (a way in which the agents might play the different games that satisfies all the assumptions) but in whether specific facts hold about *all* satisfying assignments. This type of problem is co-NP-complete, because it is the complement of an NP-complete problem.

In Section 5, we consider the problem of inferring SIs under restrictions on what type of assumptions we make. In particular, we ask whether under such restrictions, "local" inference rules (such as the aforementioned transitivity rule) are complete. We note that most natural generic assumptions about outcome correspondence (such as: isomorphic games are played isomorphically; dominated actions can be removed) satisfy a condition called *max-closedness* known from the CSP literature [9]. We show that under this condition, the local rules of inference are complete (Theorem 4). From this, we obtain an efficient algorithm for finding SIs. We show that completeness continues to hold even if we consider infinitely many games and outcome correspondence assumptions (Section 5.3). We demonstrate the usefulness of the results about completeness of repeated inference by refining Oesterheld et al.'s [14] NP-completeness result for a specific setting.

2 Preliminaries

Normal-form games. We assume familiarity with basic game-theoretic concepts, but briefly introduce some notation. An *n-player (normal-form) game* is a tuple $\Gamma = (A_1,...,A_n,u_1,...,u_n)$ of sets of *(pure) strategies* or *actions* $A_1,...,A_n$ and *utility functions* $u_i\colon A_1 \times ... \times A_n \to \mathbb{R}$ mapping *(pure) strategy profiles* or *outcomes* of the game onto utilities. For a strategy profile $\boldsymbol{a}$, let $\boldsymbol{a}_{-i}$ denote the strategy profile consisting of all pure strategies in $\boldsymbol{a}$ except Player i's strategy a_i. For $a_i \in A_i$, we denote by $(\boldsymbol{a}_{-i},a_i) \in A_1 \times ... \times A_n$ the strategy profile that arises from combining $\boldsymbol{a}_{-i}$ and a_i in the natural way. We say that one outcome $\boldsymbol{a}$ *is Pareto-better than* another outcome $\boldsymbol{a}'$ if $u_i(\boldsymbol{a}) \geq u_i(\boldsymbol{a}')$ for all i, with strictness for some i.

A *(Nash) equilibrium of* Γ is a strategy profile $\boldsymbol{a}$ s.t. for every player i and all pure strategies $a'_i \in A_i - \{a_i\}$ we have $u_i(a'_i,\boldsymbol{a}_{-i}) \leq u_i(\boldsymbol{a})$. Further, $\boldsymbol{a}$ is a *strict (Nash) equilibrium* if all the inequalities are strict. We say that a strategy a_i *strictly dominates* another strategy a'_i if for all $\boldsymbol{a}_{-i}$ we have $u_i(a_i,\boldsymbol{a}_{-i}) > u_i(a'_i,\boldsymbol{a}_{-i})$. We call $(\Phi_i\colon A_i \to A'_i)_{i=1,...,n}$ an *isomorphism* between games Γ and Γ' if there are $(\lambda_i \in \mathbb{R}_+)_i, (b_i \in \mathbb{R})_i$ such that for all strategy profiles $\boldsymbol{a}$ and players i we have $u_i(\boldsymbol{a}) = \lambda_i u'_i(\Phi_1(a_1),...,\Phi_n(a_n)) + b_i$.

Set-valued functions. We use set-valued functions to express outcome correspondences. A set-valued function $\Phi\colon X \multimap Y$ is a binary relation between X and Y, i.e., Φ is a subset of $X \times Y$. For any $x \in X$, we write $\Phi(x) \coloneqq \{y \in Y \mid (x,y) \in \Phi\}$. For $M \subseteq X$, write $\Phi(M) \coloneqq \bigcup_{x \in M} \Phi(x)$. For $\Phi\colon X \multimap Y$ we define the inverse as $\Phi^{-1}\colon Y \multimap X\colon y \mapsto \{x \in X \mid y \in \Phi(x)\}$. For two multi-valued functions $\Phi\colon X \multimap Y$ and

$\Psi\colon Y \multimap Z$, we define their composition as $\Psi \circ \Phi\colon x \mapsto \Psi(\Phi(x))$. Finally, for $\Phi, \Psi\colon X \multimap Y$ we define $\Phi \cap \Psi\colon X \multimap Y\colon x \mapsto \Phi(x) \cap \Psi(x)$ to be the intersection of Φ and Ψ.

Binary constraint structures. A *binary constraint structure (BCS)* is a triplet $(\mathscr{X}, \mathscr{D}, \mathscr{A})$, where $\mathscr{X}$ is a set of variables; $\mathscr{D} = (D_X)_{X \in \mathscr{X}}$ is a family of finite domains, one for each variable; and $\mathscr{A}$ is a set of binary constraints. Each binary constraint is a triplet (X, Y, Φ) where X and Y are variables and $\Phi \subseteq D_X \times D_Y$ is a subset of the product of the domains corresponding to X and Y. Following the notation of Oesterheld et al. [14], we generally write constraints as $X \sim_\Phi Y$. An *assignment* for $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ associates with each variable X a value v_X in D_X. We say that an assignment *satisfies* a BCS if for each constraint $(X, Y, \Phi) \in \mathscr{A}$ we have that $(v_X, v_Y) \in \Phi$. In this paper, we also call the constraints *outcome correspondences* (OCs). We call $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ finite if $\mathscr{X}$ and $\mathscr{A}$ are finite. The *binary constraint satisfaction problem (CSP)* consists in deciding whether there exists a satisfying assignment for a given BCS. The problem is well-known to be NP-complete.

Inference in binary constraint structures. Instead of constraint satisfaction, we will usually be concerned with problems of *inference* on BCSs. That is, given a BCS, we will be interested in identifying additional constraints (not given $\mathscr{A}$) that all satisfying assignments of the BCS must satisfy. One specific method for inference is to apply specific sets of rules that take a small number of existing constraints and construct a new constraint from them. In the literature on CSP, this is known as *constraint propagation*. In this paper, we will primarily consider the following set of rules [cf. 14]. In the language of the CSP literature, these are the rules needed to impose *path consistency* [13].

Lemma 1. *Let X, Y and Z be variables of a BCS with domains D_X, D_Y and D_Z, and $\Phi, \Xi\colon D_X \multimap D_Y$, $\Psi\colon D_Y \multimap D_Z$. Then: 1. Transitivity: If $X \sim_\Phi Y$ and $Y \sim_\Psi Z$, then $X \sim_{\Psi\circ\Phi} Z$. 2. Intersection: If $X \sim_\Phi Y$ and $X \sim_\Xi Y$, then $X \sim_{\Phi\cap\Xi} Y$. 3. Reflexivity: $X \sim_{\mathrm{id}_{D_X}} X$, where $\mathrm{id}_{D_X}\colon D_X \multimap D_X\colon x \mapsto \{x\}$. 4. Symmetry: If $X \sim_\Phi Y$, then $Y \sim_{\Phi^{-1}} X$. 5. $X \sim_{\mathrm{all}_{D_X, D_Y}} Y$, where $\mathrm{all}_{D_X, D_Y}\colon D_X \multimap D_Y\colon x \mapsto D_Y$.*

3 Safe (Pareto) improvements and outcome correspondence

A generic definition of safe (Pareto) improvements. We now introduce a new formalism for safe (Pareto) improvements, which generalizes that of Oesterheld et al. [14]. (We give a brief comparison in Section 6; and a detailed comparison in Appendix B.) In the following, we first describe safe improvements in generic binary constraint structures. We then discuss safe (Pareto) improvements for games in particular.

Imagine that we face a choice between some set of variables $X_0, X_1, ..., X_n$. Each variable can result in one of a number of possible values or *outcomes*. We are uncertain which outcome any given variable might result in. We have some preferences $\succeq$ over the outcomes of the different variables. These preferences may be incomplete, i.e., there may be pairs of outcomes of which neither is preferred to the other. For instance, we may be choosing as a group of people with diverging preferences and our group preferences are only definite under consensus.

The most standard approaches to choosing between the variables would involve assigning probabilities to outcomes. But in some cases such probabilistic beliefs are controversial or very difficult to assign. Suppose that instead of probabilistic beliefs, we are only willing or able to act on qualitative beliefs about how the outcomes of different variables relate. For instance, we might strongly believe that if X_i were to result in outcome o, then X_j would result in outcome o'. Then we may still be able to adjudicate between some pairs of variables. In particular, from the outcome correspondence assumptions we may be able to infer that one variable will always yield an outcome that is preferred over the outcome of another variable.

By interpreting the variables, outcomes, and outcome correspondence assumptions as a BCS, we can formalize this as follows.

Definition 1. *Let $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ be a binary constraint structure. Let $\succeq$ be a partial order on $\bigcup \mathscr{D}$ (i.e., on the union of all the domains in $\mathscr{D}$). Let $X, Y \in \mathscr{X}$ be two variables. We say that Y is a safe improvement (SI) on X w.r.t. $\succeq$ if each satisfying assignment $(v_Z)_{Z \in \mathscr{X}}$ of $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ satisfies $v_Y \succeq v_X$. We say that Y is a strict safe improvement on X if $v_Y \succ v_X$ for each satisfying assignment.*

For this definition we use the canonical definition of $\succ$ as the antisymmetric version of $\succeq$, i.e., $a \succ b$ if and only if $a \succeq b$ but not $b \succeq a$.

Note that the SI condition is itself an outcome correspondence claim [cf. 14, Theorem 3]: X' is an SI on X if and only if $X \sim_\Phi X'$, where $\Phi \colon D_X \to D_{X'} \colon o \mapsto \{o' \colon o' \succeq o\}$ associates any outcome o in X with all outcomes o' of X' that are preferred over o.

Our notion of strictness is not the only possible version. One could also consider a notion of strictness that *existentially* quantify over satisfying assignments. E.g., in addition to the (non-strict) SI condition one could require that there is a satisfying assignment v s.t. $v_Y \succ v_X$. Such definitions give rise to different hardness results than in our paper. All the results about polynomial-time decidability transfer, however. We leave a detailed investigation to future work.

One perspective on the safe improvement relation is that it is an application of decision-theoretic dominance with preferences $\succeq$, with uncertainty over the satisfying assignments (see Section 6).

Safe (Pareto) improvements in strategic settings. Throughout this paper, we primarily consider SIs in strategic settings. That is, we imagine that each variable in $\mathscr{X}$ corresponds to a possible strategic interaction between a group of *agents*. For simplicity, we assume specifically that the interaction consists of the agents playing a normal-form game. The domains $\mathscr{D}$ are the possible outcomes of the normal-form games. We imagine that a *principal* with preferences $\succeq$ chooses which game is played by the agents. Note that $\succeq$ compares outcomes across different normal-form games and $\succeq$ is not necessarily (though in this paper often will be) related to the agents' utilities in the underlying games.

In strategic settings, we will generally use the term safe *Pareto* improvement if $\succeq$ reflects the Pareto ordering of the outcomes w.r.t. to the utilities specified by the respective games. In general, this differs somewhat from Oesterheld et al. [14]; see Appendix B. Note that we are thereby assuming that the utilities are (at least ordinally) comparable across the games in question. That is, if $u_1(o)$ is Player 1's utility in outcome o of game Γ under consideration, and $u_1'(o')$ is Player 1's utility in outcome o' of another game Γ', when Player 1 prefers o over o' if $u_1(o) > u_1'(o')$. In some sense, we assume that we are given a player's single utility function over all outcomes of all games.

One reason for the principal to be uncertain about the outcomes of games is the multiplicity of solutions (e.g., of Nash equilibria or non-dominated strategies). This uncertainty seems especially intractable – we consider equilibrium selection an unsolved and probably unsolvable problem. Therefore, it is natural to be unwilling to assign probabilistic beliefs about which equilibrium (if any) of a game will materialize.

While it is hard to judge how the multiplicity of solutions is resolved in a single game, it is natural to make some qualitative assumptions about how the outcomes of different games relate. We now give two specific assumptions about outcome correspondence that we have already discussed in the introduction. These assumptions are plausible in many settings and have long been discussed in the game-theoretic literature [1, 10, 15, 7]. The first assumption is that we can remove strictly dominated strategies from any given game and obtain a game that will be played equivalently by the agents. The second assumption is that isomorphic games are played isomorphically.

Assumption 1 ([cf. 14, Assumption 1])**.** *Let $\Gamma = (A_1, ..., A_n, \mathbf{u})$ be a game. For $i = 1, ..., n$, let $A_i' \subseteq A_i$ be a subset of actions s.t. all actions in $A_i - A_i'$ are strictly dominated in Γ. Then consider the game $\Gamma' = (A_1', ..., A_n', \mathbf{u}_{|A_1' \times ... \times A_n'})$, which can be obtained from Γ by elimination of dominated strategies. Then $\Gamma \sim_\Phi \Gamma'$, where Φ is defined by $\Phi(\boldsymbol{a}) = \{\boldsymbol{a}\}$ if $a \in \bigtimes_i A_i'$ and $\Phi(\boldsymbol{a}) = \emptyset$ otherwise.*

Assumption 2 ([cf. 14, Assumption 2]). *Let $\Gamma = (A_1, ..., A_n, \mathbf{u}), \Gamma' = (A'_1, ..., A'_n, \mathbf{u}')$ be two isomorphic games that do not contain any strictly dominated strategies. Let $\mathscr{I}$ be the set of isomorphisms between Γ and Γ'. Then $\Gamma \sim_I \Gamma'$, where I is defined by $I(\boldsymbol{a}) = \{\Phi_1(a_1) \mid \Phi \in \mathscr{I}\} \times ... \times \{\Phi_n(a_n) \mid \Phi \in \mathscr{I}\}$.*

Let's illustrate our formalism using the three games $\Gamma_a, \Gamma_b, \Gamma_c$ in Figure 1. Assumption 1 gives us the outcome correspondence between Γ_a and Γ_b that we described informally in the introduction. That is, $\Gamma_a \sim_\Phi \Gamma_b$, where $\Phi\colon (D,D) \mapsto \{(D,D)\}, (D,C) \mapsto \{(D,D)\}, (C,D) \mapsto \{(C,D)\}, (D,D) \mapsto \{(D,D)\}, (C',D) \mapsto \emptyset, (C',C) \mapsto \emptyset$. Note that there is only one isomorphism between Γ_b and Γ_c. Thus, Assumption 2 allows us to obtain that $\Gamma_b \sim_\Psi \Gamma_c$, where $\Psi\colon (D,D) \mapsto \{(F,F)\}, (D,C) \mapsto \{(F,E)\}, (C,D) \mapsto \{E,F\}, (C,C) \mapsto \{(E,E)\}$. Using the transitivity rule (Lemma 1.1) we get that $\Gamma_a \sim_{\Psi\circ\Phi} \Gamma_c$, where $\Psi \circ \Phi\colon (D,D) \mapsto \{(F,F)\}, (D,C) \mapsto \{(F,E)\}, (C,D) \mapsto \{E,F\}, (C,C) \mapsto \{(E,E)\}, (C',D) \mapsto \emptyset, (C',C) \mapsto \emptyset$. Clearly, $\Psi \circ \Phi$ is strictly Pareto improving with respect to the Pareto order on the respective players' utilities. Thus Γ_c is a strict SPI on Γ_a.

4 Complexity results under general outcome correspondence assumptions

We consider the following computational problem: Given a satisfiable, finite BCS, some partial order over outcomes, and two variables in the BCS, decide whether one of the variables is a safe improvement on the other. We also consider the question of whether *any* variable in the BCS safely improves on a given variable; as well as whether there exist any *two* variables in the BCS such that one is a safe improvement on the other. The restriction to satisfiable BCS is natural given that in our context BCSs express beliefs about (and is thus satisfied by) the agents' behavior. We show that these problems are co-NP-complete, even under various restrictions that are natural in the context of games. To warm up, we first give a weaker version for generic satisfiable BCSs and arbitrary partial orders $\succeq$:

Proposition 2. *The following decision problems are co-NP-complete. Given a satisfiable, finite BCS $(\mathscr{X}, \mathscr{D}, \mathscr{A})$, a partial order $\succeq$ over $\bigcup \mathscr{D}$, and two variables $X, X' \in \mathscr{X}$, decide whether: 1. X' is a (strict) SI on X. 2. there is $\tilde{X} \in \mathscr{X}$ s.t. $\tilde{X}$ is a (strict) SI on X. 3. there are $\hat{X}, \tilde{X} \in \mathscr{X}$ s.t. $\tilde{X}$ is a (strict) SI on $\hat{X}$.*

The result is based on the close connection between inferring SIs and the complement of binary CSP, which is co-NP-complete. It's especially easy to see that we can reduce from the first problem to the complement of binary CSP: Simply add to $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ a constraint that the outcome of X' cannot be better than the outcome of X. Then the resulting BCS is satisfiable if and only if X' is *not* an SI on X. Since Proposition 2 follows from Theorem 3 below, we omit a detailed proof.

We now further refine Proposition 2. One natural question is whether the correspondence between CSP and reasoning about OC persists if we restrict attention to assumptions about outcome correspondences that one might plausibly believe in a game-theoretic context. For example, what happens if we exclude assumptions that imply irrational behavior (such as cooperation in the one-shot Prisoner's Dilemma)?

To (partially) enforce plausibility of the assumptions, we define the following. Let $\mathscr{G}$ be a finite set of games. Let $\mathscr{A}_*(\mathscr{G})$ be the set of OCs that arise from applying Assumptions 1 and 2. Besides restrictions on the assumptions, we might also ask what happens if we consider specific types of preferences $\succeq$, e.g., if we require $\succeq$ to be Pareto preferences w.r.t. the utilities in the underlying normal-form games. The following result summarizes our co-NP-completeness result under all of these restrictions.

Theorem 3. *The following decision problems are co-NP-complete: Given a finite set of games $\mathscr{G}$, a finite satisfiable set $\mathscr{A}$ of outcome correspondences that includes $\mathscr{A}_*(\mathscr{G})$, two games $\Gamma, \Gamma' \in \mathscr{G}$, and a preference ordering $\succeq$, decide whether 1. $\mathscr{A}$ implies that Γ' is a (strict) SI on Γ; 2. $\mathscr{G}$ contains a (strict) SI on Γ; 3. $\mathscr{G}$ contains any Γ', Γ s.t. Γ' is a (strict) SI on Γ. The problems remain co-NP-complete if we restrict $\succeq$ to be Player 1's preferences or Pareto preferences, and/or we bound the sizes of the games in $\mathscr{G}$.*

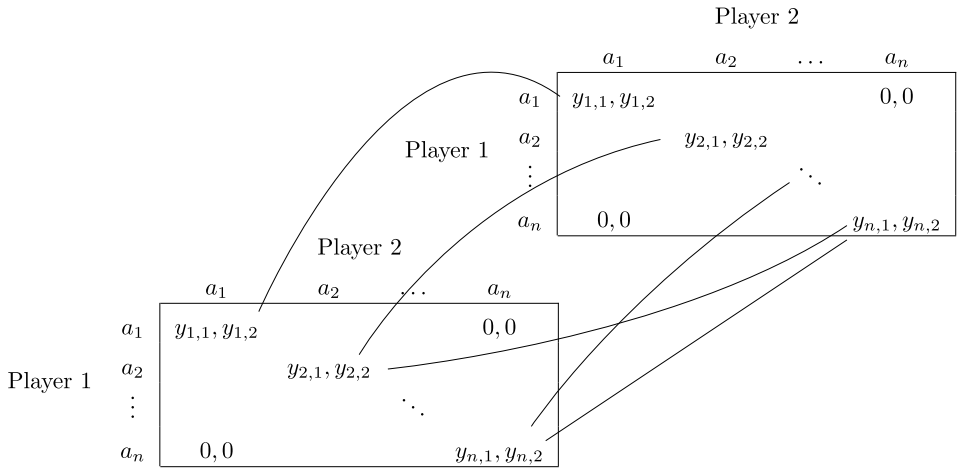


Figure 2: A type of outcome correspondence to which we can reduce binary constraint satisfaction problems without producing wildly implausible outcome correspondences.

Since we cannot expect to give a full account of what a plausible set $\mathscr{A}$ looks like, we here also give a brief account of what types of games and OC assumptions we use for our reduction from (the complement of) binary CSP. These games are sketched in Figure 2. Essentially, these are games with n pure, Pareto-optimal equilibria along the diagonal. We assume that the agents successfully coordinate on one of these equilibria (and thus play rationally in some sense) but we do not know which. Our assumptions about outcome correspondence give us information about how the agents select equilibria across games. From considering only the payoff matrices, it is unclear how one would ever arrive at such judgments. However, we might imagine further contextual information (à la *Schelling* or *focal points* [17, pp. 54–58]) under which these outcome correspondences would be plausible.

Repeated application of rules like Lemma 1 can be done in polynomial time (see Appendix D). Therefore, unless P = NP, repeated application of such rules is insufficient for deriving SIs. We discuss this in detail in Appendix E, where we also provide an example where the rules of Lemma 1 are incomplete.

5 Conditions for completeness of rule-based inference

In this section, we provide conditions under which inference as per Lemma 1 is complete and thus SIs can be found in polynomial time.

5.1 Further outcome correspondence assumptions

To guide our results, we first give some additional structurally interesting, natural assumptions about outcome correspondence. As a first example, one might assume that agents will play a Nash equilibrium.

Assumption 3. *Let $\Gamma = (A, \mathbf{u})$ be a game that has at least one pure Nash equilibrium. Then $\Gamma \sim_\Phi \Gamma$, where $\Phi\colon A \multimap A$ is defined by $\Phi(\boldsymbol{a}) = \{\boldsymbol{a}\}$ if $\boldsymbol{a}$ is a Nash equilibrium and $\Phi(\boldsymbol{a}) = \emptyset$ otherwise.*

While to some extent orthodox, we find this assumption questionable in games with multiple equilibria. It nevertheless has illustrative value. There is also one technical issue: As is well known, some normal-form games (e.g., Rock–Paper–Scissors) have some (or only) mixed equilibria. The language of outcome

correspondence cannot express that the players play mixed strategies. (To address this we would have to restrict attention to games in which all relevant mixed strategies are represented as pure strategies – but this is cumbersome to set up.) We therefore assume that the players play a *pure* equilibrium (which renders the assumption even less plausible) and restrict it to games that have at least one such pure equilibrium.

A hierarchical outcome correspondence assumption. To motivate our next assumption, consider the game on the left of Table 1 (essentially a Stag Hunt). The game has two pure equilibria: (a_H, a_H) and (a_L, a_L). The former Pareto-dominates the latter, but a_L is "playing it safe". In many contexts, it is unclear what actions the players will choose. Next consider the game on the right of Table 1. Qualitatively, this game is similar to the game on the left, but with slightly different payoffs. If we look carefully at the payoffs, we find that from left to right, action a_H becomes more attractive to both players. (To see this, consider the changes between the two games and note that each of them favors playing a_H. For instance, in the right-hand game, playing a_H gives Player 1 an additional payoff of 1, regardless of what Player 2 does. This is an extra argument for Player 1 to play a_H.) We thus argue for the outcome correspondence $\Gamma_1 \sim_{(\Phi_1,\Phi_2)} \Gamma_2$, where $\Phi_i \colon a_H \mapsto \{a_H\}, a_L \mapsto \{a_L, a_H\}$. We now give a generalized outcome correspondence assumption based on this claim. For simplicity we stick to 2-player 2x2 games.

Assumption 4. *Let* $\Gamma = (\{a_1^1, a_1^2\}, \{a_2^1, a_2^2\}, u_1, u_2)$ *and* $\hat{\Gamma} = (\{\hat{a}_1^1, \hat{a}_1^2\}, \{\hat{a}_2^1, \hat{a}_2^2\}, \hat{u}_1, \hat{u}_2)$. *Let* (a_1^1, a_2^1) *and* (a_1^2, a_2^2) *be strict Nash equilibria of* Γ *and let* $(\hat{a}_1^1, \hat{a}_2^1)$ *and* $(\hat{a}_1^2, \hat{a}_2^2)$ *be strict Nash equilibria of* $\hat{\Gamma}$. *Let* (a_1^1, a_2^1) *strictly Pareto dominate* (a_1^2, a_2^2) *(in terms of* u_1, u_2*), and let* $(\hat{a}_1^1, \hat{a}_2^1)$ *strictly Pareto dominate* $(\hat{a}_1^2, \hat{a}_2^2)$. *Further, let* $\hat{u}_i(\hat{a}_i^1, \hat{a}_{-i}^k) \geq u_i(a_i^1, a_{-i}^k)$ *and* $\hat{u}_i(\hat{a}_i^2, \hat{a}_{-i}^k) \leq u_i(a_i^2, a_{-i}^k)$ *for* $i = 1, 2$ *and* $k = 1, 2$. *Then* $\Gamma \sim_{(\Phi_1,\Phi_2)} \hat{\Gamma}$, *where* $\Phi_i \colon a_i^1 \mapsto \{\hat{a}_i^1\}, a_i^2 \mapsto \{\hat{a}_i^1, \hat{a}_i^2\}$.

5.2 Inference is complete if the outcome correspondence assumptions are max-closed

We now give a sufficient condition under which inference as per Lemma 1 is complete.

Definition 2 ([9, Def. 2.5])**.** *We call a BCS* $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ *max-closed if there exist total orders* $(\geq_X)_{X \in \mathscr{X}}$ *s.t. each outcome correspondence* $\Phi \in \mathscr{A}$ *has* $(\max(x_1, x_2), \max(y_1, y_2)) \in \Phi$ *whenever* $(x_1, y_1), (x_2, y_2) \in \Phi$.

Figure 3 illustrates what kind of outcome correspondences are not allowed. An outcome correspondence violates max-closedness w.r.t. a given order, if it has outcomes associated as per the solid lines and potentially the dashed line. To make this max-closed, we would need x_1 and y_1 to also be associated, or we would need to delete one of the "cross" associations (x_1, y_2) and (x_2, y_1). Note that the orders $\geq_X$ in the definition of max-closedness are unrelated to the ordering $\succeq$ used for defining SIs, which are primarily used for comparing outcomes between *different* variables.

Jeavons et al. [9] prove that under max-closedness, *satisfiability* of a BCS is decidable in polynomial time. In particular, their proof shows that iterated application of the rules of Lemma 1 will infer an empty outcome correspondence if and only if the BCS is unsatisfiable. We generalize this result to show that iterated application of the rules of Lemma 1 is also complete for inferring outcome correspondences, i.e., for deciding whether all satisfying assignments satisfy some further OC such as an SI relationship.

Theorem 4. *Let* $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ *be a max-closed BCS. Then inference by repeated application of Lemma 1 is complete for* $(\mathscr{X}, \mathscr{D}, \mathscr{A})$. *Consequently, the three SI decision problems of Proposition 2 and Theorem 3 can be decided in polynomial time.*

The following shows that all our assumptions are max-closed. Consequently, by Theorem 4, inference as per Lemma 1 is complete for our assumptions.

Proposition 5. *Consider any finite set of games* $\mathscr{G}$ *and set of outcome correspondences* $\mathscr{A}$ *resulting from the application of any subset of Assumptions 1 to 4 to any subsets of games. Then* $(\mathscr{G}, \mathscr{A})$ *is max-closed.*

	a_H	a_L
a_H	8,8	0,4
a_L	4,0	7,7

	a_H	a_L
a_H	9,8	1,3
a_L	4,1	7,7

Table 1: Example to illustrate Assumption 4. From left to right, a_H only becomes more attractive for both players.

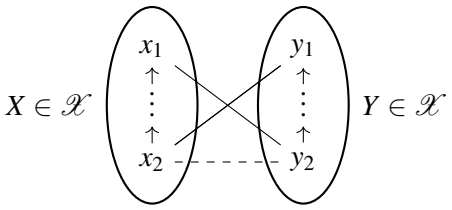


Figure 3: An illustration of the kinds of OCs that are *not* allowed under max closedness. Arrows indicate orderings ($\geq_X$, $\geq_Y$); lines indicate OCs.

Corollary 6. *Consider any finite set of games $\mathscr{G}$ and assumptions $\mathscr{A}$ resulting from the application of Assumptions 1 to 4. Then repeated inference as per Lemma 1 on $(\mathscr{G}, \mathscr{A})$ is complete on $\mathscr{G}$.*

Jeavons et al. [8] provide a generalization of Definition 2 to partial orders under which satisfiability can still be decided in polynomial time using repeated application of Lemma 1. It follows that inference of OCs can also still be done in polynomial time under this condition. However, repeated application of the rules of Lemma 1 is insufficient in this case. We discuss this in detail in Appendix A.

5.3 Generalizing completeness to infinite binary constraint structures

In many cases, the set of variables or games under considerations is infinite. For instance, Oesterheld et al. [14] consider a setting in which the principals can specify the utility functions of the agents. In principle, there are infinitely many utility functions and thus infinitely many possible games to have the agents play.

It is hard to say anything general about the complexity of inferring SIs on infinite BCSs. The answer depends heavily on how we construct finite representations of infinite binary constraint structures. For instance, note that any language for such finite-length representations will only be able to represent a countable subset of the uncountable set of infinite binary constraint structures. In the following, we essentially show a *compactness* result (similar to compactness results for propositional logic) for inference on binary constraint structures. Roughly, the result shows that if an SI can be inferred in some infinite BCS, then it can already be inferred from a finite subset of that BCS.

Theorem 7. *Let $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ be a BCS with countable $\mathscr{X}$ and finite domains. Let $X, X' \in \mathscr{X}$ be variables such that $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ implies $X \sim_\Phi X'$. Then there exists a finite subset $\mathscr{X}'$ of variables and a finite subset of the constraints $\mathscr{A}' \subseteq \mathscr{A}$, such that $X \sim_\Phi X'$ follows from $(\mathscr{X}', \mathscr{D}, \mathscr{A}')$. Assuming the axiom of choice, this continues to hold for uncountable $\mathscr{X}$.*

It follows that if inference rules like those in Lemma 1 are complete on all finite subsets of an infinite binary constraint structure, they are also complete for the infinite binary constraint structure itself:

Corollary 8. *Let $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ be a countable BCS with finite domains. If the rules of Lemma 1 are complete for every finite sub-BCS (defined by $\mathscr{X}' \subseteq \mathscr{X}$ and constraints $\mathscr{A}' \subseteq \mathscr{A}$ over $\mathscr{X}'$), then they are also complete for $(\mathscr{X}, \mathscr{D}, \mathscr{A})$.*

Oesterheld et al. [14] state a result about the complexity of deciding the existence of safe Pareto improvements that are provable by the rules in Lemma 1 (less item 2). Using Corollary 8, we can show that the result holds not just for inference via Lemma 1 but for inference period. For details, see Appendix G.

6 Related work

Constraint satisfaction problems. Technically, our work is most closely related to the literature on (binary) constraint satisfaction problems and in particular constraint propagation. We have used some ideas and results from that literature, such as the NP-completeness of CSP and the max-closedness condition [9, 8]. Our perspective differs from the CSP literature in two ways. First, we are interested in *inference* on binary constraint structures, as opposed to finding satisfying assignments. More specifically, we are interested in inferring SIs. This inference problem is different from the problem of finding satisfiable assignments, sometimes in subtle ways (e.g., see Appendix A). Second, because we work in a specific domain (strategic interactions), we are interested in specific types of constraints. Still, we have established a novel and direct connection between safe improvements in game theory and the CSP literature, which we consider to be an important part of our contribution.

Prior work on safe Pareto improvements. In terms of the game-theoretic motivation, our work is most closely related to prior work on SPIs [14]. Our basic setup and definition of safe (Pareto) improvements generalizes Oesterheld et al.'s [14] setup by allowing arbitrary assumptions about outcome correspondence, arbitrary interventions on how the agents play the game, and arbitrary preferences over outcomes. Appendix B gives a more detailed comparison of the setups.

Our hardness results are different from those of Oesterheld et al. [14] and Sauerberg et al. [16]. In Oesterheld et al.'s NP-hardness result (Theorem 9), $\mathscr{X}$ is infinitely large and the result hinges on $\mathscr{X}$ being more than polynomially large as a function of the problem input size (since otherwise the existence of a polynomial-time algorithm follows from max-closedness). Our co-NP-hardness results are driven by allowing a wider range of outcome correspondence assumptions on finite, explicitly represented sets of games, whereas other work [14, 16] always consider specific assumptions (Assumptions 1 and 2). Prior work on SPIs has not considered the question of completeness of inference as per Lemma 1 at all.

One perspective on the concept of safe improvements is that it is an application of the decision-theoretic dominance principle to choice under uncertainty about how the agents resolve equilibrium selection problems. That is, let the set of states be the satisfying assignments of $(\mathscr{X}, \mathscr{D}, \mathscr{A})$. Let the actions be the variables in $\mathscr{X}$. And let the outcome of choosing action X in a given state (assignment) $(v_X)_{X \in \mathscr{X}}$ be the outcome v_X. Then X' is a safe improvement on X if X' dominates X w.r.t. the given ordering $\succeq$. While conceptually there is a close relationship between the SI concept and the dominance principle, we are unaware of technical work on decision-theoretic dominance that is related to the technical concepts developed in our paper.

Acknowledgments

We thank Emery Cooper, Nathaniel Sauerberg and our anonymous referees for helpful discussions. Caspar Oesterheld's work on this project was funded by the FLI AI existential risk PhD fellowship.

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), pp. 1–48, doi:10.2202/1534-5971.1141.

[2] T. Baumann (2017): *Using surrogate goals to deflect threats*. Available at `https://s-risks.org/using-surrogate-goals-to-deflect-threats/`.

[3] B. Bukh (2012): *Walk through Combinatorics: Compactness principle*. Available at www.borisbukh.org/DiscreteMath12/notes_compactness.pdf. Notes for CMU Course 21-701.

[4] J. Clifton (2020): *Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda*. Available at https://longtermrisk.org/files/Cooperation-Conflict-and-Transformative-Artificial-Intelligence-A-Research-Agenda.pdf.

[5] R. Cowen (1998): *A compactness theorem for infinite constraint satisfaction*. Reports on Mathematical Logic 32, pp. 97–107.

[6] I. Gilboa, E. Kalai & E. Zemel (1990): *On the order of eliminating dominated strategies*. Operations Research Letters 9(2), pp. 85–89, doi:10.1016/0167-6377(90)90046-8.

[7] John C. Harsanyi & Reinhard Selten (1988): *A General Theory of Equilibrium Selection in Games*. The MIT Press, Cambridge, MA.

[8] P. Jeavons, D. Cohen & M. Gyssens (1995): *A unifying framework for tractable constraints*. In: Proceedings of the First International Conference on Principles and Practice of Constraint Programming, LNCS, Springer, pp. 276–291, doi:10.1007/3-540-60299-2_17.

[9] P.G. Jeavons & M.C. Cooper (1995): *Tractable constraints on ordered domains*. Artificial Intelligence 79(2), pp. 327–339, doi:10.1016/0004-3702(95)00107-7.

[10] Elon Kohlberg & Jean-Francois Mertens (1986): *On the Strategic Stability of Equilibria*. Econometrica 54(5), pp. 1003–1037. Available at http://www.jstor.org/stable/1912320.

[11] A. Malcev (1936): *Untersuchungen aus dem Gebiete der mathematischen Logik*. Recueil Mathématique 1(43)(3), pp. 323–336.

[12] D. Monderer & M. Tennenholtz (2004): *K-Implementation*. Journal of Artificial Intelligence Research 21, pp. 37–62, doi:10.5555/1622467.1622470.

[13] U. Montanari (1974): *Networks of constraints: Fundamental properties and applications to picture processing*. Information sciences 7, pp. 95–132, doi:10.1016/0020-0255(74)90008-5.

[14] C. Oesterheld & V. Conitzer (2022): *Safe Pareto improvements for delegated game playing*. Autonomous Agents and Multi-Agent Systems 36(2), p. 46, doi:10.1007/s10458-022-09574-6.

[15] D.G. Pearce (1984): *Rationalizable Strategic Behavior and the Problem of Perfection*. Econometrica 54(4), pp. 1029–1050, doi:10.2307/1911197.

[16] N. Sauerberg & C. Oesterheld (2025): *Promises Made, Promises Kept: Safe Pareto Improvements via Ex Post Verifiable Commitments*. Available at https://arxiv.org/abs/2505.00783.

[17] T.C. Schelling (1960): *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

[18] J.K. Truss (1997): *Foundations of Mathematical Analysis*. Clarendon Press, Oxford.

A Completeness and incompleteness of inference as per Lemma 1 and computational complexity under natural generalizations of max-closedness

The condition of max-closedness is based on a complete order on each domain. What happens if instead we consider a closedness condition based on partial orders? Of course, to define the closedness condition

we at least need something like the max of two elements to be well-defined. Therefore, a natural generalization of linear orders is to consider *join lattices*, i.e., partial orders under which every two elements have a least upper bound (a join). Instead of defining join closedness based on the underlying semi-lattice, we (equivalently) define join-closedness directly using the least upper bound operator ($\sqcup$). This is sometimes called the *algebraic* approach to lattices. Specifically, we say that $\sqcup\colon M \times M \to M$ is a *join* operator on a set M if for $x_a, x_b, x_c \in M$ we have: $\sqcup(\sqcup(x_a, x_b), x_c) = \sqcup(x_a, \sqcup(x_b, x_c))$ (Associativity); $\sqcup(x_a, x_b) = \sqcup(x_b, x_a)$ (Commutativity); $\sqcup(x_a, x_a) = x_a$ (Idempotency).

Definition 3 ([8])**.** *We say that a BCS $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ is (semi-lattice) join-closed if there is a family of join operators $(\sqcup_X\colon D_X \times D_X \to D_X)_{X \in \mathscr{X}}$ on the domains of the variables s.t. for every constraint $\Phi \in \mathscr{A}$ between X and Y, we have that if $(x_a, y_a), (x_b, y_b) \in \Phi$, then $(\sqcup_X(x_a, x_b), \sqcup_Y(y_a, y_b)) \in \Phi$.*

A.1 Refutation completeness of inference as per Lemma 1

Jeavons et al. [8] show that under join-closedness, *satisfiability* of a given constraint structure can be decided in polynomial time. (This result implies that *inference* can also be solved in polynomial time, see Appendix A.3.) In fact, their proof shows that inference as per Lemma 1 is sufficient to decide satisfiability. For convenience, we reprove this result here.

Theorem 9 ([8, Theorem 16])**.** *Let $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ be semi-lattice-join-closed binary-constraint structure. Then the following two are equivalent: 1. We can infer an empty outcome correspondence using the rules of Lemma 1. 2. $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ is unsatisfiable.*

Theorem 9 can be viewed as stating a kind of *refutation completeness* (similar to the refutation completeness of the resolution rule). To prove Theorem 9 we use the following lemma (which is the analogue of Lemma 18 for semi-lattice join closedness).

Lemma 10. *Let $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ be semi-lattice join-closed. Let $(\mathscr{X}, \mathscr{D}, \mathscr{A}^+)$ result from application of Lemma 1 to $(\mathscr{X}, \mathscr{D}, \mathscr{A})$. Then $(\mathscr{X}, \mathscr{D}, \mathscr{A}^+)$ is semi-lattice join-closed.*

Proof of Lemma 10. Lemma 1.3 to 5 are easy. So we only consider the transitivity and intersection rules.
Transitivity: Let $\Gamma \sim_\Phi \Gamma'$ and $\Gamma' \sim_\Psi \hat{\Gamma}$. Then consider the outcome correspondence $\Gamma \sim_{\Psi \circ \Phi} \hat{\Gamma}$. We need to show that whenever $\hat{o}^a \in (\Psi \circ \Phi)(o^a)$ and $\hat{o}^b \in (\Psi \circ \Phi)(o^b)$, then $\sqcup(\hat{o}^a, \hat{o}^b) \in (\Psi \circ \Phi)(\sqcup(o^a, o^b))$. To prove this, note first that if $\hat{o}^a \in (\Psi \circ \Phi)(o^a)$, then there is o'^a in Γ' s.t. $o'^a \in \Phi(o^a)$ and $\hat{o}^a \in \Phi(o'^a)$. There must also be an analogous o'^b. Now by max-closedness we have that $\sqcup(o'^a, o'^b) \in \Phi(\sqcup(o^a, o^b))$, and $\sqcup(\hat{o}^a, \hat{o}^b) \in \Psi(\sqcup(o'^a, o'^b))$. It follows (from the definition of composition) that $\sqcup(\hat{o}^a, \hat{o}^b) \in (\Psi \circ \Phi)(\sqcup(o^a, o^b))$.

Intersection: Let $\Gamma \sim_\Phi \Gamma'$ and $\Gamma \sim_\Psi \Gamma'$. Consider the outcome correspondence $\Gamma \sim_{\Psi \cap \Phi} \Gamma'$. We need to show that whenever $o'^a \in (\Psi \cap \Phi)(o^a)$ and $o'^b \in (\Psi \cap \Phi)(o^b)$, then $\sqcup(o'^a, o'^b) \in (\Psi \cap \Phi)(\sqcup(o^a, o^b))$. Note that if $o'^a \in (\Psi \cap \Phi)(o^a)$, then in particular $o'^a \in \Psi(o^a)$. Similarly, $o'^b \in \Psi(o^b)$. Thus, by max closure of Ψ, we have that $\sqcup(o'^a, o'^b) \in \Psi(\sqcup(o^a, o^b))$. Analogously, we get that $\sqcup(o'^a, o'^b) \in \Phi(\sqcup(o^a, o^b))$. It follows that $\sqcup(o'^a, o'^b) \in (\Psi \cap \Phi)(\sqcup(o^a, o^b))$. □

Proof of Theorem 9. The implication from 1 to 2 is trivial. We have left to prove the other direction. We prove this by proving its contrapositive: If repeated application of Lemma 1 does not yield an empty outcome correspondence, then $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ is satisfiable.

So let $(\Psi^{i,j})_{i,j}$ be the minimal outcome correspondences between the variables in $\mathscr{X}$ that can be inferred by repeated application of Lemma 1. (Because of the intersection rule, there are unique minimal inferable outcome correspondences.) Then we construct a satisfying assignment as follows. First assign to X_1 be the largest element o_1 in the domain of X_1 s.t. $\Psi^{1,1}(o_1) \neq \emptyset$. (There is a unique maximal such element since for every two $\hat{o}_1, \tilde{o}_1$ with this property, $\sqcup(\hat{o}_1, \tilde{o}_1)$ also has this property by Lemma 10.)

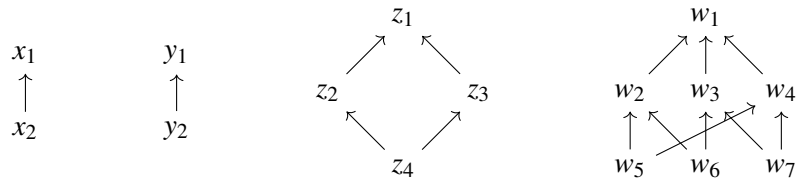


Figure 4: Semilattices for the example for the proof of Proposition 11

For $i \geq 2$ choose o_i to be the maximal element of $\Psi^{1,i}(o_1)$. Note that $\Psi^{1,i}(o_1)$ is nonempty, since otherwise one can infer using Lemma 1 that we cannot assign o_1 to X_1 and so it would have to be $\Psi^{1,1}(o_1) = \emptyset$. Also the maximal element is unique because for any two elements $\tilde{o}_i, \hat{o}^i \in \Psi^{1,i}(o_1)$, we have that $\sqcup(\tilde{o}_i, \hat{o}^i) \in \Psi^{1,i}(o_1)$ by Lemma 10.

We have left to show that the assignment of o_i thus constructed satisfies $(\mathscr{X}, \mathscr{D}, \mathscr{A})$. That is, for every outcome correspondence $X_i \sim_\Phi X_j$ we need to show that $o_j \in \Phi(o_i)$. Notice that since $o_j \in \Psi^{1,j}(o_1)$, there has to be some $o'_i \in \Psi^{1,i}(o_1)$ s.t. $o'_i \in \Psi^{j,i}(o_j)$ for X_i. After all, if this weren't the case, we could apply the transitivity rule applied to $X_1 \sim_{\Psi^{1,i}} X_i \sim_{\Psi^{i,j}} X_j$ to infer that o_j is not compatible with o_1. Further, by choice of o_i we must have $o'_i \leq o_i$. By an analogous argument there must be $o'_j \leq o_j$ for X_j s.t. $o'_j \in \Psi^{j,i}(o_i)$. By join-closedness we have that $o_i \in \Psi^{j,i}(o_j)$ and therefore $o_j \in \Phi(o_i)$. □

A.2 Incompleteness of Lemma 1 under join-closedness and similar conditions

Unfortunately, inference as per Lemma 1 alone is *not* sufficient for solving the kinds of inference problems we are concerned with in this paper, i.e., for deciding whether all satisfying assignments of a given constraint structure also satisfy some additional outcome correspondence.

Proposition 11. *There exists a join-closed BCS in which there exists an SPI, but the SPI cannot be inferred using repeated application of the rules of Lemma 1.*

Proof. We prove this using the following example. We take four variables X, Y, Z, W. In Figure 4, we give the domains, as well as join lattices on the domain, given by the transitive closures of the graphs in the figure. The operators $\sqcup_X, \sqcup_Y, \sqcup_Z, \sqcup_W$ are defined as the least upper bounds on these join-lattices (e.g., $\sqcup_W(w_5, w_7) = w_4$, $\sqcup_W(w_7, w_2) = w_1$). It is easy to verify that these are indeed join lattices.

The outcome correspondence assumptions are given as follows: Between X, Z, assume that the following outcome correspondence holds: $x_2 \mapsto \{z_2, z_4\}$, $x_1 \mapsto \{z_1, z_2, z_3, z_4\}$. Between Y, Z assume $y_2 \mapsto \{z_3, z_4\}$, $y_1 \mapsto \{z_1, z_2, z_3, z_4\}$. Between X, W assume $x_1 \mapsto \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$, $x_2 \mapsto \{w_2, w_5, w_6\}$. Between Y, W assume $y_1 \mapsto \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$, $y_2 \mapsto \{w_3, w_6, w_7\}$. Between Z, W, assume $z_4 \mapsto \{w_7, w_5, w_4\}$, and for $i = 1, 2, 3$: $z_i \mapsto \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$.

It is easy to verify that no progress can be made by applying Lemma 1 to the assumptions. In particular, we can't infer an outcome correspondence between X and Y that disallows the pair (x_2, y_2). (To verify this, one needs to go through all the different triplets of variables and apply the transitivity rule to see whether any pairs of outcomes that are permitted by the above can be excluded from consideration.)

However, we can see as follows that the OC assumptions imply that no satisfying assignment can assign x_2 to X and Y to y_2. To prove this, assume for contradiction that both x_2 and y_2 are part of some satisfying assignment. By the OCs between X and Z and Y and Z, we can infer that in the assignment the value z_4 must be assigned to Z. Finally, we can see that the intersection of the sets values corresponding to x_2, y_2 and z_4 is empty. (Note that the *pairwise* intersections are all non-empty.) □

A.3 Inference under join-closedness can be solved in polynomial time

Repeated inference as per Lemma 1 alone is insufficient for inferring all true outcome correspondences under join-closedness. However, because the rules of Lemma 1 are refutation complete (complete for deciding satisfiability), we can still use them to decide inference questions in polynomial time. Roughly, to decide whether $X \sim_\Phi Y$ follows from $(\mathscr{X}, \mathscr{D}, \mathscr{A})$, we need to add $X \sim_{\Phi^C} Y$ to $(\mathscr{X}, \mathscr{D}, \mathscr{A})$, where $\Phi^C = D_X \times D_Y - \Phi$ is the complement of Φ, and then perform inference as per Lemma 1 to decide whether $X \sim_{\Phi^C} Y$ is satisfiable in $(\mathscr{X}, \mathscr{D}, \mathscr{A})$. We thus obtain the following result.

Theorem 12. *Given semi-lattice-join-closed $(\mathscr{X}, \mathscr{D}, \mathscr{A})$ and preferences $\succeq$, the following problems are solvable in polynomial time. (1) Decide whether there exist $X, X' \in \mathscr{X}$ s.t. $X \neq X'$ and X' is an SI on X. (2) Given a variable $X \in \mathscr{X}$, decide whether $\mathscr{X} - \{X\}$ contains a variable that is an SI on X. (3) Given two variables X and X' in $\mathscr{X}$, decide whether X' is an SI on X. This continues to hold if we consider the analogous questions for strict SI.*

B Relation between the formalism of Oesterheld et al. [14] and ours

We here relate our formalism to the formalism of Oesterheld et al. [14] in more detail. As described in the main text, the most important differences between the respective setups are the following. We allow arbitrary sets of variables $\mathscr{X}$ or games $\mathscr{G}$ to choose from. In contrast, Oesterheld et al. [14] consider specific sets $\mathscr{G}$, namely for arbitrary given games Γ, the set of "subset games" on Γ (games that arise from removing strategies in Γ, and assigning new utilities to outcomes). Similarly, we allow making general outcome correspondence assumptions $\mathscr{A}$, whereas the prior paper only considers specific outcome correspondence assumptions (roughly Assumptions 1 and 2). Finally, we consider safe improvements w.r.t. arbitrary preferences $\succeq$ over outcomes, whereas Oesterheld et al. [14] consider safe improvement w.r.t. specific preferences, namely the principals' Pareto preferences over outcomes.

The last difference is a little subtle and may confuse comparisons between the two papers. In particular, in the introduction we have considered a safe Pareto improvement where $\succeq$ is defined by the Pareto order of the utilities in the underlying games. That is, if we have two games $(A_1^1, A_2^1, u_1^1, u_2^1)$ and $(A_1^2, A_2^2, u_1^1, u_2^2)$, and we let $(a_1, a_2) \succeq (a_1', a_2')$ if and only if $u_i^1(a_1^1, a_2^2) \geq u_i^2(a_1^2, a_2^2)$ for $i = 1, 2$. In particular, note that $\succeq$ thus defined is thereby based on the agent's utility functions in each of the games in $\mathscr{G}$. In contrast, in Oesterheld et al.'s [14] framework, the original players' (principals') Pareto preferences $\succeq$ typically aren't related to the utility functions that the representatives (agents) act on. Instead they are given by the original game, which is only possible because all games under consideration share the same set of possible outcomes (namely those of the original game). So, the simplest definition of an SPI in our setting for an arbitrary game $\mathscr{G}$ is different from the notion of SPI as per Oesterheld et al. [14].

For further illustration, we now analyze Oesterheld et al.'s [14, Table 1] central example (the Demand Game) in our formalism. We will use A_1 and A_2 to denote the action sets in the Demand Game, and let u_1, u_2 be the utility functions of the Demand Game. The set of games $\mathscr{G}$ is then $\{(A_1^s, A_2^s, u_1^s, u_2^s) \mid A_1^s \subseteq A_1, A_2^s \subseteq A_2, u_1^s, u_2^s \colon A_1^s \times A_2^s \to \mathbb{R}\}$, the (uncountably) infinite set of games that have action sets that are subsets of A_1 and A_2 and that have arbitrary utility functions over outcomes.

We now characterize the principals' preference relation $\succeq$ to match the relevant preference relation of Oesterheld et al. [14]. Note that the games in $\mathscr{G}$ all have outcomes in $A_1 \times A_2$. So our preference relation $\succeq$ is merely over $A_1 \times A_2$. Specifically, for all $a_1', a_1 \in A_1, a_2, a_2' \in A_2$ we let $(a_1', a_2') \succeq (a_1, a_2)$ if and only if $u_i(a_1', a_2') \geq u_i(a_1, a_2)$ for $i = 1, 2$. We can similarly define the strict relation $\succ$.

Note that for many games $(A_1', A_2', u_1', u_2') \in \mathscr{G}$, the preference relation $\succeq$ over $A_1' \times A_2'$ is essentially unrelated to the utility functions u_1' and u_2'. In the context of Oesterheld et al. [14] [as well as, e.g., 2] this

makes perfect sense. However, it is somewhat contrary to some of the examples in our paper in which we have often assumed the principals' preferences to align with those of the agents.

Now in this setting, we ask whether there is a (non-trivial, strict) safe improvement (in $\mathscr{G}$) w.r.t. $\succeq$ on Γ. The answer and the way it is derived are straightforwardly the same as in Oesterheld et al. [14].

C Incompleteness of inference as per Lemma 1

From the co-NP-completeness it follows that unless P = NP, inference rules like the above cannot be complete. That is, if P≠NP, then repeatedly applying inference rules like the above will in some cases not be able to infer constraints that do follow from the given ones. We spell this reasoning out in more detail for inference on outcome correspondence in Appendix E.2. We here give Montanari's [13] example of a binary constraint structure for which the above rules are incomplete.

Proposition 13 ([13])**.** *There exists a BCS $(\mathscr{X},\mathscr{D},\mathscr{A})$ with $|\mathscr{X}| = 4$ and $|\mathscr{D}_i| = 4$ for $i = 1,...,4$ and a constraint (X_j,X_i,A) such that every satisfying assignment of (X,D,C) also satisfies (X_j,X_i,A) but (X_j,X_i,A) cannot be inferred from (X,D,C) by repeated application of Lemma 1.*

Proof. Consider the problem of 3-coloring a complete graph with four vertices. (That is, $X = \{X_1,X_2,X_3,X_4\}$, $D = \{1,2,3\}$, $C = \{(X_i,X_j,\{(v_i,v_j) \in \{1,2,3\} \times \{1,2,3\} \mid v_i \neq v_j\}$.) This is not satisfiable. So, in particular, every satisfying of this also satisfies, say, $(X_1,X_2,\emptyset)$. However, no inferences can be made using the transitivity or intersection rule.

□

D On the complexity of inference as per Lemma 1 and similar sets of rules

Lemma 14. *Let* INFERENCE *be a function that takes a set of all outcome correspondences between k games as input and outputs new outcome correspondences that are subsets of the old ones. Let* INFERENCE *run in time $\leq f(m)$, where m is the maximum number of outcomes of each game. Then iterated application of* INFERENCE *to a fixed point (see Algorithm 1) can be done in time at most $m^2n^{k+2}f(m)$, where n is the number of games in the input.*

For the algorithm, we use the following notation: For any set $\mathscr{A}$ of assumptions about outcome correspondence and set of games M let $\mathscr{A}_{|M}$ be the set of outcome correspondences $\Gamma \sim_\Phi \Gamma'$ in $\mathscr{G}$.

Algorithm 1 Algorithm for repeated rule-based inference

$\mathscr{A}' \leftarrow \mathscr{A}$
repeat
  progress $\leftarrow$ false
  for each $M \subset \mathscr{G}$ of size k **do**
    $\mathscr{A}^{\text{new}}_M \leftarrow \text{INFERENCE}(M, \mathscr{A}'_{|M})$
    if $\mathscr{A}^{\text{new}}_M \neq \mathscr{A}'_{|M}$ **then**
      progress $\leftarrow$ true
      $\mathscr{A}' \leftarrow (\mathscr{A}' - \mathscr{A}'_{|M}) \cup \mathscr{A}^{\text{new}}_M$
until not progress
return $\mathscr{A}'$

Proof. First we analyze the innermost block of Algorithm 1. By assumption, INFERENCE runs in $f(m)$ steps. Now consider the inner loop (over subsets of size k). There are $\binom{n}{k}$ such subsets, so at each iteration of the outer loop there are $\binom{n}{k}$ iterations of the inner loop. Finally we must consider the outer loop. Note that at every iteration of the outer loop except for the last, at least one outcome correspondence is narrowed. Clearly, such narrowing cannot go on forever (in a discrete, finite space). From this we will obtain an upper bound on the number of iterations of the outer loop as follows. Each element of $\mathscr{A}$ (and $\mathscr{A}'$) can be viewed as a set of pairs of game outcomes. Thus, each element of $\mathscr{A}$ has at most m^2 elements. $\mathscr{A}$ itself meanwhile contains an outcome correspondence for each pair of games. So, $\mathscr{A}$ has n^2 elements. So overall $\mathscr{A}$'s elements contain at most n^2m^2 elements. Each iteration of the outer loop (except for the last) removes at least one of these elements. It follows that the outer loop has at most n^2m^2 iterations. Putting this all together, we have shown that the above algorithm terminates in $n^2m^2\binom{n}{k}f(m)$ steps. □

The inference rules in Lemma 1 can be represented by a function INFERENCE that takes three games as input. Thus, repeated inference as per Lemma 1 can be done in polynomial time using Algorithm 1.

E Incompleteness of inference as per Lemma 1 and similar sets of rules

E.1 Incompleteness of the rules of Lemma 1 for inferring SPIs

In general, the rules of inference in Lemma 1 are incomplete for inferring SPIs.

Proposition 15. *There exists a set of games $\mathscr{G}$, a set $\mathscr{A}$ of assumptions about OC on $\mathscr{G}$ and an OC $\Gamma \sim_\Phi \Gamma'$ between $\Gamma, \Gamma' \in \mathscr{G}$ s.t.: $\mathscr{A}$ is satisfiable; $\mathscr{A}$ includes $\mathscr{A}_*(\mathscr{G})$ (i.e., the assumptions resulting from the application of Assumptions 1 to 3); Φ is Pareto-improving; every satisfying assignment for $\mathscr{A}$ also satisfies $\Gamma \sim_\Phi \Gamma'$; using the rules in Lemma 1, we cannot derive $\Gamma \sim_\Phi \Gamma'$ from $\mathscr{A}$.*

This can be proved by using Montanari's [13] example (Proposition 13) and adapting it with constructions similar to those in the proof of Theorem 3 (see Appendix F.2).

E.2 Generalizing incompleteness to other sets of rules

One might ask whether completeness could be achieved by adding rules similar to those in Lemma 1. Perhaps we have simply omitted some important rule (similar to the way that Oesterheld et al. [14] omitted the intersection rule (Lemma 1.2), which makes inference stronger). Whether this is true depends, of course, on what we are willing to call an inference rule. For instance, we can achieve completeness trivially if we allow inference rules that take as input an arbitrary number of existing outcome correspondences $\mathscr{A}'$ and then determine by brute force which outcome correspondences are satisfied by all satisfying assignments of $\mathscr{A}'$. To obtain a non-trivial result, one must restrict the notion of inference rule. We here consider two restrictions. The first is to restrict attention to rules that take in a bounded-size set of OCs and output a new OC. The following result shows that no such inference rule exists.

Theorem 16 (informal)**.** *There exists no complete set of inference rules that take a bounded number of outcome correspondences as input and infer a new valid outcome correspondence if there exists one.*

Proof sketch. We can consider generalized versions of the counterexample in the proof of Proposition 13 with n variables with domains of size $n-1$ and pairwise inequality constraints. No subset of less than n variables permits any further inferences. □

Another notion of an inference rule is that it runs in polynomial time to narrow down the OCs within *any* set of variables (if possible). Again, we get that no such rule can exist, assuming P≠NP, because otherwise we would obtain a polynomial-time algorithm (as in Lemma 14) for deciding the satisfiability of binary constraint structures.

F Proofs of results in the main text

F.1 Proof of Lemma 1

Proof. We here prove the intersection rule (Item 2). For proofs of the other claims, refer to Oesterheld et al. [14, Lemma 2]. By definition of $\sim$, $X \sim_\Psi Y$ means that if the agents play (a_1, a_2) in X, they play an element of $\Psi(a_1, a_2)$ in Y. Analogously for $X \sim_\Xi Y$. Therefore, we have that if the agents play (a_1, a_2) in X, they play an element of $\Psi(a_1, a_2)$ and an element of $\Xi(a_1, a_2)$ in Y. In other words, if the agents play (a_1, a_2) in X, they play an element of $\Psi(a_1, a_2) \cap \Xi(a_1, a_2)$ in Y. Thus, $X \sim_{\Psi \cap \Xi} Y$. □

F.2 Proof of Theorem 3

We first prove co-NP-*membership*.

Proposition 17. *The following problems are all in co-NP: Given a finite BCS $(\mathscr{G}, \mathscr{D}, \mathscr{A})$, two games $\Gamma, \Gamma' \in \mathscr{G}$, a preference ordering $\succeq$ over the outcomes of the games in $\mathscr{G}$, decide 1. ... whether Γ' is an SI on Γ. 2. ... whether there exists a game Γ' s.t. $\mathscr{A}$ implies that Γ' is an SI on Γ. 3. ... whether there exist games Γ, Γ' s.t. $\mathscr{A}$ implies that Γ' is an SI on Γ. This continues to hold if we require the SIs to be strict.*

Proof. We here prove co-NP-membership of the first problem. All other claims are proved analogously. To prove membership in co-NP, we need to show membership in NP of the complement. The complement of the problem is as follows: Decide whether there exists an assignment of outcomes to games in $\mathscr{G}$ that satisfies $\mathscr{A}$ but that assigns outcomes o to Γ and o' to Γ' s.t. $o' \not\succeq o$. Clearly this problem is in NP using the satisfying assignment as a witness. □

Proof of Theorem 3. First consider the case of a comparison between a single given pair of games. To show that this problem is co-NP-hard, we show that its complement is NP-hard. That is, we show that the following problem is NP-hard: Given $\mathscr{G}, \mathscr{A}, \succeq$ that satisfy the assumptions and given $\Gamma, \Gamma' \in \mathscr{G}$, decide whether Γ' is *not* an SPI on Γ, i.e., whether there is a satisfying assignment of $\mathscr{G}, \mathscr{A}$ s.t. $o' \not\succeq o$, where o, o' are the outcomes assigned to Γ, Γ', respectively. To prove this, we reduce from BCSP, which is NP-hard.

So let $(\mathscr{X} = \{X_1, ..., X_n\}, \mathscr{D} = (D_i)_{i=1,...,n}, \mathscr{A})$ be any BCS. Then construct the following: $\mathscr{G} = \{\Gamma, \Gamma', \Gamma_1, ..., \Gamma_n\}$, where Γ is as in Table 2 and Γ' is a game with a single outcome with utilities $3, 2$. For $i = 1, ..., n$, construct Γ_i as follows. Let X_i have k_i outcomes named $o^i_1, ..., o^i_k$. Then Γ_i is a $(k+1)$-by-$(k+1)$ game as per Figure 2 with $y_{1,1} > y_{2,1} > ... > y_{k+1,1}$ and $y_{1,2} < y_{2,2} < ... < y_{k+1,2}$. Further, ensure with the choice of the y variables that there are no isomorphisms between any pair of games with more than one outcome. For any outcome correspondence Φ in $\mathscr{A}$ between games Γ_i, Γ_j, we add an outcome correspondence $\Phi' = \{((a_{k_i+1}, a_{k_i+1}), (a_{k_j+1}, a_{k_j+1}))\} \cup \{((a_{l_i}, a_{l_i}), (a_{l_j}, a_{l_j})) \mid (o^i_{l_i}, o^j_{l_j}) \in \Phi\}$. Further, for $i = 1, ..., n$ add the following outcome correspondence between Γ and Γ_i: $\{((a_1, a_1), (a_{l_i}, a_{l_i})) \mid l_i = 1, ..., k_i\} \cup \{((a_2, a_2), (a_{k_i+1}, a_{k_i+1}))\}$ and the following outcome correspondence between Γ' and Γ_i: $\{((a_1, a_1), (a_{l_i}, a_{l_i})) \mid l_i = 1, ..., k_i + 1\}$. Finally we assume the following between Γ, Γ': $\{((a_1, a_1), (a_1, a_1)), ((a_1, a_1), (a_2, a_2))\}$. It's easy to see that the above satisfies all the given requirements.

		Player 2	
		a_1	a_2
Player 1	a_1	$4,0$	$0,0$
	a_2	$0,0$	$2,1$

Table 2: Game Γ used in the proof of Theorem 3

Now it is easy to see that Γ' is *not* SPI on Γ if and only if $(\mathscr{X},\mathscr{D},\mathscr{A})$ is satisfiable. To see this notice first that Γ' is *not* an SPI on Γ if and only if (a_1,a_1) can occur in Γ. But for (a_1,a_1) to possible in Γ, there must be a satisfying assignment of the Γ_i consisting only of outcomes (a_l,a_l) with $l \in \{1,...,k_i\}$ (i.e., $l \neq k_i+1$). It's easy to see that such an assignment would induce a valid assignment for $(\mathscr{X},\mathscr{D},\mathscr{A})$.

This concludes our proof of the co-NP-hardness of the problem for given Γ,Γ'.

We now describe how to extend the proof to also show the co-NP-hardness of the other two problems. First consider the problem of deciding whether *any* game in $\mathscr{G}$ is an SPI on *any* other game in $\mathscr{G}$. To prove the co-NP-completeness, we modify the construction to ensure that the only games that could possibly be in an SPI relation to each other are Γ and Γ'. To do this, all we need to ensure is that the utilities of the outcomes (a_{k_i+1},a_{k_i+1}) in Γ^i are Pareto-incompatible with each other and with $(2,1)$ and $(3,2)$. So for instance, these utilities could be $(3+\varepsilon,1-\varepsilon),(3+2\varepsilon,1-2\varepsilon),...$ and so on for some small ε. After, all these outcomes form one valid assignment in our construction. Thus, if these outcomes of any two games are incomparable, then the games are incomparable in terms of the SPI relation.

It is easy to show that the same construction also shows that it's co-NP-hard to decide whether there is an SPI on some given Γ. □

F.3 Proof of Theorem 4

We will use a couple of lemmas.

Lemma 18. *If $\mathscr{A}$ is max-closed, then minimal $\Psi^{i,j}$ as per Lemma 1 is max-closed.*

Since this proof is easy and we prove a similar result as Lemma 10, we omit a proof here.

Lemma 19. *Let Ψs satisfy max-closedness. Then let $a \in \Gamma_1$. Let h_2 be the max in Γ_2 that's compatible with a. Let h_3 be the max in Γ_3 that's compatible with a. Then h_2 and h_3 must be compatible with each other.*

Proof. h_2 must be compatible with something in Γ_3 that is compatible with a. (Otherwise, we could exclude h_2 from $\Psi^{1,2}(a)$ via Γ_3.) Therefore h_2 must be compatible with something in Γ_3 that's $\leq h_3$. Similarly, h_3 is compatible with something in Γ_2 that's $\leq h_2$. From max-closedness it follows that h_2 is compatible with h_3. □

Proof of Theorem 4. Let $\Psi^{i,j}$ be the minimal inferred OC between games G_i,G_j. Then we need to show that whenever $b \in \Psi^{i,j}(a)$, there is in fact a satisfying assignment with $\Gamma_i \leftarrow a$ and $\Gamma_j \leftarrow b$. We prove this by construction. WLOG, let $i=1$ and $j=2$. For $k=3,4,...$, assign to Γ_k the minimum of the max'es of the sets corresponding to a and b in Γ_k, i.e., the minimum of $h^k_a := \max(\Psi^{1,k}(a))$ and $h^k_b := \max(\Psi^{2,k}(b))$. We will need to show that this assignment satisfies all the constraints.

First, let's prove that these are all compatible with a and b. Note that for this it is sufficient to show that $\Psi^{1,k}(a)$ and $\Psi^{2,k}(b)$ are not disjoint. This must clearly be true, because if they were disjoint, then we

could use the rules of Lemma 1 to infer that a and b are incompatible. In particular, $(\Psi^{2,k})^{-1}(\Psi^{1,k}(a))$ will not contain b.

Now, we need to show that for $k,l \in \{2,3,...\}$, $\min(h_a^k, h_b^k)$ and $\min(h_a^l, h_b^l)$ are compatible with each other. Note first that by Lemma 19, h_a^k and h_a^l are compatible with each other. Thus, if $\min(h_a^k, h_b^k) = h_a^k$ and $\min(h_a^l, h_b^l) = h_a^l$, it follows immediately $\min(h_a^k, h_b^k)$ and $\min(h_a^l, h_b^l)$ are compatible. The case of $\min(h_a^k, h_b^k) = h_b^k$ and $\min(h_a^l, h_b^l) = h_b^l$ is covered analogously.

We are left to cover the case where (WLOG) $\min(h_a^k, h_b^k) = h_a^k$ and $\min(h_a^l, h_b^l) = h_b^l$, i.e., $h_a^k \leq h_b^k$ and $h_b^l \leq h_a^l$. With the above facts about compatibility, we can apply max-closedness to obtain that h_b^k and h_a^l are compatible. Thus $\min(h_a^k, h_b^k)$ and $\min(h_a^l, h_b^l)$ are also compatible with each other. □

F.4 Proof of Proposition 5

Proof. Note that Assumption 3 is always max-closed w.r.t. any ordering. We can therefore ignore them in the following.

Consider a single pair of games and their OC as per Assumption 4. This OC is max-closed under the following ordering: $(a_H, a_H) >_\Gamma (a_L, a_L) >_\Gamma (a_L, a_H) >_\Gamma (a_H, a_L)$ and similarly $(\hat{a}_H, \hat{a}_H) >_{\hat{\Gamma}} (\hat{a}_L, \hat{a}_L) >_{\hat{\Gamma}} (\hat{a}_L, \hat{a}_H) >_{\hat{\Gamma}} (\hat{a}_H, \hat{a}_L)$. Thus we order all games that are subject to Assumption 4 in this way. Further order any any game that's isomorphic one of these games to be isomorphically ordered.

Now let's construct the orderings of all other fully reduced games. We do this iteratively. Pick a game Γ that does not yet have an order. Now partition the outcomes according to the following equivalence relation: $\boldsymbol{a}$ and $\boldsymbol{a}'$ are related if for each i there is an isomorphism from Γ to itself that maps a_i onto a'_i. Order the outcomes of Γ in any way that has all the outcomes in one equivalence class appear in a row. Now take any game Γ' that is isomorphic to Γ. Partition Γ' in the same way as Γ. Note that the Assumption 2 OC between Γ and Γ' will map the equivalence classes of Γ 1-to-1 onto equivalence classes of Γ', with each outcome in an equivalence class in Γ being associated with each outcome in the corresponding equivalence class of Γ'. Order the equivalence classes of Γ' in the same way as the equivalence classes of Γ. Repeat this process (of picking an unordered game, ordering it and its isomorphic games) until all fully reduced games are ordered. It's easy to see that the constructed orderings prove max-closedness of the considered isomorphism OCs.

Next consider the non-reduced games, i.e., games that contain strictly dominated strategies. First take each non-reduced game Γ. Note that each game reduces onto a unique fully reduced game Γ' [e.g. 15, 6, 1]. Then let the ordering in Γ coincide with the ordering in Γ' on the shared outcomes, leaving open for now how the remaining orderings fit in. If Γ' is not in $\mathscr{G}$ (and thus doesn't have an ordering) we treat the outcomes in Γ' the same as the eliminable outcomes in Γ.

From all non-reduced games take all outcomes that have not yet been ordered. Order all of these outcomes arbitrarily. Call these ordering $>_*$. Then within each unreduced games, order the so-far-unordered outcomes according to $>_*$ and let them all be smaller than the already-ordered outcomes. □

F.5 Proof of Theorem 7

This result is essentially a *compactness theorem* for BCS. The proof of Theorem 7 in the countable case is given below. The compactness theorem for propositional logic for the uncountable case was proved by Malcev [11, §1] [see also, e.g., 18, Sect. 13.4]. In fact, strictly weaker assumptions than the axiom of choice are sufficient for this result: the ultrafilter lemma or, equivalently, the Boolean prime ideal theorem. Cowen [5] gives a proof of a compactness theorem for constraints satisfaction problems. A relatively accessible proof based on Zorn's lemma (which is equivalent to the axiom of choice) of a compactness

theorem for the r-coloring problem that could easily be transferred to our setting is given by Bukh [3]. We here omit a proof of Theorem 7 for the uncountable case.

Proof for the countable case of Theorem 7. Let M_i be the set of assignments for all of $\mathscr{X}$ satisfying constraints between the constraints between the first i variables. Clearly, $M_1 \supseteq M_2 \supseteq M_3 ...$ Also, let M be the set of assignments that satisfy all the constraints in $\mathscr{A}$. Note that $M = M_1 \cap M_2 \cap ...$ Now let's say that each assignment in M satisfies $X \sim_\Phi X'$. Then we show that there is M_i s.t. every assignment in M_i satisfies $X \sim_\Phi X'$. Assume for contraposition that every M_i contains an assignment that violates $X \sim_\Phi X'$. Then for every n (s.t. X, X' are among the first n variables) there must be an assignment of the first n variables that violates $X \sim_\Phi X'$ and that is allowed by all M_i. (This uses the fact that the domain is finite and the $\supseteq$ relationship between the M_i.) Moreover, if $(o_1, ..., o_n)$ is in all M_i and $m > n$, then there is $o_{n+1}, ..., o_m$ s.t. $(o_1, ..., o_m)$ is in all M_i. Thus, we can construct $o_1, o_2, ...$ that is in all M_i and violates $X \sim_\Phi X'$. It follows that $o_1, o_2, ...$ is in M, i.e., that there is a valid assignment that violates $X \sim_\Phi X'$. □

G Strengthening the complexity result of Oesterheld and Conitzer (2022)

Oesterheld et al. [14] prove a hardness result for SPIs under a specific type of (infinite) set $\mathscr{G}$ and (variants of) Assumptions 1 and 2. However, their result is about deciding whether an SPI can be inferred with Lemma 1. Using our results about the completeness of inference, we can show that in this case inference as per Lemma 1 is equivalent to inference period, and so in particular, we can prove the hardness of the regular inference problem. To keep it brief, we omit the strict versions of the problems and leave details of the setting to the appendix.

Theorem 20. *The following decision problem is NP-complete: Given a game Γ, decide whether there exists a subset game Γ^s of Γ s.t.: 1. (Non-triviality:) If we iteratively remove all strictly dominated actions from Γ and Γ^s, respectively, then the resulting games are not equal. (They are allowed to be isomorphic.) 2. $\Gamma \sim_\Phi \Gamma^s$ is implied by Assumptions 1 and 2. 3. For all $\boldsymbol{a}$ in Γ, we have $\mathbf{u}(\Phi(\boldsymbol{a})) \geq \mathbf{u}(\boldsymbol{a})$.*

Proof. Oesterheld et al.'s [14] proof of their Theorem 9 shows that it's NP-complete to decide whether one can infer using the rules of Lemma 1 that Γ^s is an SPI on Γ. (Note that they use a slightly different version of Assumption 2. But those differences never matter, because the games in their proof aren't action-symmetric, i.e., the relevant isomorphisms are always unique. They also use a somewhat different set of rules Lemma 1. Here, too, it's easy to show that these differences do not matter for the construction in their proof.) From Corollary 6 together with Theorem 7, we know that the question about inference via Lemma 1 is equivalent to the question of inference period. □

Skating System Unveiled: Exploring Preference Aggregation in Ballroom Tournaments

Laryssa Horn Paul Nüsken Jörg Rothe
Tessa Seeger
Institut für Informatik, MNF
Heinrich-Heine-Universität Düsseldorf
Düsseldorf, Germany
{laryssa.horn, paul.nuesken, rothe, tessa.seeger}@hhu.de

The *Skating System*, which originated from the scrutineering system in dance sport tournaments, can be formulated as a voting system: We introduce and formalize the *Skating System Single* (*SkS*, for short), a new voting system embedded into the framework of computational social choice. Although SkS has similarities with Bucklin voting, it differs from it because it is subject to additional constraints when determining the election winners. Through an analysis of the axiomatic properties of SkS and of its vulnerability to manipulative and electoral control attacks, we compare SkS with Bucklin voting and provide insights into its potential strengths and weaknesses. In particular, we show that SkS satisfies nondictatorship as well as the majority criterion, positive responsiveness, monotonicity, and citizens' sovereignty but violates the Condorcet criterion, strong monotonicity, independence of clones, consistency, participation, resoluteness, and strategy-proofness. Further, we study manipulation, i.e., where (groups of) voters vote strategically to improve the outcome of an election in their favor, showing that the constructive coalitional weighted manipulation problem for SkS is NP-complete, while the destructive variant can be solved in polynomial time. Lastly, we initiate the study of electoral control, where an external agent attempts to change the election outcome by interfering with the structure of the election. Here, we show NP-completeness for constructive and destructive control by deleting candidates as well as for constructive control by adding voters, whereas we show that the problem of destructive control by adding voters can be solved in polynomial time.

1 Introduction

Voting plays a fundamental role in collective decision-making and has an abundance of practical usages. One of these applications is winner determination in *ballroom tournaments*. Ballroom dancing is a style of partner dance that originated in the 16th century in the European Royal courts [1, 20] and has since evolved, incorporating influences from various cultures. It is characterized by specific partner dance positions, coordinated movements, and a combination of various dance styles. Today, it continues to be popular worldwide, both as a social activity and a competitive sport with standardized rules and judging criteria. A typical ballroom tournament proceeds in several rounds in which couples dance simultaneously and where the so-called adjudicators cast their votes on the participating couples. After each round, at most half of the participating dancing couples drop out, whereas the remaining couples proceed to the next round. Eventually, the tournament progresses through the quarter-finals and semi-finals to the final. A final consists of at least two and at most six couples competing in multiple dances.[1] The adjudicators evaluate the dance couples in two different ways depending on the phase of the tournament. During all rounds except the final, the adjudicators approve or disapprove the participants. However, they can give

[1] In exceptional cases, the tournament office may allow more than six couples.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 271–285, doi:10.4204/EPTCS.437.23

© L. Horn et al.
This work is licensed under the
Creative Commons Attribution License.

an approval to at most half of the participants. The evaluation format changes in the finals, though, where the adjudicators are asked to cast a strict linear ranking of all participants per dance. This ranking must be created independently for every dance, i.e., having, e.g., four dances in the final, all adjudicators must rank all couples four times.

The intricate and complex rule set the adjudicators must adhere to is known as the *Skating System*, described in detail by Williams [31]. The system's name has its roots in figure skating and was adapted and later used for dance sport tournaments. Records show that the first major competition that used an early form of the Skating System was the British Championship in Blackpool in 1937 (see [31]). The rule set covers the following topics: Instructions to the adjudicators on how to evaluate the participants, winner determination of the single dances, winner determination of the tournament, and additional rules to break possible ties. That is, the Skating System consists of multiple evaluation methods, each producing an output to be used for the next evaluation step. The following example shows how to rank the couples—based on the rankings of the adjudicators—to determine the winner of one dance in a dance tournament final according to the Skating System.

Example 1. *Consider a dance tournament final with a single dance, six competing couples, and five adjudicators—A, B, C, D, and E—each ranking the couples from first (best) to sixth (last) place as shown on the right. Intuitively, couple* 31 *could be winning for having the most first positions, or couples* 32 *or* 34 *for having several of first, second, and third positions. However, the actual winner according to the Skating System is couple* 33*. Couple* 32 *is ranked second,* 34 *is ranked third, followed by couples* 36*, then* 35*, and lastly, couple* 31 *is even in the last position.*

Couples	*A*	*B*	*C*	*D*	*E*
31	1	1	6	6	6
32	4	2	3	3	1
33	2	6	2	4	2
34	3	3	1	2	5
35	6	4	5	6	4
36	5	5	4	1	3

As shown in the example, the Skating System allows for quite unintuitive results, making a formalization and an extensive axiomatic analysis particularly important to either justify or dismiss its usage both in ballroom tournaments and as a general voting rule. Until today, there are only a handful of investigations of the Skating System [22, 23]. Therefore, the Skating System is mostly unknown in the area of COMSOC.

We introduce a new voting rule—called *Skating System Single* (*SkS*, for short)—that corresponds to and formalizes parts of the rule set of the Skating System, namely the winner determination of each dance in the final round. To the best of our knowledge, the Skating System has not been formalized as a social choice function yet. However, such a formalization is necessary to be able to perform axiomatic or complexity-theoretic analyses. Since the rule set of the Skating System is already widely used in real-life tournaments worldwide,[2] we deem it necessary to start such studies. Therefore, we provide an extensive axiomatic analysis of SkS, considering properties such as the *Condorcet criterion*, *majority*, *positive responsiveness*, *(strong) monotonicity*, *independence of clones*, and *participation*.

Moreover, we study the complexity of *manipulation* for SkS: In an election, instead of all voters casting their votes sincerely when ranking the candidates from most to least preferred, some voters might also vote *strategically* to change the election outcome in their favor. That is, based on complete knowledge of all sincere voters' preferences, they might cast ballots that misrepresent their honest preferences with the intention of making a favored candidate the winner of an election (*constructive case*) or prevent a despised candidate from becoming the winner (*destructive case*). A voting system is called *strategy-proof* if no voter can improve the election outcome by voting strategically. However, Gibbard [17] and Satterthwaite [26] showed that every voting system that satisfies certain reasonable axioms like *nondicta-*

[2]Specifically, the Skating System is used by the World DanceSport Federation and the International Olympic Committee.

torship cannot be strategy-proof. Bartholdi et al. [3, 4] introduced and studied constructive manipulation by a single manipulator who aims at making a favored candidate win an unweighted election. Conitzer et al. [7] added the destructive goal, generalized this for coalitions of manipulators and also considered weighted elections, i.e., elections where the contribution of each vote is weighted by some integer. In their book chapter, Conitzer and Walsh [8] survey the computational barriers to manipulation in voting.

Another type of attack on elections, for which we initiate the study of SkS in this work, is *electoral control*, introduced by Bartholdi et al. [5] in the constructive and Hemaspaandra et al. [18] in the destructive case. In control scenarios, an external agent, also called the *(election) chair*, attempts to make a distinguished candidate win (constructive case) or not win (destructive case) an election by interfering with the structure of the election, e.g., by adding, deleting, or partitioning voters and candidates.

If the corresponding decision problem is NP-hard, we say that the voting rule is *resistant* to that kind of attack. Otherwise, if the decision problem can be solved in polynomial time, we call this voting rule *vulnerable* to that attack type.

Related Work. Manipulation and control are also closely related to bribery [13], a third attack type on elections, surveyed together with control by Faliszewski and Rothe [15]. In bribery scenarios, an external agent tries to make a given candidate win an election (in the constructive case) or, not win (in the destructive case) by bribing voters to change their vote within a given budget. The book chapter by Baumeister and Rothe [6] covers all three types of attacks on elections.

The newly introduced SkS is most closely related to the *Bucklin* voting system. For Bucklin voting, manipulation and bribery have been studied by Faliszewski et al. [14] and electoral control by Erdélyi et al. [11, 12]. While the work of Bartholdi et al. [3, 4] marks the very beginnings of the area of COMSOC, this—and, in particular, the study of strategic voting, manipulation, and control of elections—is still a very active field of current research (see, e.g., [9, 10, 21]).

2 Preliminaries

Before we turn to elections, we introduce some general assumptions and notation. For a logical expression P, we define $\mathbb{1}_P$ to be 1 if P is true, and 0 otherwise. For $n \in \mathbb{N} \setminus \{0\}$, we write $[n]$ for the set $\{1,\dots,n\}$. We assume that the reader is familiar with the basic concepts of complexity theory (for more background, the textbooks by see Papadimitriou [24] and Rothe [25]).

An election is a pair (C,V), where $C = \{c_1,\dots,c_m\}$ for $m \in \mathbb{N}$ is a set of candidates and $V = (v_1,\dots,v_n)$ for $n \in \mathbb{N}$ is a list of votes specifying the voters' preferences over the candidates in C. Voters can express their preferences over the candidates in many different ways. As is most common in social choice theory, SkS requires rankings of the candidates, i.e., each vote is given by a (strict) linear order over C. We write $c \succ_v d$ if a voter v prefers candidate c to candidate d, omitting the subscript when v is clear from the context: $c \succ d$. We use $\mathrm{pos}_v(c)$ to denote the position of candidate c in vote v. A voting system maps any given election (C,V) to a subset of C, the *winner(s) of* (C,V).

For $i \in [m]$ and $c \in C$, let

$$\mathrm{score}^i(c) = \sum_{v \in V} \mathbb{1}_{\mathrm{pos}_v(c) \le i}$$

be the number of votes in which c is ranked among the top-i positions in the votes from V. Further, let

$$\mathrm{sum_pos}^i(c) = \sum_{v \in V} \mathrm{pos}_v(c) \cdot \mathbb{1}_{\mathrm{pos}_v(c) \le i}$$

be the sum of positions candidate c achieves up to stage i in the votes V. Only if needed, we use (C,V) as a subscript and explicitly write $\text{score}^i_{(C,V)}(c)$ and $\text{sum_pos}^i_{(C,V)}(c)$.

We now define the well-known *Bucklin voting system* which, given an election (C,V), proceeds in stages until at least one candidate has reached (or exceeded) the *majority threshold* $\text{maj}(V) = \lfloor |V|/2 \rfloor + 1$, i.e., is among at least $\text{maj}(V)$ top-i positions in the votes V. That is, in each stage i, we check if at least one candidate reaches the majority threshold. If none of the candidates reached it in the current stage, we proceed to the next stage. For any candidate $c \in C$ and for any stage $i \in [m]$, with $m = |C|$, the *Bucklin score of c in stage i* is defined by $\text{score}^i_{\text{B}}(c) = \text{score}^i(c)$. Let i^* be the smallest stage in which at least one candidate reaches $\text{maj}(V)$, and return the candidate(s) with the highest Bucklin score in this stage i^* as the (set of) *Bucklin winner(s)*,

Obviously, every Bucklin winner who is already found in the first stage must be a unique winner (i.e., wins alone), and if we were to proceed to the very last stage $m = |C|$, all candidates in C are Bucklin winners. In particular, we always have at least one Bucklin winner.

Example 2. *Let (C,V) be an election with $C = \{U,X,Y,Z\}$ and the following votes shown on the right. In stage $i = 1$, we have* $\text{score}^1_{\text{B}}(U) = 2$ *and* $\text{score}^1_{\text{B}}(c) = 1$ *for $c \in \{X,Y,Z\}$. None of the candidates reaches the majority threshold* $\text{maj}(V) = 3$*; thus we proceed with stage 2 in which both U and X reach this threshold with a score of* 3*. Hence, both U and X are the Bucklin winners of the election.*

Anne:	$U \succ X \succ Y \succ Z$
Bob:	$U \succ X \succ Y \succ Z$
Caro:	$X \succ U \succ Y \succ Z$
Dave:	$Y \succ Z \succ U \succ X$
Ella:	$Z \succ Y \succ U \succ X$

3 Skating System Single

Example 2 shows a Bucklin election with multiple winners. A keen reader may have noticed that, even though both winners share the same Bucklin score, one of the candidates (namely U) is ranked higher than the other in two of the three decisive votes (and four of the five voters in total). While in this example the difference in the sum of positions is only one, we later show in Theorem 1 that this difference can be arbitrarily large. Surely, this makes a strong case that U deserves to win alone, does it not?

In this section, we—first intuitively and then more formally—introduce the *Skating System Single* (*SkS*, for short), a new voting system inspired by the rule set of the Skating System [31]. Just as Bucklin voting, SkS proceeds in stages. While proceeding through each stage, various conditions are checked until a winner is found. Essentially, most of the conditions can be seen as a form of tie-breaking if more than one candidate qualifies as a possible winner. However, even though this intrinsic tie-breaking exists, SkS is not a resolute rule. Before formally defining winner determination, let us first give an overview of the stages in SkS, given an election (C,V) with $m = |C|$ candidates.

Stage 1: Is any candidate $c \in C$ in the top position of at least $\text{maj}(V)$ votes?

Yes: Candidate c is the unique SkS winner.

No: Continue with stage 2.

Stage i, $2 \leq i \leq m$: Are there any candidates $C' \subseteq C$ in the top-i positions of at least $\text{maj}(V)$ votes?

Yes: Let $C^* \subseteq C'$ be the candidates from C' with the highest score (to be formally defined below). Is it true that $|C^*| = 1$?

Yes: This candidate in C^* is the unique SkS winner.

No: Does a single candidate $c \in C^*$ have the lowest sum of top-i positions (to be formally defined below)?

Yes: This candidate c is the unique SkS winner.

No: If $i = m$ (i.e., currently we are already in the last stage), return all candidates C^* as SkS winners.
Otherwise, continue with stage $i+1$, but only with candidates from C^* who tie for the lowest sum of top-i positions.[3]

No: Continue with stage $i+1$ (with all candidates).

Having a first idea of what the complete SkS voting system looks like, we can now start with its formally precise definition.

Again, we use the majority threshold $\mathrm{maj}(V) = \lfloor |V|/2 \rfloor + 1$. Next, we define the scoring functions used by SkS. In each stage, the score of a candidate will consist of two components: First, the number of votes in which the candidate was ranked in the top-i positions and, second, the sum of these positions, allowing for a more refined distinction of candidates. The higher a candidate is ranked in a voter's preference, the less impact this vote has on the sum of positions, i.e., a lower ranking in the top-i positions will increase the sum of positions more than being the top candidate in a vote, so a lower sum of positions is favorable.

Definition 1 (SkS). Let (C,V) be an election with $m = |C|$ candidates. For any candidate c, let

$$i_c^* = \min_{i \in [m]} \{ i \mid \mathrm{score}^i(c) \geq \mathrm{maj}(V) \}$$

be the minimal stage in which c reaches the majority threshold, and let $i^* = \min_{c \in C}(i_c^*)$ be the smallest stage in which at least one candidate reaches the majority threshold. The SkS voting rule looks for the minimal stage $j \in [i^*, m]$ in which one single candidate has the highest score. To this end, consider the following cases for the set $C_j = \arg\max_{c \in C}(\mathrm{score}^j(c))$.

Case $|C_j| = 1$: The sole candidate in C_j is the unique SkS winner.

Case $|C_j| > 1$: SkS compares the sum of positions for all candidates in C_j: $C_j' = \arg\min_{c' \in C_j}(\mathrm{sum_pos}^j(c'))$.

Case $|C_j'| = 1$: The sole candidate in C_j' is the unique SkS winner.

Case $|C_j'| > 1$: If more than one candidate has the lowest sum of positions, we proceed to the next stage $j+1$ and repeat the steps mentioned above with only the candidates in C_j' (i.e., set $C = C_j'$). As soon as $j = m$, all candidates C_j' are the SkS winners.

The following example illustrates how SkS works.

Example 3. *Let (C,V) be an election with the set of candidates $C = \{X,Y,Z\}$ and the voters Anne, Bob, Caro, and Dave having the following preferences: Anne: $X \succ Y \succ Z$, Bob: $Y \succ X \succ Z$, Caro: $Y \succ Z \succ X$, and Dave: $Z \succ X \succ Y$. We have* $\mathrm{maj}(V) = \lfloor 4/2 \rfloor + 1 = 3$. *The stages of the SkS election are depicted in Table 1. Candidates within the top-i positions of the votes in stage i are marked by a ✓. The two rows at the bottom of the tables show the resulting scores.*

Since in the first stage, no candidate receives enough votes to reach the majority threshold, the election proceeds to the second stage. Here, for each candidate, we also show in which position of the given vote the candidate is ranked in case this candidate was ranked within the top-i positions. That is, we write ✓(2), if X is ranked in the second position. In the second stage, both X and Y reach the majority threshold. Since candidate Y has a lower sum of positions, Y is the unique SkS winner.

[3] Proceeding with only the candidates from C^* who tie for the lowest sum of top-i positions instead of all candidates from C^* leads to an interesting interaction. A candidate $c \in C^*$ may be excluded based on the sum of positions, even if c would beat all candidates who tie for the lowest sum of positions by score in the next stage.

Stage 1	Votes	X	Y	Z
Anne:	$X \succ Y \succ Z$	✓	-	-
Bob:	$Y \succ X \succ Z$	-	✓	-
Caro:	$Y \succ Z \succ X$	-	✓	-
Dave:	$Z \succ X \succ Y$	-	-	✓
	$\text{score}^1(c)$	1	2	1
	$\text{sum_pos}^1(c)$	1	2	1

(a) Stage 1 of the election.

Stage 2	Votes	X	Y	Z
Anne:	$X \succ Y \succ Z$	✓(1)	✓(2)	-
Bob:	$Y \succ X \succ Z$	✓(2)	✓(1)	-
Caro:	$Y \succ Z \succ X$	-	✓(1)	✓(2)
Dave:	$Z \succ X \succ Y$	✓(2)	-	✓(1)
	$\text{score}^2(c)$	3	3	2
	$\text{sum_pos}^2(c)$	5	4	3

(b) Stage 2 of the election.

Table 1: Stages of the SkS election in Example 3.

As mentioned above, SkS resembles Bucklin voting: They both proceed in stages and use the majority principle. Therefore, let us compare them in some more detail and discuss their similarities and differences. The first and most obvious similarity is that the first part of the SkS score equals the Bucklin score: $\text{score}^i(c) = \text{score}^i_{\text{B}}(c) = \sum_{v \in V} \mathbb{1}_{\text{pos}_v(c) \leq i}$. Second, the first stage of both voting systems coincides: A candidate c with $\text{score}^1(c) \geq \text{maj}(V)$ wins alone in both SkS and Bucklin voting. Beyond the obvious, unique Bucklin winners reaching the majority threshold at a later stage will also be unique SkS winners. The reverse, however, is not always the case: Even if an SkS winner is unique, there can be multiple Bucklin winners of the election. For example, in the election from Example 3, we have seen that Y is the unique SkS winner, but it is easy to see that both X and Y would be the Bucklin winners. Furthermore, all SkS winners are also Bucklin winners; again, the reverse is not always the case (as we have just seen). We summarize and prove these and other properties below.

Theorem 1. *(i) Every unique Bucklin winner is a unique SkS winner.*

(ii) Every SkS winner is a Bucklin winner.

(iii) For all $k \geq 1$, there is an election with the same k SkS and Bucklin winners.

(iv) The difference in the sum of positions of Bucklin winners can be arbitrarily large in the decisive stage.

Proof. (i) Assume that $c_w \in C$ is the unique Bucklin winner of an election (C,V). Let i^* be the minimal stage in which at least one candidate reaches the majority threshold in Bucklin voting. Clearly, c_w reaches the majority threshold in stage i^*. The Bucklin winner is defined as $\arg\max_{c^* \in C}(\text{score}^{i^*}_{\text{B}}(c^*))$. Since c_w is the sole Bucklin winner, c_w has a greater Bucklin score than all other candidates in stage i^*. The majority threshold is the same for both SkS and Bucklin voting and, thus, we know that c_w also reaches the majority threshold in an SkS election in stage i^* and stage i^* is, again, the smallest stage in which a candidate has reached the majority threshold. By definition of $\text{score}^i(c) = \sum_{v \in V} \mathbb{1}_{\text{pos}_v(c) \leq i} = \text{score}^i_{\text{B}}(c)$, it immediately follows that $\arg\max_{c \in C}(\text{score}^{i^*}(c)) = \{c_w\}$, and therefore, c_w is also the unique SkS winner of (C,V).

(ii) Let (C,V) be an election. First, if (C,V) has a unique Bucklin winner, we know from the first statement that the SkS and Bucklin winners of (C,V) are the same. So assume that (C,V) has two or more Bucklin winners. It remains to show that every SkS winner is also a Bucklin winner in this case. We show this by contradiction. Let C^B_w be the set of Bucklin winners and C^S_w the set of SkS winners of (C,V). For the sake of contradiction, assume that $C^S_w \not\subseteq C^B_w$, i.e., at least one SkS

winner, say c^*, is not a Bucklin winner. Since c^* is an SkS winner, we have $\text{score}^{i^*}(c^*) \geq \text{maj}(V)$, where i^* is the minimal stage in which any candidate reached the majority threshold. Since c^* is not a Bucklin winner, we also know that $\text{score}_{\text{B}}^{i^*}(c^*) < \text{score}_{\text{B}}^{i^*}(c)$ and $\text{score}_{\text{B}}^{i^*}(c) \geq \text{maj}(V)$ for all c in C_w^B. Immediately, we have $\text{score}^{i^*}(c^*) < \text{score}^{i^*}(c)$ for all c in C_w^B. Since $c^* \in C_w^S$, we have $c^* \in \arg\max_{c \in C}(\text{score}^{i^*}(c))$. This is a contradiction.

(iii) For $k = 1$, this directly follows from the first statement. For $k > 1$, construct an election (C,V) as follows. Let $C = \{c_1, c_2, \ldots, c_k\}$. Construct k votes by setting

$$\begin{array}{ccccccccc}
c_1 & \succ & c_2 & \succ & \cdots & \succ & c_{k-1} & \succ & c_k \\
c_2 & \succ & c_3 & \succ & \cdots & \succ & c_k & \succ & c_1 \\
 & & & & \vdots & & & & \\
c_k & \succ & c_1 & \succ & \cdots & \succ & c_{k-2} & \succ & c_{k-1}
\end{array}$$

Due to the cyclic permutation of the votes, each candidate is ranked at each position exactly once and receives exactly one additional point in each stage. Therefore, all candidates reach the majority threshold $\text{maj}(V)$ at the same time: in stage $\frac{k+1}{2}$ for odd k, and in stage $\frac{k}{2}+1$ for even k. Since all candidates tie in their scores, they all are Bucklin winners. For SkS, the positional scores of all candidates in each stage i is $\sum_{\ell \in [i]} \ell$ and, as in Bucklin voting, all candidates tie in each stage. Therefore, SkS proceeds through all stages without excluding any candidate and, finally, declares all of them winners.

(iv) We give a construction for an election in which the difference in the sum of positions between two Bucklin winners can be controlled by setting two integer parameters. Let $i, n \in \mathbb{N}$ be two integers with $i \geq 3$ and $n \geq 0$ and let $C' = \{c_1, \ldots, c_{2n}\}$ and $D = \{d_1, \ldots, d_{n+2}\}$ be two sets of candidates. Construct an election (C,V) with the candidates $C = \{a,b\} \cup C' \cup D$ and the following votes—note that we only depict the first $n+2$ positions of each vote, as the remainder of the vote can be arbitrary:

Vote list V	
$a \succ c_1 \succ \ldots \succ c_n \succ b \succ \ldots$	$\Big\}(i-1)$
$\vdots$	
$a \succ c_1 \succ \ldots \succ c_n \succ b \succ \ldots$	
$b \succ c_{n+1} \succ \ldots \succ c_{2n} \succ a \succ \ldots$	
$d_1 \succ d_2 \succ \ldots \succ d_{n+1} \succ d_{n+2} \succ \ldots$	$\Big\}(i-1)$
$\vdots$	
$d_1 \succ d_2 \succ \ldots \succ d_{n+1} \succ d_{n+2} \succ \ldots$	

We have $\text{maj}(V) = \left\lfloor i - \frac{1}{2} \right\rfloor + 1 = i$. It is clear, that each candidate from C' and D occurs at most $i-1$ times in the top $n+2$ positions. Candidate a is ranked in the first postion a total of $i-1$ times and then once again in position $n+2$, reaching the majority threshold in stage $n+2$. The inverse is true for candidate b, who is ranked only once in the first position and then $i-1$ times in position $n+2$, also reaching the majority threshold in stage $n+2$. Thus both a and b are the Bucklin winners. The difference between the two winners is immediately clear, though: Calculating the

sum of positions in stage $n+2$ for a and b, we have

$$\text{sum_pos}^{n+2}(a) = i-1+n+2,$$
$$\text{sum_pos}^{n+2}(b) = 1+(i-1)(n+2).$$

By increasing either i (and thus the number of votes) or n (and thus the number of candidates) or both, we can boost the difference in the sum of positions between a and b arbitrarily. □

SkS and Bucklin voting differ in their behavior whenever some candidates reach the majority threshold due to the same number of votes. In such cases, Bucklin returns all these candidates as winners, yet SkS instead first compares their sums of positions: If they are the same, it moves on to the next stage with the relevant candidates, repeating the procedure either until one candidate is the sole winner or until the last stage reached with multiple SkS winners. Following this, we state the following lemma on winner determination in SkS. Intuitively, Lemma 1 states that once candidates have tied both in terms of the highest score and lowest sum of positions (while also reaching the majority threshold) leading to a reduced candidate set in the next stage, any unique winner beats all other remaining candidates by a strictly higher score.

Lemma 1. *Let (C,V) be an election and $C' \subseteq C$ some candidates. Assume that, for some stage i, we have $\text{score}^i(c) = \text{score}^i(d) \geq \text{maj}(V)$, $\text{score}^i(c) > \text{score}^i(x)$, and $\text{sum_pos}^i(c) = \text{sum_pos}^i(d)$ for all $c,d \in C', c \neq d$, and $x \in (C \setminus C')$, i.e., SkS proceeds to stage $i+1$ with the reduced candidate set C'. Then it holds that*

(i) for a unique SkS winner c, we have $\text{score}^j(c) > \text{score}^j(d)$ for some $i < j \leq |C|$ and $d \in (C' \setminus \{c\})$;

(ii) if there is no unique SkS winner, we have $\text{score}^j(c) = \text{score}^j(d)$ for all $i < j \leq |C|$ and for some $c \in C', d \in C' \setminus \{c\}$.

Proof. (i) Let SkS be in stage i and assume we have a reduced candidate set $C' \subseteq C$. Since we have a reduced candidate set, all candidates from the reduced set must have tied both the score and sum of positions in the previous round. Assume that c is the first unique SkS winner in stage j with $i < j \leq |C|$ but $\text{score}^j(c) \leq \text{score}^j(d)$ for some $d \in C^* \subseteq (C' \setminus \{c\})$. Since c is a winner there can be no candidate d with a higher score and thus we have $\text{score}^j(c) = \text{score}^j(d)$. Additionally, since c is the unique SkS winner, the sum of positions for c is the uniquely lowest sum of positions among all candidates from C^*. Since all candidates from C^* must have tied both in highest score and sum of positions in stage $j-1$, we have that c either decreased their sum of positions or gained the same increase in score as the other candidates while simultaneously receiving less sum of positions. As both are not possible, we thus have a contradiction.

(ii) This follows directly from the definition of SkS. □

4 Axiomatic Analysis of SkS

In this section, we consider the following axiomatic properties of voting systems: the Condorcet criterion, the majority criterion, monotonicity, positive responsiveness, strong monotonicity, independence of irrelevant alternatives, independence of clones, consistency, participation, nondictatorship, citizens' sovereignty, resoluteness, and strategy-proofness. These have been motivated and discussed in depth in the social choice literature and have been investigated for an abundance of voting systems (see, e.g., [2, 6, 30, 32]). In the following, we study which of them are satisfied by SkS and which are not.

Theorem 2. *SkS satisfies the majority criterion, positive responsiveness, monotonicity, nondictatorship, and citizens' sovereignty. On the other hand, SkS does not satisfy the Condorcet criterion, strong monotonicity, independence of irrelevant alternatives, independence of clones, consistency, participation, resoluteness, nor strategy-proofness.*

Proof. Due to space limitations, among the axioms satisfied by SkS, we will only present the proofs for monotonicity and positive responsiveness, two closely related, well-known axioms from the social choice literature. Recall that a voting system is *monotonic* if for every election (C,V) and for each candidate $c \in C$, if (i) c wins the election, and (ii) a new election (C,V') is generated from (C,V) by improving the position of c in some votes, leaving the relative rankings of all other candidates unchanged, then c also wins this new election. Further, *positive responsiveness* means that a (possibly tied) winner c turns a unique winner whenever c is raised in some votes, again leaving everything else the same.

Monotonicity: Assume a candidate c wins an SkS election (C,V). Now, we improve the positions of c in some votes, leaving the relative rankings of all other candidates unchanged. Let W be the resulting list of new votes. For the sake of contradiction, assume that c does no longer win the SkS election (C,W). However, since c has only improved the positions in some votes, if c has won (C,V) already in the first stage due to a majority of top positions, c must also win (C,W) in the first stage due to a majority of top positions. If c has won (C,V) at a later stage, let $i_c^* > 1$ be the stage where c meets or exceeds the majority threshold in (C,V). Now, since c has only improved the positions in some votes while none of the other candidates can have a better position in the votes of W than in those of V, it follows that (i) c reaches the majority threshold at a stage $j_c^* \leq i_c^*$ in (C,W) and no other candidate can reach the majority threshold at an earlier stage than in (C,V), (ii) $\text{score}^{j_c^*}_{(C,W)}(c) \geq \text{score}^{i_c^*}_{(C,V)}(c)$, and (iii) $\text{sum_pos}^{j_c^*}_{(C,W)}(c) \leq \text{sum_pos}^{i_c^*}_{(C,V)}(c)$. Hence, c again is an SkS winner in (C,W), a contradiction.

Positive responsiveness: Let (C,V) be an election and let $c,d \in C$ be two SkS winners of (C,V). Recall from Lemma 1 that if multiple winners exist in an SkS election, then once the majority threshold is reached at some stage i, candidates c and d will continue to tie in both score and sum of positions for every following stage. Let V' be a vote list derived from V by improving, without loss of generality, c's position in at least one ballot, while leaving the relative rankings of all other candidates unchanged. Assume that c is not a unique winner of election (C,V'). Then either (i) c does not win at all or (ii) c does win alongside someone else, say $x \in C$. Since SkS satisfies monotonicity, (i) immediately leads to a contradiction, which leaves us with case (ii). Due to c's improvement, there must be some stage j with $1 \leq j \leq |C|$ where $\text{score}^j_{V'}(c) > \text{score}^j_V(c)$ and $\text{sum_pos}^i_{V'}(c) < \text{sum_pos}^i_V(c)$. Note that for any other candidate $y \in C \setminus \{c\}$ and every stage i, we have $\text{score}^i_{V'}(y) \leq \text{score}^i_V(y)$ and $\text{sum_pos}^i_{V'}(y) \geq \text{sum_pos}^i_V(y)$. In particular, x's position did not improve. Let $i' \leq i$ be the stage in which c reaches the majority threshold in (C,V'). If $i' < i$, we have that c reached the majority threshold earlier than in the original election. Since x is also a winner, we have $\text{score}^{i'}_{V'}(c) = \text{score}^{i'}_{V'}(x)$, a contradiction. Lastly, if $i' = i$, due to Lemma 1, we have $\text{score}^j_{V'}(c) = \text{score}^j_{V'}(x)$ and $\text{sum_pos}^j_{V'}(c) = \text{sum_pos}^j_{V'}(y)$ for $i' \leq j \leq |C|$, again a contradiction.

Condorcet criterion, independence of clones, and independence of irrelevant alternatives: Again, due to space limitations, we will present just one counterexample in Table 2 showing that SkS satisfies neither the Condorcet criterion nor independence of clones nor independence of irrelevant alternatives (referring to, e.g., [2, 6, 30, 32] for the formal definitions of these properties). Let $C = \{X,Y,Z\}$ and $C' = \{X,Y,Z,Z'\}$. The list V of votes over C is depicted in Table 2a. Since $|V| = 7$, we have $\text{maj}(V) = 4$. Further, we have $\text{score}^1_{(C,V)}(c) = 2$ for $c \in \{X,Y\}$ and $\text{score}^1_{(C,V)}(Z) = 3$. Since neither candidate reaches the required threshold in stage $i = 1$, we continue with the next stage. Here, we have $\text{score}^2_{(C,V)}(X) = 3$,

Anne:	$X \succ Y \succ Z$
Bob:	$X \succ Y \succ Z$
Caro:	$Y \succ X \succ Z$
Dave:	$Y \succ Z \succ X$
Ella:	$Z \succ Y \succ X$
Fred:	$Z \succ Y \succ X$
Gail:	$Z \succ Y \succ X$

(a) Vote list V of the original election.

Anne:	$X \succ Y \succ Z \succ Z'$
Bob:	$X \succ Y \succ Z \succ Z'$
Caro:	$Y \succ X \succ Z \succ Z'$
Dave:	$Y \succ Z \succ Z' \succ X$
Ella:	$Z \succ Z' \succ Y \succ X$
Fred:	$Z \succ Z' \succ Y \succ X$
Gail:	$Z \succ Z' \succ Y \succ X$

(b) Vote list V' after adding the clone Z'.

Table 2: Vote lists of the two elections in the proof of Theorem 2.

$\text{score}^2_{(C,V)}(Z) = 4$ and $\text{score}^2_{(C,V)}(Y) = 7$. Now, Y is the unique SkS winner because $\text{score}^2_{(C,V)}(c) \geq \text{maj}(V)$ for $c \in \{Y,Z\}$, and $\text{score}^2_{(C,V)}(Y) > \text{score}^2_{(C,V)}(Z)$. Adding a clone Z' of candidate Z yields the election (C',V') depicted in Table 2b. Note that the clone Z' is always ranked in the position behind Z and the rest of the votes is left untouched. The scores in the first stage of the election remain the same. However, in the second stage, the scores change with respect to the election (C,V). That is, the scores of X and Z are unchanged, while Y loses three points to the clone Z'. Now, we have $\text{score}^2_{(C',V')}(c) = 4$ for $c \in \{Y,Z\}$ and $\text{score}^2_{(C',V')}(d) = 3$ for $d \in \{X,Z'\}$. Since $\text{maj}(V') = \text{score}^2_{(C',V')}(Y) = \text{score}^2_{(C',V')}(Z)$, we have to consider the sum of positions for Y and Z to determine the winner. We have $\text{sum_pos}^2_{(C',V')}(Z) = 5 < 6 = \text{sum_pos}^2_{(C',V')}(Y)$. Thus, after the addition of the clone Z', the winner of the SkS election changed from Y to Z, and SkS satisfies neither independence of clones nor independence of irrelevant alternatives. Note that, in the second election, Y is still a Condorcet winner and yet loses the SkS election to Z now. Therefore, SkS does not satisfy the Condorcet criterion. □

As Bucklin and SkS are heavily related, their axiomatic properties often align. Nonetheless, they are not always the same, as shown by positive responsiveness, which SkS satisfies while Bucklin does not.[4] A comprehensive overview of all properties considered here is given in Table 3.

	Condorcet Criterion	Majority	Monotonicity	Positive Responsiveness	Strong Monotonicity	Independence of Irrelevant Alternatives	Independence of Clones	Consistency	Participation	Nondictatorship	Citizens' Sovereignty	Resoluteness	Strategy-Proofness
Bucklin	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
SkS	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗

Table 3: Comparison of axiomatic properties between SkS and Bucklin (see, e.g., [6]).

[4] It is easy to see that Bucklin violates positive responsiveness by a small counterexample. Consider the candidates $C = \{a,b,c\}$ and the votes $a \succ b \succ c \succ d$, $b \succ a \succ c \succ d$, and $c \succ d \succ a \succ b$. Both a and b are the Bucklin winners in stage two. We can now improve b's position in the last vote to $c \succ d \succ b \succ a$ without changing the election result.

5 Complexity of Manipulation in SkS

In this section, we will study the complexity of manipulation in SkS. As mentioned in the introduction, Bartholdi et al. [3, 4] initiated the study of manipulating voting systems in terms of determining the computational complexity of the corresponding decision problems. They considered the most restricted problem variant: constructive manipulation by a single manipulator seeking to make a preferred candidate win an unweighted election.

The most general problem, due to Conitzer et al. [7], considers constructive coalitional weighted manipulation. This means that there is a group (or *coalition*) of manipulators who jointly seek to make their preferred candidate win a weighted election. In a *weighted* election, voters may not all have the same voting power; rather, each voter is assigned a nonnegative integer weight, and if the voting system (as is the case for SkS) uses scores, these weights need to be taken into account when determining the candidates' scores. For SkS, a weighted election is represented as (C,V,W_V), where $W_V = (w_1,\ldots,w_n)$ is a list of nonnegative integer weights with w_i being the weight of the i-th voter. The majority threshold now is $\mathrm{maj}(V) = \lfloor (\Sigma_{v\in V} w_v)/2 \rfloor + 1$, and $\mathrm{score}^i(c)$ for $i \in [m]$ and $c \in C$ now is the sum of the weights of those voters who rank c among their top-i positions, i.e., $\mathrm{score}^i(c) = \sum_{v\in V} \left(\mathbb{1}_{\mathrm{pos}_v(c)\le i} \cdot w_v \right)$. The definition of $\mathrm{sum_pos}^i(c)$ is analogously weighted, i.e., $\mathrm{sum_pos}^i(c) = \sum_{v\in V} \left(\mathrm{pos}_v(c) \cdot \mathbb{1}_{\mathrm{pos}_v(c)\le i} \cdot w_v \right)$, and SkS for weighted elections works as described in Definition 1. Now, we are ready to define the most general manipulation problem for any voting system $\mathscr{E}$:

$\mathscr{E}$-CONSTRUCTIVE-COALITIONAL-WEIGHTED-MANIPULATION ($\mathscr{E}$-CCWM)	
Given:	A set C of candidates, a list V of nonmanipulative votes over C each having a nonnegative integer weight collected in a list W_V, a list W_S of the manipulators' nonnegative integer weights, where the manipulators' votes in S are as yet unspecified, $V \cap S = \emptyset$, and a distinguished candidate $c \in C$.
Question:	Can the votes in S be set such that c is the unique $\mathscr{E}$ winner of the weighted election $(C, V \cup S, W_{V\cup S})$?

The following special variants of $\mathscr{E}$-CCWM have been studied in COMSOC (see, e.g., [8, 6]): $\mathscr{E}$-CONSTRUCTIVE-COALITIONAL-MANIPULATION ($\mathscr{E}$-CCM) where the considered elections are *unweighted* (i.e., every voter has unit weight) and $\mathscr{E}$-CONSTRUCTIVE-MANIPULATION ($\mathscr{E}$-CM) where the considered elections are *unweighted* and there is only a *single manipulator*. Further, when the goal of the manipulation is not to make a distinguished candidate c win alone but to prevent c from becoming the sole winner, we obtain the *destructive* variants of the above three problems: $\mathscr{E}$-DESTRUCTIVE-COALITIONAL-WEIGHTED-MANIPULATION ($\mathscr{E}$-DCWM), $\mathscr{E}$-DESTRUCTIVE-COALITIONAL-MANIPULATION ($\mathscr{E}$-DCM), and $\mathscr{E}$-DESTRUCTIVE-MANIPULATION ($\mathscr{E}$-DM).

We focus on the *unique-winner model*, as defined in the above problems.[5] We expect the reader to be familiar with the notions of NP-hardness and NP-completeness based on the polynomial-time many-one reducibility $\le^{\mathrm{p}}_{\mathrm{m}}$ (see, e.g., [16, 19, 24, 25])

Theorem 3. *SkS-*CCWM *is* NP*-complete, even for instances with only four candidates.*

NP-hardness of SkS-CCWM can be shown by suitably extending the proof of Bucklin-CCWM being NP-hard, which is due to Faliszeswki et al. [14], via a reduction from the well-known NP-complete problem PARTITION: Given a list $(a_1,a_2,\ldots,a_k)$ of positive integers with $\sum_{i=1}^{k} a_i = 2K$, does there exist a subset $A \subseteq [k]$ such that $\sum_{i\in A} a_i = \sum_{i\in[k]\setminus A} a_i = K$? In particular, we have to take care of all cases in which the sum of positions plays a role in the winner determination.

[5] One could also study the *nonunique-winner model* where the constructive goal is merely to make c a winner (possibly among several winners), and the destructive goal is to ensure that c does not win at all.

The complexity of SkS-CCM and SkS-CM remains open.

We now turn to the destructive case and study the complexity of SkS-DCWM. Unlike its constructive variant, this problem is easy to solve, which can be achieved by employing the algorithm of Faliszeswki et al. [14] for Bucklin-DCWM and by appropriately handling the cases in which the sum of positions plays a role for determining the SkS winners. Since DCM is a special case of DCWM and DM in turn is a special case of DCM (see, e.g., [6]), we obviously have: DM $\leq_{\mathrm{m}}^{\mathrm{p}}$ DCM $\leq_{\mathrm{m}}^{\mathrm{p}}$ DCWM.

Theorem 4. *SkS-*DCWM *(and hence, SkS-*DCM *and SkS-*DM*) are in* P.

6 Complexity of Control in SkS

In the previous section, a coalition of manipulators seeked to influenced the election in their favor. In a similar fashion, an external agent (called the *chair*) having the power to change the structure of an election, may exert this control power, again with the goal of making some favored candidate win (or preventing a despised candidate's victory). The study of electoral control was initiated by Bartholdi et al. [5] for the constructive case and by Hemaspaandra et al. [18] for the destructive case. Since then, control has been investigated for many voting systems (see, e.g., [6, 15]), and in particular for Bucklin voting by Erdélyi et al. [11, 12]. Out of the many standard control types from the literature (again, see [6, 15]), we focus on a few select cases, namely on (constructive and destructive) control by either adding voters or deleting candidates, starting with the latter type of control for any voting system $\mathscr{E}$:

$\mathscr{E}$-CONSTRUCTIVE-CONTROL-BY-DELETING-CANDIDATES ($\mathscr{E}$-CCDC)	
Given:	A set C of candidates, a list V of votes over C, a distinguished candidate $c \in C$, and a nonnegative integer k.
Question:	Does there exist a subset of candidates $C' \subseteq C$, $\lvert C' \rvert \leq k$, such that c is the unique $\mathscr{E}$ winner of the election $(C \setminus C', V)$, where the votes in V are now restricted to the candidates in $C \setminus C'$?

$\mathscr{E}$-DESTRUCTIVE-CONTROL-BY-DELETING-CANDIDATES ($\mathscr{E}$-DCDC), the destructive variant of $\mathscr{E}$-CCDC, is defined analogously, except that now the chair's goal is to ensure that the distinguished candidate c is *not* a unique winner of $(C \setminus C', V)$ (i.e., c may still win, but only along with other candidates), and it is not allowed to simply delete c (i.e., we require $c \notin C'$). In $\mathscr{E}$-CONSTRUCTIVE-CONTROL-BY-ADDING-VOTERS ($\mathscr{E}$-CCAV), in addition to the candidates C, the votes V over C, the distinguished candidate $c \in C$, and a nonnegative integer k (now as an addition limit), we are given a list of additional votes U over C of as yet *unregistered* voters, and we ask whether there exists a sublist of votes $U' \subseteq U$ with $\lvert U' \rvert \leq k$ such that c is the unique $\mathscr{E}$ winner of the election $(C, V \cup U')$. Again, the destructive variant of $\mathscr{E}$-CCAV, $\mathscr{E}$-DESTRUCTIVE-CONTROL-BY-ADDING-VOTERS ($\mathscr{E}$-DCAV), is defined analogously, except that the chair's goal now is to prevent the distinguished candidate c from becoming a unique winner. Again, we focus on the unique-winner model.

We now present our results on the above control problems for SkS.

Theorem 5. *SkS-*CCDC*, SkS-*DCDC*, and SkS-*CCAV *are* NP*-complete in the unique-winner model.*

For all of the above control problems we can either directly apply or adapt the proof of the respective control problem for Bucklin voting due to Erdélyi et al. [11].

In contrast, SkS is vulnerable to *destructive* control by adding voters. We note that Bucklin-DCAV is in P as well, i.e., Erdélyi et al. [11] provide a polynomial-time algorithm solving this problem. We know that by Theorem 1, any unique Bucklin winner is also a unique SkS winner and thus any Yes-instance of

Table 4: Overview of complexity results for manipulation and control in SkS. R means resistance and V means vulnerability.

	CWM	CDC	CAV
Constructive (C)	R [Theorem 3]	R [Theorem 5]	R [Theorem 5]
Destructive (D)	V [Theorem 4]	R [Theorem 5]	V [Theorem 6]

Bucklin-DCAV is also a Yes-instance of SkS-DCAV in the unique-winner model. The reverse, however, is not always true, since in the case of multiple Bucklin winners, the despised candidate could still be a unique winner in SkS due to the sum of positions. Therefore, in order to adapt it to SkS, we need to extend the existing algorithm for Bucklin-DCAV so as to take each additional case into consideration.

Theorem 6. *In the unique-winner model, SkS-*DCAV *is in* P.

Due to limited space, we only provide a sketch of the adaptions and the resulting algorithm. Recall that by Theorem 1, a unique Bucklin winner is also a unique SkS winner, and every SkS winner is also a Bucklin winner. Additionally, Bucklin-DCAV is in P for both winner models, as has been shown by Erdélyi et al. [11]. We can use these facts by first running the Bucklin-DCAV algorithms before handling the cases specific to SkS in which the sum of positions comes into play. Essentially, if destructive control by adding voters is possible for Bucklin in the nonunique-winner model (i.e, if the despised candidate is not among the Bucklin winners after this control action), it immediately follows that destructive control by adding voters is possible for SkS in the unique-winner model as well. Moreover, if destructive control by adding voters for Bucklin is impossible in both winner models, it immediately follows that this control action is also impossible for SkS in the unique-winner model.

We now consider the only remaining case where destructive control by adding voters for Bucklin is impossible in the nonunique-winner model but possible in the unique-winner model. That is, no candidate strictly beats the despised candidate, but it is possible to find a list of votes to add such that at least one candidate wins together with the despised candidate as a Bucklin co-winner. For SkS, this is a necessary, yet not a sufficient condition for destructive control by adding voters to be possible in the unique-winner model: If two candidates tie in Bucklin score during some stage, since the despised candidate can still have a lower sum of positions, we cannot guarantee that the despised candidate is not a unique SkS winner. Fortunately, it is possible—even though nontrivial—to check in polynomial time whether the current Bucklin co-winner can actually beat the despised candidate in SkS.

7 Conclusion and Future Work

We have introduced the *Skating System Single*, a voting system that is inspired by the Skating System which is used in ballroom tournaments, i.e., dance competitions. We have studied SkS from an axiomatic point of view and have analyzed the computational complexity of manipulation for it; in addition, we have started the complexity-theoretic analysis of electoral control for SkS (see Tables 3 and 4 for an overview of our results). Since SkS has many similarities with Bucklin voting, it is not surprising that these two voting systems share many of their properties. However, this is not always the case: For the property of positive responsiveness, we show that SkS satisfies it, whereas Bucklin does not. While SkS certainly stands on its own, it can also be viewed as a refinement of Bucklin voting (as demonstrated by Theorem 1 and Lemma 1) and can thus be applied in many general election settings aside from ballroom

tournament finals. Further, the proofs for SkS are technically more complicated due to the tie-breaking rules involving not only scores but also the candidates' sums of positions in the votes.

As to future research, it would be interesting to study further axiomatic properties of SkS (e.g., the *Smith criterion* [29][6]) and to expand our study of attacking SkS to other manipulation and control scenarios as well as to bribery, which have been studied intensively for Bucklin voting [11, 12, 14]. In particular, it would be highly interesting to detect further differences between the results on SkS and those on Bucklin voting.

Acknowledgements. This work was supported in part by Deutsche Forschungsgemeinschaft under DFG research grant RO-1202/21-2 (project 438204498).

References

[1] T. Arbeau (1589): *Orchésographie*. Unknown publisher, Langres, France. A later edition published in 1888 by F. Vieweg, libraire-éditeur, E. Bouillon & E. Vieweg, Successeurs. In French.

[2] K. Arrow, A. Sen & K. Suzumura, editors (2011): *Handbook of Social Choice and Welfare*. 2, North-Holland. Available at `https://shop.elsevier.com/books/handbook-of-social-choice-and-welfare/arrow/978-0-444-50894-2`.

[3] J. Bartholdi III & J. Orlin (1991): *Single Transferable Vote Resists Strategic Voting*. *Social Choice and Welfare* 8(4), pp. 341–354, doi:10.1007/BF00183045.

[4] J. Bartholdi III, C. Tovey & M. Trick (1989): *The Computational Difficulty of Manipulating an Election*. *Social Choice and Welfare* 6(3), pp. 227–241, doi:10.1007/BF00295861.

[5] J. Bartholdi III, C. Tovey & M. Trick (1992): *How Hard Is It to Control an Election?* *Mathematical and Computer Modelling* 16(8/9), pp. 27–40, doi:10.1016/0895-7177(92)90085-Y.

[6] D. Baumeister & J. Rothe (2024): *Preference Aggregation by Voting*. In J. Rothe, editor: *Economics and Computation. An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, 2nd edition, chapter 4, Classroom Companion: Economics, Springer, Berlin and Heidelberg, Germany, pp. 233–367, doi:10.1007/978-3-662-47904-9_4.

[7] V. Conitzer, T. Sandholm & J. Lang (2007): *When Are Elections with Few Candidates Hard to Manipulate?* *Journal of the ACM* 54(3):14, doi:10.1145/1236457.1236461.

[8] V. Conitzer & T. Walsh (2016): *Barriers to Manipulation in Voting*. In F. Brandt, V. Conitzer, U. Endriss, J. Lang & A. Procaccia, editors: *Handbook of Computational Social Choice*, chapter 6, Cambridge University Press, Cambridge, United Kingdom, pp. 127–145, doi:10.1017/CBO9781107446984.007.

[9] A. Eggers & T. Nowacki (2024): *Susceptibility to Strategic Voting: A Comparison of Plurality and Instant-Runoff Elections*. *The Journal of Politics* 86(2), pp. 521–534, doi:10.1086/726943.

[10] E. Elkind, S. Obraztsova & N. Teh (2024): *Temporal Fairness in Multiwinner Voting*. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, AAAI Press, Vancouver, Canada, pp. 22633–22640, doi:10.1609/aaai.v38i20.30273.

[11] G. Erdélyi, M. Fellows, J. Rothe & L. Schend (2015): *Control Complexity in Bucklin and Fallback Voting: A Theoretical Analysis*. *Journal of Computer and System Sciences* 81(4), pp. 632–660, doi:10.1016/j.jcss.2014.11.002.

[12] G. Erdélyi, M. Fellows, J. Rothe & L. Schend (2015): *Control Complexity in Bucklin and Fallback Voting: An Experimental Analysis*. *Journal of Computer and System Sciences* 81(4), pp. 661–670, doi:10.1016/j.jcss.2014.11.003.

[6] Note that the *strict Smith set* is also known as the *Schwartz set* [27, 28].

[13] P. Faliszewski, E. Hemaspaandra & L. Hemaspaandra (2009): *How Hard Is Bribery in Elections?* Journal of Artificial Intelligence Research 35, pp. 485–532, doi:10.1613/jair.2676.

[14] P. Faliszewski, Y. Reisch, J. Rothe & L. Schend (2015): *Complexity of Manipulation, Bribery, and Campaign Management in Bucklin and Fallback Voting*. Journal of Autonomous Agents and Multi-Agent Systems 29(6), pp. 1091–1124, doi:10.1007/s10458-014-9277-x.

[15] P. Faliszewski & J. Rothe (2016): *Control and Bribery in Voting*. In F. Brandt, V. Conitzer, U. Endriss, J. Lang & A. Procaccia, editors: Handbook of Computational Social Choice, chapter 7, Cambridge University Press, Cambridge, United Kingdom, pp. 146–168, doi:10.1017/CBO9781107446984.008.

[16] M. Garey & D. Johnson (1979): *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, doi:10.5555/574848.

[17] A. Gibbard (1973): *Manipulation of Voting Schemes: A General Result*. Econometrica 41(4), pp. 587–601, doi:10.2307/1914083.

[18] E. Hemaspaandra, L. Hemaspaandra & J. Rothe (2007): *Anyone But Him: The Complexity of Precluding an Alternative*. Artificial Intelligence 171(5–6), pp. 255–285, doi:10.1016/j.artint.2007.01.005.

[19] R. Karp (1972): *Reducibility among Combinatorial Problems*. In R. Miller, J. Thatcher & J. Bohlinger, editors: Complexity of Computer Computations, Plenum Press, New York City, NY, USA, pp. 85–103, doi:10.1007/978-1-4684-2001-2_9.

[20] J. Marion (2008): *Ballroom: Culture and Costume in Competitive Dance*. Bloomsbury Publishing, London, United Kingdom, doi:10.2752/9781847887115.

[21] C. Maushagen, D. Niclaus, P. Nüsken, J. Rothe & T. Seeger (2024): *Toward Completing the Picture of Control in Schulze and Ranked Pairs Elections*. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence, ijcai.org, pp. 2940–2948, doi:10.24963/ijcai.2024/326.

[22] X. Mora (2001): *The Skating System*. Universitat Autònoma de Barcelona, Spain. Available from: https://mat.uab.cat/~xmora/escrutini/skating2en.pdf.

[23] X. Mora (2004): *Improving the Skating System II: Methods and Paradoxes from a Broader Perspecive*. Universitat Autònoma de Barcelona, Spain. Available from: https://mat.uab.cat/~xmora/escrutini/iss2en.pdf.

[24] C. Papadimitriou (1995): *Computational Complexity*, second edition. Addison-Wesley, Boston, MA, USA.

[25] J. Rothe (2005): *Complexity Theory and Cryptology. An Introduction to Cryptocomplexity*. EATCS Texts in Theoretical Computer Science, Springer-Verlag, Heidelberg and Berlin, Germany, doi:10.1007/3-540-28520-2.

[26] M. Satterthwaite (1975): *Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions*. Journal of Economic Theory 10(2), pp. 187–217, doi:10.1016/0022-0531(75)90050-2.

[27] T. Schwartz (1970): *On the Possibility of Rational Policy Evaluation*. Theory and Decision 1(1), pp. 89–106, doi:10.1007/BF00132454.

[28] T. Schwartz (1972): *Rationality and the Myth of the Maximum*. Noûs 6(2), pp. 97–117, doi:10.2307/2216143.

[29] J. Smith (1973): *Aggregation of Preferences with Variable Electorate*. Econometrica 41(6), pp. 1027–1041, doi:10.2307/1914033.

[30] N. Tideman (1987): *Independence of Clones as a Criterion for Voting Rules*. Social Choice and Welfare 4(3), pp. 185–206, doi:10.1007/BF00433944.

[31] M. Williams (2018): *The Skating System Study Guide*. Available from: https://dancesport.org.au/accreditation/candidate_info/scrutineering_tutorial.pdf.

[32] W. Zwicker (2016): *Introduction to the Theory of Voting*. In F. Brandt, V. Conitzer, U. Endriss, J. Lang & A. Procaccia, editors: Handbook of Computational Social Choice, chapter 2, Cambridge University Press, Cambridge, United Kingdom, pp. 23–56, doi:10.1017/CBO9781107446984.003.

Logic of (Common or Distributed)* Knowledge

Chenwei Shi*
Department of Philosophy
Tsinghua University
Beijing, China
scw@tsinghua.edu.cn

In this paper, we generalize epistemic logic so that it can help reason about ways of combining common knowledge and distributed knowledge such as "common distributed knowledge", "distributed common knowledge", "distributed common distributed knowledge" and so on. Moreover, we study the logic of its dynamic update by arbitrary reading events. We axiomatize these logics and prove their soundness and completeness.

1 Introduction

In epistemic logic [8], common knowledge and distributed knowledge are two notions of group knowledge which have been studied extensively. In a Kripke model $(W, \{R_a\}_{a\in A}, V)$ where each binary relation R_a respresents one agent a's epistemic state, the distributed knowedge of a group $G \subseteq A$ of agents is represented by $\bigcap_{a\in G} R_a$; the common knowledge of G is represented by the transitive relfexive closure of the union of R_a for $a \in G$, i.e. $(\bigcup_{a\in G} R_a)^*$, the smallest set of pairs which contains $\bigcup_{a\in G} R_a$ and is closed under transitivity and reflexivity, or equivalently, the set of pairs which can be connected by a path of finite steps of relation $\bigcup_{a\in G} R_a$.

A new notion of group knowledge, called "common distributed knowledge", is studied by a recent paper [3]. As the name indicated, this notion combines common knowledge and distributed knowledge somehow. In fact, it is a notion of knowledge for a group of groups of agents. Taking a group of groups $G = \{G_1, \ldots, G_n\}$, the notion of common distributed knowledge is represented by the relation $(\bigcup_{i=1}^{n} \bigcap_{a\in G_i} R_a)^*$. Individual agents' knowledge in the notion of common knowledge is replaced by groups' distributed knowledge, but the definion of common knowledge is kept the same.

The notion "common distributed knowledge" is only one way of combining common knowledge and distributed knowledge. It indicates other possible ways. For example, distributed common knowledge, distributed common distributed knowledge and so on. In this paper, we develop a logic which can help reason about all these ways of combining common knowledge and distributed knowledge.

The multiple iterations of common knowledge and distributed knowledge may sound too complex and abstract to make any sense. The following example hopefully makes it clear how these complex epistemic notions can naturally show up.

Example 1.1. One day, Alice, Bob and Carter are playing card. Their father helps distribute a deck of cards (52) to them. In the current round, Bob and Carter team up, which means that they share one card, in addition to the 25 cards assigned to each of them. They can check the shared card together so that they commonly know what it is. Alice holds the remaining card.

Now according to the rules of the card game they are playing, Alice can choose to check the shared card of Bob and Carter. In return for this, Bob and Carter can check Alice's card.

*This research is supported by Beijing Philosophy and Social Science Foundation (Grant number: 24DTR012)

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 286–300, doi:10.4204/EPTCS.437.24

© C. Shi
This work is licensed under the
Creative Commons Attribution License.

Alice exercises her right and checks the card. Bob and Carter also check Alice's card. In the scenario, the actions of learning are public while what is learned is not (at least to the audience, for example, their father).

Restricting our attention to the information relevant to the current round of the card game, that is, who has what cards, it is not hard to figure out that Bob and Carter commonly know more than what Alice knows after they check each other's card.

On one hand, what Alice knows after she checked the shared card is a simple instance of "distributed common knoledge". It is Alice's common knowledge (for an individual agent, her common knowledge is the same as her knowledge) and Bob and Carter's common knowledge that undergo the operation of distributed knowledge. On the other hand, what Bob and Carter commonly know after they check Alice's card is a simple instance of "common distributed knowledge". It is Bob and Alice's distributed knowledge and Bob and Carter's distributed knowledge that undergo the operation of common knowledge.

So, in this example, there is more common distrubuted knowledge of the group $\{$Bob,Alice$\}$ and the group $\{$Carter,Alice$\}$ than the distributed common knowledge of Alice and the team $\{$Carter,Alice$\}$. But is it always the case?

To have a logic for reasoning about the above type of questions, it is key to observe that the intersection of two equivalence relations/reflexive and transitive relations and the reflexive and transtive closure of the union of two equivalence relations/reflexive and transitive relations can serve as meet and join respectively in a lattice. This observation serves as the cornerstone of our axiomatization of the logic and its dynamic extensions.

In section two, we present the logic where iteration of the operations for defining distributed and common knowledge is allowed. In section three, we extend the logic with a dynamic operator for semi-public reading events. In section four, we consider a more general type of dynamics than semi-public reading events, that is, arbitrary reading events. For the static logic, we provide a sound and complete axiomatization. For both of its dynamic extensions, we find their reduction laws, as usually done in dynamic epistemic logic (c.f. [7] and [2]).

2 Syntax, Semantics and Axiomatization

In this section, we present our logic's syntax, semantics and axiom system.

Definition 2.1 (Language $\mathscr{L}$)**.**

$$\mathsf{T} \ni \tau ::= x \mid \tau + \tau \mid \tau \cdot \tau$$

where $x \in \mathsf{AT}$, a finite set of atomic terms.

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle\tau\rangle\varphi$$

where $p \in \mathsf{AF}$, a countable set of atomic formulas, and $\tau \in \mathsf{T}$. $\wedge$, $\rightarrow$ and $\leftrightarrow$ are defined as usual, and $[\tau]\varphi := \neg\langle\tau\rangle\neg\varphi$.

The simplest form of distributed knowledge/common knowledge is expressed as $[x_1 \cdot x_2]\varphi / [x_1 + x_2]\varphi$ respectively. The common distributed knowledge in example 1.1 is expressed as $[(b \cdot a) + (c \cdot a)]\,\varphi$, while the distributed common knowledge is expressed as $[a \cdot (b + c)]\,\varphi$. It is very hard to read formulas of the form $[\tau]\,\varphi$ in natural language when τ becomes complex. However, it should be not very hard to see how the formal language can help us express those complex ways of combining distributed and common knowledge in an efficient way.

Definition 2.2 (Regular model)**.** A model $\mathbf{M} = (W, \{R_\tau \mid \tau \in \mathsf{T}\}, V)$ is tuple where W is a set of possible states, R_τ is a binary relation on W for each term τ in T and $V : \mathsf{AF} \to \wp(W)$ is a vaulation function. A regular model is a model which satisfies

1. $R_{\sigma\cdot\tau} = R_\sigma \cap R_\tau$
2. $R_{\sigma+\tau} = (R_\sigma \cup R_\tau)^*$

for all $\sigma, \tau \in \mathsf{T}$. When all the relations are equivalence relations/reflexive and transitive relations, a regular model is called an "S5/S4-regular model" respectively.

A key observation is that in an S5/S4-regular model, $(\{R_\tau \subseteq W \times W \mid \tau \in \mathsf{T}\}, \cdot, +)$ where $R_\sigma \cdot R_\tau = R_\sigma \cap R_\tau$ and $R_\sigma + R_\tau = (R_\sigma \cup R_\tau)^*$ consititutes a lattice. That is, the sublattice of the lattice of all equivalence relations/preorders over W which is generated by $\{R_x \mid x \in \mathsf{AT}\}$.

Definition 2.3 (Truth condition)**.** For a model $\mathbf{M}$ and a possible state w in it,

$$\begin{array}{lll} \mathbf{M}, w \models p & \text{iff} & w \in V(p) \\ \mathbf{M}, w \models \neg\varphi & \text{iff} & \mathbf{M}, w \not\models \varphi \\ \mathbf{M}, w \models \varphi \vee \psi & \text{iff} & \mathbf{M}, w \models \varphi \text{ or } \mathbf{M}, w \models \psi \\ \mathbf{M}, w \models \langle\tau\rangle\varphi & \text{iff} & \mathbf{M}, v \models \varphi \text{ for some } v \in R_\tau(w) \end{array}$$

Definition 2.4 (Axiom system $S5\mathsf{LCDK}$)**.**

(CPL)	All tautologies in CPL	
(Rules)	Mpdus Ponens	Necessitation rule
(S5)		
K	$[\tau](\varphi \to \psi) \to ([\tau]\varphi \to [\tau]\psi)$	
T	$[\tau]\varphi \to \varphi$	
4	$[\tau]\varphi \to [\tau][\tau]\varphi$	
5	$\neg[\tau]\varphi \to [\tau]\neg[\tau]\varphi$	
dual	$\langle\tau\rangle\varphi \leftrightarrow \neg[\tau]\neg\varphi$	
(LATTICE)		
Idempotency	$\langle\tau\cdot\tau\rangle\varphi \leftrightarrow \langle\tau\rangle\varphi$	$\langle\tau+\tau\rangle\varphi \leftrightarrow \langle\tau\rangle\varphi$
Commutativity	$\langle\tau\cdot\sigma\rangle\varphi \leftrightarrow \langle\sigma\cdot\tau\rangle\varphi$	$\langle\tau+\sigma\rangle\varphi \leftrightarrow \langle\sigma+\tau\rangle\varphi$
Associativity	$\langle(\rho\cdot\sigma)\cdot\tau\rangle\varphi \leftrightarrow \langle\rho\cdot(\sigma\cdot\tau)\rangle\varphi$	$\langle(\rho+\sigma)+\tau\rangle\varphi \leftrightarrow \langle\rho+(\sigma+\tau)\rangle\varphi$
Absorption	$\langle\tau\cdot(\tau+\sigma)\rangle\varphi \leftrightarrow \langle\tau\rangle\varphi$	$\langle\tau+(\tau\cdot\sigma)\rangle\varphi \leftrightarrow \langle\tau\rangle\varphi$
(STAR)		
FP	$\langle\tau+\sigma\rangle\varphi \leftrightarrow (\varphi \vee (\langle\tau\rangle\langle\tau+\sigma\rangle\varphi \vee \langle\sigma\rangle\langle\tau+\sigma\rangle\varphi))$	
INDUC	$[\tau+\sigma](\varphi \to ([\tau]\varphi \wedge [\sigma]\varphi)) \to (\varphi \to [\tau+\sigma]\varphi)$	

When the axiom 5 is removed from $S5\mathsf{LCDK}$, the system is called $S4\mathsf{LCDK}$.

We define a preorder over terms based on the above axiom system:

$$\tau \leq \sigma \qquad \text{iff}_{def} \qquad \vdash \langle\tau\rangle\varphi \to \langle\sigma\rangle\varphi$$

which will be quite handy in our proof of the next completeness theorem.

Theorem 2.1. *$S5\mathsf{LCDK}$ is sound and weakly complete with repsect to the class of S5-regular frames. $S4\mathsf{LCDK}$ is sound and weakly complete with respect to the class of S4-regular frames.*

The proof makes use of the method of step by step, the details of which can be found in the appendix.

With the help of this sound and complete logic, the question in Exapmle 1.1 can be easily answered. To see this, it is enough to realize that the formal counterpart of the question in our logic is whether $\vdash [(b+c)\cdot a]\varphi \to [b\cdot a + c\cdot a]\varphi$ is the case. By basics of lattice theory, $b\cdot a + c\cdot a \leq (b+c)\cdot a$ is the case. So the answer is yes.

3 Semi-public Reading Events

In this section, we turn to a type of dynamic update proposed in [3] and demonstrate its natural generalization in our new logic. We also show how Example 1.1 can be formalized in this dynamic setting.

As [3] puts it

> We call all these actions *semi-public 'reading' events*. In all of them, some agents get to access ('read') some other agents' knowledge base(s). But *the fact that this access is gained (or not)* is public: it is common knowledge who can 'read' whose knowledge base during these events.

A semi-public 'reading' event is represented by a reading map.

> A *reading map* is a function $\alpha : A \to \wp(A)$, mapping agents $a \in \mathsf{A}$ to sets of agents $\alpha(a) \subseteq \mathsf{AT}$, subject to the constraint that
>
> $$a \in \alpha(a) \quad \text{(for every } a \in A\text{)}.$$
>
> Intuitively, $\alpha(a)$ is *the set of agents whose information is accessed by a* during this action. So this last constraint means that *every agent a can always re-read her own knowledge base*.

In our logic, the reading map can be generalized as follows

$$\beta : \mathsf{AT} \to \mathsf{T} \ .$$

We could have required that

$$\beta(a) \leq a$$

to make sure that every agent a can always re-read her own knowledge base. However, to stay as general as possible, we refrain from this extra condition. According to generalized reading maps, agents can read not only other agents' knowledge but also other more complex forms of knowledge, for example, a group's common knowledge.

Given a reading map β, let $!\beta$ denote a semi-public event corresponding to it. Given any S5/S4-regular model $\mathbf{M} = (W, \{R_\tau \mid \tau \in \mathsf{T}\}, V)$, the event $!\beta$ returns an updated S5/S4-regular model $\mathbf{M}^{!\beta} = (W, \{R_\tau^{!\beta} \mid \tau \in \mathsf{T}\}, V)$ which has the same set of possible states W, the same valuation function V but has new equivalence relations/preorders $R_\tau^{!\beta}$ given by

$$R_\tau^{!\beta} := R_{\beta'(\tau)} \ .$$

where β' is a function mapping terms in T to terms in T which satisfies the following conditions:

1. $\beta'(x) = \beta(x)$ for $x \in \mathsf{AT}$;
2. $\beta'(\tau \cdot \sigma) = \beta'(\tau) \cdot \beta'(\sigma)$;
3. $\beta'(\tau + \sigma) = \beta'(\tau) + \beta'(\sigma)$.

So β' is a homomorphism from T to T which extends β on AT. By this fact, it implies that the updated model is indeed S5/S4-regular.

For each generalized rading map β, we add dynamic modalities $[!\beta]\varphi$ into the syntax of our logic (resulting in a new language $\mathscr{L}_!$) and evaluate them at a state as follows:

$$\mathbf{M}, w \models [!\beta]\varphi \quad \text{iff} \quad \mathbf{M}^{!\beta}, w \models \varphi \ .$$

All the above definitions in this section are straightforward generalization of their counterpart in [3]. They also result in quite straightforward reduction laws:

- $[!\beta]\, p \leftrightarrow p$;
- $[!\beta]\, \neg\varphi \leftrightarrow \neg\, [!\beta]\, \varphi$;
- $[!\beta]\, (\varphi \wedge \psi) \leftrightarrow [!\beta]\, \varphi \wedge [!\beta]\, \psi$;
- $[!\beta]\, [\tau]\, \varphi \leftrightarrow [\beta'(\tau)]\, [!\beta]\, \varphi$.

Putting the above reduction laws together with the axiom system $S5\mathsf{LCDK}/S4\mathsf{LCDK}$, we get the axiom system for the dynamic logic of semi-public reading, $S5\mathsf{LCDK}!/S4\mathsf{LCDK}!$.

Theorem 3.1. $S5\mathsf{LCDK}!/S4\mathsf{LCDK}!$ *is sound and complete with respect to the class of S5/S4-regular frames.*

The soundness is easy to check, from which we get the equal expressivity of $\mathscr{L}$ and $\mathscr{L}_!$ with respect to the class of regular models. So the completeness of $S5\mathsf{LCDK}!$ and $S4\mathsf{LCDK}!$ follows from the completeness of $S5\mathsf{LCDK}$ and $S4\mathsf{LCDK}$ respectively.

In the dynamic logic of semi-public reading, we can formalize the scenario in Example 1.1. Let $a,b,c \in \mathsf{AT}$. Consider a generalized reading map β satisfying $\beta(a) = a \cdot (b+c)$ (Alice semi-publicly learns what Bob and Carter commonly know), $\beta(b) = b \cdot a$ (Bob semi-publicly learns what Alice knows) and $\beta(c) = c \cdot a$ (Carter semi-publicly learns what Alice knows). It does not matter what the value of $\beta(x)$ for $x \neq a,b,c$ is. The question is whether $[!\beta]\, ([a]\, \varphi \to [b+c]\, \varphi)$ is derivable in $S5\mathsf{LCDK}$ or $S4\mathsf{LCDK}$. Applying the reduction laws, we can see that $[!\beta]\, ([a]\, \varphi \to [b+c]\, \varphi)$ is equivalent to

$$[a \cdot (b+c)]\, [!\beta]\, \varphi \to [b \cdot a + c \cdot a]\, [!\beta]\varphi$$

which brings us back to the end of Section 2.

Speaking more generally, the connection between lattice theory and epistemic logic is key to the work in this paper. In the next section, we demonstrate the benefit of this connection by modeling epistemic update by arbitrary reading events.

4 Arbitrary reading events

In a semi-public reading event, an agent's action of reading others' knowledge is public while what she learns is not. In an arbitrary reading event, the action of reading others' knowledge is not necesarily public either.

The uncertainty of an agent about other agents' reading actions can be modelled by epistemic indistinguishability relation between generalized reading maps. Generalizing the definition of *reading event model* in [3], let a *S5 regular reading event model* be a structure $\mathbf{E} = (E, \{R_\tau \mid \tau \in T\}, S)$ where E is a finite set of 'events', $S : E \to F(\mathsf{AT}, \mathsf{T})$ is a function mapping events e in E to a generlized reading map $S(e) \in F(\mathsf{AT}, \mathsf{T})$. R_τ are equivalence relations on E which satisfy the two conditions of regularity as in a regular model and the following condition:

$$eR_x f \text{ implies } [S(e)](x) \equiv [S(f)](x) \ .$$

The symbol $\equiv$ is the abbreviation of $\leq \cap \geq$ where $\leq$ is the preorder we define over the set of terms T at the end of Section 2.

The epistemic update of a regular model $\mathbf{M} = (W, \{R^{\mathbf{M}}_\tau \mid \tau \in \mathsf{T}\}, V^{\mathbf{M}})$ by a S5 regular reading event model $\mathbf{E} = (E, \{R^{\mathbf{E}}_\tau \mid \tau \in T\}, S)$ is defined by a straightforward generalization of the product update proposed in [3]. That is, $\mathbf{M} \otimes \mathbf{E} = (S \times E, \{R^{\mathbf{M}\otimes\mathbf{E}}_\tau \mid \tau \in \mathsf{T}\}, V^{\mathbf{M}\otimes\mathbf{E}})$, where

- $S \times E = \{(w,e) \mid w \in W, e \in E\}$
- $V^{\mathbf{M}\otimes\mathbf{E}} = V^{\mathbf{M}}$
- $R_\tau^{\mathbf{M}\otimes\mathbf{E}}$ is defined recursively :
  - $(w,e)R_x^{\mathbf{M}\otimes\mathbf{E}}(w',e')$ iff $wR^{\mathbf{M}}_{[S(e)](x)}w'$ and $eR_x^{\mathbf{E}}e'$
  - $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1\cdot\sigma_2} := R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1} \cap R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_2}$
  - $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2} := (R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1} \cup R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_2})^*$

The following characterization of $R_\tau^{\mathbf{M}\otimes\mathbf{E}}$ will play a key role in finding reduction laws and proving their validity.

Proposition 4.1. *Given an updated model $\mathbf{M}\otimes\mathbf{E} = (W\times E, \{R_\tau^{\mathbf{M}\otimes\mathbf{E}} \mid \tau \in \mathsf{T}\})$ where $\mathbf{E}$ is an S5 regular reading event model and $\mathbf{M}$ is S5/S4 regular model, let $\gamma : E\times\mathsf{T}\to\mathsf{T}$ be a function defined as follows:*

- $\gamma(e,x) = [S(e)](x)$
- $\gamma(e,\sigma_1\cdot\sigma_2) = \gamma(e,\sigma_1)\cdot\gamma(e,\sigma_2)$
- $\gamma(e,\sigma_1+\sigma_2) = \sum_{eR^{\mathbf{E}}_{\sigma_1+\sigma_2}f}(\gamma(f,\sigma_1)+\gamma(f,\sigma_2))$

We have the following characterization of $R_\tau^{\mathbf{M}\otimes\mathbf{E}}$ in terms of $R^{\mathbf{M}}_{\gamma(e,\tau)}$ and $R^{\mathbf{E}}_\tau$:

$$(w,e)R_\tau^{\mathbf{M}\otimes\mathbf{E}}(w',e') \quad \textit{iff} \quad wR^{\mathbf{M}}_{\gamma(e,\tau)}w' \textit{ and } eR^{\mathbf{E}}_\tau e'$$

Proof. We prove by induction on the structure of τ. The basic case follows directly from the definition of $R_x^{\mathbf{M}\otimes\mathbf{E}}$. The case for $\sigma_1\cdot\sigma_2$ follows directly from the definition of $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1\cdot\sigma_2}$, the inductive hypothesis and the regularity of $\mathbf{M}$ and $\mathbf{E}$.

Now consider the case where τ is of the form $\sigma_1+\sigma_2$.

Assume that $(w,e)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w',e')$. There is a path

$$(w,e)R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_1}(w_1,e_1)R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_2}\ldots R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_n}(w_n,e_n)R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_{n+1}}(w',e') \ .$$

Each $R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_i}$ in the path is either $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1}$ or $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_2}$. So it follows from inductive hypothesis that

$$eR^{\mathbf{E}}_{\tau_1}e_1R^{\mathbf{E}}_{\tau_2}\ldots R^{\mathbf{E}}_{\tau_n}e_nR^{\mathbf{E}}_{\tau_{n+1}}e' \quad\text{and}\quad wR^{\mathbf{M}}_{\gamma(e,\tau_1)}w_1R^{\mathbf{M}}_{\gamma(e_1,\tau_2)}\ldots R^{\mathbf{M}}_{\gamma(e_{n-1},\tau_n)}w_nR^{\mathbf{M}}_{\gamma(e_n,\tau_{n+1})}w' \ .$$

Each τ_i in $R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_i}$ is either σ_1 or σ_2. So $eR^{\mathbf{E}}_{\sigma_1+\sigma_2}e'$ and $wR^{\mathbf{M}}_{\sum_{eR^{\mathbf{E}}_{\sigma_1+\sigma_2}f}(\gamma(f,\sigma_1)+\gamma(f,\sigma_2))}w'$.

Assume that $eR^{\mathbf{E}}_{\sigma_1+\sigma_2}e'$ and $wR^{\mathbf{M}}_{\sum_{eR^{\mathbf{E}}_{\sigma_1+\sigma_2}f}(\gamma(f,\sigma_1)+\gamma(f,\sigma_2))}w'$. So there is a path

$$\bigstar \qquad wR^{\mathbf{M}}_{\gamma(f_0,\tau_1)}w_1R^{\mathbf{M}}_{\gamma(f_1,\tau_2)}\ldots R^{\mathbf{M}}_{\gamma(f_{n-1},\tau_n)}w_nR^{\mathbf{M}}_{\gamma(f_n,\tau_{n+1})}w' \ .$$

Since $eR^{\mathbf{E}}_{\sigma_1+\sigma_2}f_i$ for all f_i in the path, $eR^{\mathbf{E}}_{\sigma_1+\sigma_2}e'$ and $R^{\mathbf{E}}_{\sigma_1+\sigma_2}$ is an equivalence relation, there is a path from f_i to f_{i+1} (when $i=n$, $f_{i+1}=e'$):

$$f_iR^{\mathbf{E}}_{\tau_{i+1}}f_iR^{\mathbf{E}}_{\sigma^i_1}e^i_1R^{\mathbf{E}}_{\sigma^i_2}\ldots e^i_{m(i)}R^{\mathbf{E}}_{\sigma^i_{m(i)+1}}f_{i+1} \ .$$

Each step is $R^{\mathbf{E}}_{\sigma^i_k}$ where σ^i_k is either σ_1 or σ_2. Now, we can expand each step $w_iR^{\mathbf{M}}_{\gamma(f_i,\tau_{i+1})}w_{i+1}$ in the path $\bigstar$ to the following one (when $i=n$, $w_{i+1}=w'$):

$$w_iR^{\mathbf{M}}_{\gamma(f_i,\tau_{i+1})}w_{i+1}R^{\mathbf{M}}_{\gamma(f_i,\sigma^i_1)}w_{i+1}R^{\mathbf{M}}_{\gamma(e^i_1,\sigma^i_2)}w_{i+1}\ldots w_{i+1}R^{\mathbf{M}}_{\gamma(e^i_{m(i)},\sigma^i_{m(i)+1})}w_{i+1} \ .$$

So we get a new path from w to w'. Each step is $R^{\mathbf{M}}_{\gamma(e^i_k,\sigma^i_k)}$ or $R^{\mathbf{M}}_{\gamma(f_i,\tau_{i+1})}$ where σ^i_k and τ_{i+1} are either σ_1 or σ_2. By the inductive hypothesis, we have

$$(w_i,f_i)R^{\mathbf{M}\otimes\mathbf{E}}_{\tau_{i+1}}(w_{i+1},f_i)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma^i_1}(w_{i+1},e^i_1)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma^i_2}(w_{i+1},e^i_2)\dots R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma^i_{m(i)}}(w_{i+1},e^i_{m(i)})R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma^i_{m(i)+1}}(w_{i+1},f_{i+1})$$

which implies that $(w_i,f_i)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w_{i+1},f_{i+i})$ by the definition of $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}$. Therefore, we have

$$(w,f_0)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w_1,f_1)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}\dots R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w_n,f_n)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w',e') \ .$$

Note that f_0 is not necessarily e. However, there is a path from e to f_0 in $\mathbf{E}$ via either $R^{\mathbf{E}}_{\sigma_1}$ or $R^{\mathbf{E}}_{\sigma_2}$:

$$eR^{\mathbf{E}}_{\sigma^*_1}e^*_1R^{\mathbf{E}}_{\sigma^*_2}\dots e^*_{m(*)}R^{\mathbf{E}}_{\sigma^*_{m(*)+1}}f_0$$

and a path from w to w

$$wR^{\mathbf{M}}_{\gamma(e,\sigma^*_1)}wR^{\mathbf{M}}_{\gamma(e^*_1,\sigma_{*_2})}w\dots wR^{\mathbf{M}}_{\gamma(e^*_{m(*)},\sigma^*_{m(*)+1})}w \ .$$

So by inductive hypothesis and the definition of $R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}$, we have

$$(w,e)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w,f_0) \ .$$

Therefore, $(w,e)R^{\mathbf{M}\otimes\mathbf{E}}_{\sigma_1+\sigma_2}(w',e')$. □

To the language $\mathscr{L}$, we add dynamic operators $[\mathbf{E},e]\,\varphi$ where $\mathbf{E}$ is a finite S5 regular reading event model and $e\in\mathbf{E}$. The new language is denoted by $\mathscr{L}_{\mathfrak{E}5}$ where $\mathfrak{E}5$ denotes the class of all finite S5-regular reading event models. The truth condition of $[\mathbf{E},e]\,\varphi$ in a regular model $\mathbf{M}$ is given by

$$\mathbf{M},w\models[\mathbf{E},e]\,\varphi \quad \text{iff} \quad \mathbf{M}\otimes\mathbf{E},(w,e)\models\varphi \ .$$

The reduction laws for $\mathscr{L}_{\mathfrak{E}}$ are the following ones:

- $[\mathbf{E},e]\,p\leftrightarrow p$
- $[\mathbf{E},e]\,\neg\varphi\leftrightarrow\neg\,[\mathbf{E},e]\,\varphi$
- $[\mathbf{E},e]\,(\varphi\wedge\psi)\leftrightarrow([\mathbf{E},e]\,\varphi\wedge[\mathbf{E},e]\,\psi)$
- $[\mathbf{E},e]\,[\tau]\,\varphi\leftrightarrow\bigwedge_{eR^{\mathbf{E}}_{\tau}e'}[\gamma(e,\tau)]\,[\mathbf{E},e']\,\varphi$

Where γ is the function defined in Proposition 4.1.

Proposition 4.2. *All above reduction axioms are valid in the class of S4/S5-regular frames.*

Proof. We only prove the validity of $[\mathbf{E},e]\,[\tau]\,\varphi\leftrightarrow\bigwedge\{[\gamma(e,\tau)]\,[\mathbf{E},e']\,\varphi\mid eR^{\mathbf{E}}_{\tau}e'\}$. The validity of other reduction laws are trivial.

Take an arbitrary regular model $\mathbf{M}$ and an arbitrary world w in it.

$\mathbf{M},w\models[\mathbf{E},e]\,[\tau]\,\varphi$ iff $\mathbf{M}\otimes\mathbf{E},(w,e)\models[\tau]\,\varphi$ iff $\forall(w',e')\in\mathbf{M}\otimes\mathbf{E}:(w,e)R^{\mathbf{M}\otimes\mathbf{E}}_{\tau}(w',e')\Rightarrow(w',e')\models\varphi$ iff (by Proposition 4.1) $\forall w'\in\mathbf{M},e'\in\mathbf{E}:wR^{\mathbf{M}}_{\gamma(e,\tau)}w'\,\&\,eR^{\mathbf{E}}_{\tau}e'\Rightarrow\mathbf{M}\otimes\mathbf{E},(w',e')\models\varphi$ iff $\forall w'\in\mathbf{M},e'\in\mathbf{E}:wR^{\mathbf{M}}_{\gamma(e,\tau)}w'\,\&\,eR^{\mathbf{E}}_{\tau}e'\Rightarrow\mathbf{M},w'\models[\mathbf{E},e']\,\varphi$ iff $\forall e'\in\mathbf{E}:eR^{\mathbf{E}}_{\tau}e'\Rightarrow\forall w'\in\mathbf{M}\,(wR^{\mathbf{M}}_{\gamma(e,\tau)}w'\Rightarrow\mathbf{M},w'\models[\mathbf{E},e']\,\varphi)$ iff $\forall e'\in\mathbf{E}:eR^{\mathbf{E}}_{\tau}e'\Rightarrow\mathbf{M},w\models[\gamma(e,\tau)]\,[\mathbf{E},e']\,\varphi$ iff $\mathbf{M},w\models\bigwedge_{eR^{\mathbf{E}}_{\tau}e'}[\gamma(e,\tau)]\,[\mathbf{E},e']\,\varphi$ □

Putting the above reduction laws together with the axiom system $S5\mathsf{LCDK}/S4\mathsf{LCDK}$, we get the axiom system for the dynamic logic of arbitrary reading, $S5\mathsf{LCDK}\mathfrak{E}/S4\mathsf{LCDK}\mathfrak{E}$.

Theorem 4.1. *S*5LCDK$\mathfrak{E}$/*S*4LCDK$\mathfrak{E}$ *is sound and complete with respect to the class of S5/S4-regular frames.*

The soundness follows from the validity of the reduction laws. By the validity of the reduction laws, we also get the equal expressivity of $\mathscr{L}$ and $\mathscr{L}_{\mathfrak{E}5}$ with respect to the class of regular models. So the completeness of *S*5LCDK$\mathfrak{E}$ and *S*4LCDK$\mathfrak{E}$ follows from the completeness of *S*5LCDK and *S*4LCDK respectively.

5 Conclusion and Future Work

By viewing distributed knowledge and common knowledge as meet and join in a lattice respectively, we have shown that the epistemic logic with operators for distributed knowledge and common knowledge can be extended to express more involved types of group knowledge in the S4 and S5 cases. To characterize the epistemic update by arbitrary reading events in the extended logic LCDK, we provide its reduction laws.

The connection with lattice theory is key to the extension proposed in this paper. By attaching the epistemic interpretation to lattice theoretical notions, it becomes possible to deepen our understanding of epistemic notions by exploring lattice theoretical results. For example, what if we ask the lattice of distributed knowledge and common knowledge to be modular or even distributive? Some results in lattice theory can be found in [10].

Although this paper focuses on the epistemic interpretation of the lattice structre of the equivalence relations/reflexive and transitive relations, there are other possible interpretations. For example, questions in [4]; functional dependence in [1]; priority structures in [6], and so on. The lattice theoretical perspective of this paper can be applied to the other interpretations too.

A Soundness and Completeness

Soundness is easy to see, so we ignore the proof. The proof for weak completeness comes next.

Steps (1) to (4) are adapted from Section 4.8 of [5].

A.1 Step (1) Fischer-Ladner closure and atoms

Definition A.1 (Fischer-Ladner closure)**.** Let X be a set of formulas. Then X is FL-closed if it is closed under subformulas and satisfies the following additional constraints

1. If $\langle\tau\cdot\sigma\rangle\varphi\in X$, then $\langle\tau\rangle\varphi\wedge\langle\sigma\rangle\varphi\in X$;
2. If $\langle\tau+\sigma\rangle\varphi\in X$, then $\langle\tau\rangle\langle\tau+\sigma\rangle\varphi,\langle\sigma\rangle\langle\tau+\sigma\rangle\varphi\in X$.

If Σ is any set of formulas, then $\mathrm{FL}(\Sigma)$ is the smallest set of formulas containing Σ that is FL closed.

Definition A.2. We define $\neg\mathrm{FL}(\Sigma)$, the closure of Σ, as the smallest set containing Σ which is FL-closed and closed under single negations. The single negation of a formulas ϕ is

$$\sim\phi:=\begin{cases}\psi & \phi \text{ is of the form } \neg\psi\\ \neg\phi & \text{otherwise}\end{cases}$$

Definition A.3 (Atoms)**.** Let Σ be a set of formulas. A set of formulas A is an atom over Σ if it is a maxiaml consistent subset of $\neg\mathrm{FL}(\Sigma)$.

Lemma A.0.1. *Let Σ be any set of formulas, and A any element of $At(\Sigma)$. Then*

- *For all ϕ, exactly one of ϕ and $\sim\phi$ is in A.*
- *For all $\phi\vee\psi\in\neg\mathrm{FL}(\Sigma)$: $\phi\vee\psi\in A$ iff $\phi\in A$ or $\psi\in A$.*
- *For all $\langle\tau\rangle\phi,\langle\sigma\rangle\phi\in\neg\mathrm{FL}(\Sigma)$: if $\sigma\leq\tau$ and $\langle\sigma\rangle\phi\in A$, then $\langle\tau\rangle\phi\in A$.*
- *For all $\langle\tau+\sigma\rangle\phi\in\neg\mathrm{FL}(\Sigma)$: $\langle\tau+\sigma\rangle\phi\in A$ iff $\phi\in A$ or $\langle\tau\rangle\langle\tau+\sigma\rangle\phi\in A$ or $\langle\sigma\rangle\langle\tau+\sigma\rangle\phi\in A$.*

Lemma A.0.2. *Let $\mathscr{M}$ be the set of all MCSs, and Σ a set of formulas. $At(\Sigma)=\{\Gamma\cap\neg\mathrm{FL}(\Sigma)\mid\Gamma\in\mathscr{M}\}$.*

Lemma A.0.3. *If $\phi\in\neg\mathrm{FL}(\Sigma)$ and ϕ is consistent, then there is an $A\in At(\Sigma)$ such that $\phi\in A$.*

A.2 Step (2): Finite Canonical Model over Σ

Definition A.4. Let Σ be a finite set of formulas. The finite canonical model (FCM) over Σ is the triple $\mathscr{M}_F=(At(\Sigma),\{S^\Sigma_\tau\}_{\tau\in T},V^\Sigma)$ where for all propositional variables p, $V^\Sigma(p)=\{A\in At(\Sigma)\mid p\in A\}$, and for all atoms $A,B\in At(\Sigma)$ and for all terms τ, $AS^\Sigma_\tau B$ if $\hat{A}\wedge\langle\tau\rangle\hat{B}$ is consistent.

Lemma A.0.4. *The canonical relations are reflexive and symmetric.*

The above lemma can also be proved as a corollary to the following fact.

Proposition A.1. *For $A,B\in At(\Sigma)$, $AS_\tau B$ iff there is $a\in[\![\hat{A}]\!]_{\mathscr{M}}$ and there is $b\in[\![\hat{B}]\!]_{\mathscr{M}}$ such that $aR^{\mathscr{M}}_\tau b$ where $\mathscr{M}$ is the canonical model.*

This fact also indicates that S_τ is not necessarily transitive. Moreover, it makes the following existence lemma a corollary to the existence lemma for the canonical model.

Lemma A.0.5 (Existence lemma for FCM)**.** *Let A be an atom, and let τ be a term in T. Then for all formulas $\langle\tau\rangle\psi$ in $\neg\mathrm{FL}(\Sigma)$, $\langle\tau\rangle\psi\in A$ iff there is a $B\in At(\Sigma)$ such that $AS_\tau B$ and $\psi\in B$.*

Lemma A.0.6. $S_{\tau+\sigma}\subseteq(S_\tau\cup S_\sigma)^*$.

Lemma A.0.7. $S^\Sigma_{\sigma\cdot\tau}\subseteq S^\Sigma_\sigma\cap S^\Sigma_\tau$.

A.3 Step (3): Transitive FCM

Definition A.5 (Transitive FCM (TFCM))**.** Given a FCM $\mathscr{M}^\Sigma_F=(At(\Sigma),\{S^\Sigma_\tau\}_{\tau\in T},V^\Sigma)$, we define its transtive closure as $\mathscr{M}^\Sigma_{TF}=(At(\Sigma),\{(S^\Sigma_\tau)^t\}_{\tau\in T},V^\Sigma)$ where $(S^\Sigma_\tau)^t$ is the transitive closure of S^Σ_τ.

To prove the existence lemma for TFCM, the direction from left to right immediately follows from the existence lemma for FCM, because $S^\Sigma_\tau\subseteq(S^\Sigma_\tau)^t$. The other direction follows from the following fact.

Proposition A.2. *For $A,B\in At(\Sigma)$, $AS^\Sigma_\tau B$ implies that $A_\tau=B_\tau$ where $A_\tau=\{\varphi\mid\langle\tau\rangle\varphi\in A\}$.*

Lemma A.0.8 (Existence lemma for TFCM)**.** *Let A be an atom, and let τ be a term in T. Then for all formulas $\langle\tau\rangle\psi$ in $\neg\mathrm{FL}(\Sigma)$, $\langle\tau\rangle\psi\in A$ iff there is a $B\in At(\Sigma)$ such that $A(S_\tau)^tB$ and $\psi\in B$.*

A.4 Step (4): Quasi-regularity of TFCM

Lemma A.0.9. $(S_{\tau+\sigma})^t=(S_\tau\cup S_\sigma)^*$.

By the above lemma, TFCM only gets one of the two constraints for a regular model right. In this sense, it is quasi-regular. The other constraint $R_{\sigma\cdot\tau}=R_\sigma\cap R_\tau$ does not hold. In fact, only the direction from right to left does not necessarily hold in a TFCM.

Proposition A.3. $(S^\Sigma_{\sigma\cdot\tau})^t\subseteq(S^\Sigma_\sigma)^t\cap(S^\Sigma_\tau)^t$

A.5 Step (5): Building S5-Regular Model Step by Step

Definition A.6. A network is a triple $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, V)$ such that R_τ is a binary relation on a set N, and V is a labeling function mapping each point in N to an atom in $At(\Sigma)$ for a finite set of formulas Σ.

Definition A.7 (Walk and Path in a network)**.** Given a network $\mathscr{N}$, a walk from u to v is a sequence $(w_0, R_1 \ldots, R_n w_n)$ satisfying $w_0 = u$, $w_n = v$ and $w_i R_i w_{i+1}$ in $\mathscr{N}$; a path from u to v is a walk where $w_i \neq w_j$ for $i \neq j$. A τ-walk from u to v is a walk $(w_0, R_\tau \ldots, R_\tau w_n)$; a τ-path is a τ-walk from u to v where $w_i \neq w_j$ for $i \neq j$. We stipulate that (u) is a τ-path for any $\tau \in T$.

We will also talk about notions of walk and path in other strctures, the definitions of which are analogous.

Definition A.8. A network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, V)$ is coherent if for any $u, v \in \mathscr{N}$ and any $\tau \in T$,

(C0) R_τ is symmetric;

(C1) if $uR_\tau v$, then $\widehat{V(u)} \wedge \langle\tau\rangle \widehat{V(v)}$ is consistent;

(C2) if $uR_\tau v$ and $\tau \leq \tau'$, then $uR_{\tau'} v$;

(C3) if there is a path from u to v where $u \neq v$, then there is $\sigma \in T$ such that

$$\uparrow \sigma = \{\tau \in T \mid \text{ there is a } \tau\text{-walk from } u \text{ to } v\}$$

the corresponding relation R_σ is called the foundational bridge between u and v in $\mathscr{N}$.[1]

Here are some basic facts about walks, paths and foundational bridges in a coherent network.

Proposition A.4. *Given a coherent network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, V)$*

1. $\{\tau \in T \mid$ *there is a τ-walk from u to v*$\} = \{\tau \in T \mid$ *there is a τ-path from u to v*$\}$
2. *if $uR_{\sigma_0} w_1 R_{\sigma_1} \ldots R_{\sigma_{n-1}} w_n R_{\sigma_n} v$ is a path from u to v, then $uR_{\sum_0^n \sigma_i} w_1 R_{\sum_0^n \sigma_i} \ldots R_{\sum_0^n \sigma_i} w_n R_{\sum_0^n \sigma_i} v$ is a path.*
3. *If R_τ is the basic bridge between u and u', R_σ is the basic bridge between u' and u'' and all paths from u to u'' go via u', then $R_{\tau+\sigma}$ is the basic bridge between u and u''.*

Proof. (1) $\{\tau \in T \mid$ there is a τ-walk from u to $v\} \supseteq \{\tau \in T \mid$ there is a τ-path from u to $v\}$ is obvious.

The other direction holds because from any τ-walk from u to v, we can get a τ-path by pruning the τ-walk: traversing the τ-walk from u, every time a node w in the τ-walk appears twice, deleting the nodes between the first and the second appearances of w including the second appearance of w.

(2) follows from the condition C2.

(3) Take any path from u to u'', $uR_\delta w_1 R_\delta \ldots R_\delta w_n R_\delta u''$. It goes via u'. So there is i such that $s_i = u'$. R_τ is the foundational bridge from u to u' and R_σ is the foundational bridge between u' and u'' . So $\tau \leq \delta$ and $\sigma \leq \delta$, which imply that $\tau + \sigma \leq \delta$. So $\delta \in \uparrow (\tau + \sigma)$.

On the other hand, there is a τ-path $uR_\tau w_1 R_\tau \ldots R_\tau w_n R_\tau u'$ and a σ-path $u' R_\sigma v_1 R_\sigma \ldots R_\sigma v_m R_\tau u''$. It implies that $uR_\tau w_1 R_{\tau+\sigma} \ldots R_{\tau+\sigma} w_n R_{\tau+\sigma} u' R_{\tau+\sigma} v_1 R_{\tau+\sigma} \ldots R_{\tau+\sigma} v_m R_\tau u''$ is a $\tau + \sigma$-path.

Therefore, $R_{\tau+\sigma}$ is the foundational bridge from u to u''. □

Definition A.9. A network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, V)$ is saturated if it satisfies:

(S1) for any $u, v \in \mathscr{N}$ and any $\sigma + \sigma' \in T$, if $uR_{\sigma+\sigma'} v$, then there is a path from u to v each step of which is R_σ or $R_{\sigma'}$.

[1] Strictly speaking, there are more than one foundational bridges from u to v. However, for any two foudational bridges from u to v in a coherent network, R_σ and R_τ, $R_\sigma = R_\tau$ because of condition (C2).

(S2) $\mathscr{N}$ is modally saturated. That is, we demand that if $\langle\tau\rangle\psi \in \nu(u)$ for some $u \in N$, then there is some $v \in N$ such that $uR_\tau v$ and $\psi \in \nu(v)$.

Definition A.10. Let $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, \nu)$ be a network. The frame $\mathfrak{F}_\mathscr{N} = (N, \{R_\tau\}_{\tau\in T})$ is called the underlying frame of $\mathscr{N}$. The induced valuation $V_\mathscr{N}$ on $\mathfrak{F}_\mathscr{N}$ is defined by $V_\mathscr{N}(p) = \{s \in N \mid p \in \nu(s)\}$ for $p \in \neg FL(\Sigma)$. The structure $\mathfrak{I}_\mathscr{N} = (\mathfrak{F}, V_\mathscr{N})$ is the induced model.

A network is good if it is both coherent and saturated.

Lemma A.0.10 (Existence Lemma for good network)**.** *Let $\mathscr{N}$ be a good network. For all formulas $\langle\tau\rangle\psi$ in $\neg$FL(Σ) and $s \in N$, $\langle\tau\rangle\psi \in \nu(s)$ iff there is a $s' \in N$ such that $sR_\tau s'$ and $\psi \in \nu(s')$.*

Proposition A.5. *A good network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, \nu)$ satsifies: for any $\tau, \tau', \sigma + \sigma' \in T$ and $u, v \in N$*

1. $R_{\tau\cdot\tau'} \subseteq R_\tau \cap R_{\tau'}$;
2. $R_{\sigma+\sigma'} \subseteq (R_\sigma \cup R_{\sigma'})^*$

Proof. The first follows from (C2) and the second follows from (S1). □

Definition A.11. Given a good network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, \nu)$, let the reflexive and transitive closure of R_τ for each $\tau \in T$ be denoted by E_τ, the perfect network generated from $\mathscr{N}$ is a triple $\mathscr{P}_\mathscr{N} = (N, \{E_\tau\}_{\tau\in T}, \nu)$.

Proposition A.6. *A perfect network satsisfies C0, C2, C3, S1 and S2. Moreover, the frame underlying a perfect network is S5-regular.*

Proof. Given a perfect network $\mathscr{P}_\mathscr{N} = (N, \{E_\tau\}_{\tau\in T}, \nu)$ from the good network $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, \nu)$, we first prove that it satisfies C0, C2, C3, S1 and S2.

(C0) Assume that $uE_\tau v$, then $uR_\tau v$ or there is $w_1, \ldots, w_n \in N$ such that $uR_\tau w_1 R_\tau \ldots R_\tau w_n R_\tau v$. R_τ is symmetric, so $vE_\tau u$.

(C2) Assume that $uE_\tau v$ and $\tau \leq \tau'$. If $uR_\tau v$, then it follows that $uR_{\tau'} v$ and thus $uE_{\tau'} v$. So assume that $uR_\tau v$ is not the case. Then there is $w_1, \ldots, w_n \in N$ such that $uR_\tau w_1 R_\tau \ldots R_\tau w_n R_\tau v$. So $uR_{\tau'} w_1 R_{\tau'} \ldots R_{\tau'} w_n R_{\tau'} v$ and thus $uE_{\tau'} v$.

(C3) Assume that there is a path from u to v in $\mathscr{P}_\mathscr{N}$. It follows that there is a path in $\mathscr{N}$ from u to v. So there is $\sigma \in T$ such that σ is a foundational bridge from u to v. It is easy to see that

$$\uparrow\sigma \subseteq \{\tau \in T \mid \text{ there is a } \tau\text{-path from } u \text{ to } v \text{ in } \mathscr{P}_\mathscr{N}\}.$$

For the other direction, take any path $uE_\tau w_1 E_\tau \ldots E_\tau w_n E_\tau v$, then there must be a path between w_i and w_{i+1} in $\mathscr{N}$ for any $1 \leq i \leq n-1$: $w_i R_\tau w_i^1 R_\tau w_i^2 R_\tau \ldots R_\tau w_i^{m_i} R_\tau w_{i+1}$. So there is a τ-path from u to v in $\mathscr{N}$, which implies that $\tau \in \uparrow\sigma$. This implies that

$$\uparrow\sigma \supseteq \{\tau \in T \mid \text{ there is a } \tau\text{-path from } u \text{ to } v \text{ in } \mathscr{P}_\mathscr{N}\}.$$

(S1) Assume that $uE_{\sigma+\sigma'} v$. It follows that $u = w_0 R_{\sigma+\sigma'} w_1 R_{\sigma+\sigma'} \ldots R_{\sigma+\sigma'} w_n R_{\sigma+\sigma'} w_{n+1} = v$ for some $w_1, \ldots, w_n$. For each $w_i R_{\sigma+\sigma'} w_{i+1}$, there is a path $w_i R_{\sigma_i^1} w_i^1 R_{\sigma_i^2} w_i^2 \ldots w_i^{m_i-1} R_{\sigma_i^{m_i}} w_{i+1}$ where $R_{\sigma_i^j}$ is R_σ or $R_{\sigma'}$. So there is a path from u to v each step of which is R_σ or $R_{\sigma'}$, which implies that there is a path from u to v each of which is E_σ or $E_{\sigma'}$.

(S2) This condition obviously holds in $\mathscr{P}_\mathscr{N}$

$E_{\tau\cdot\tau'} \subseteq E_\tau \cap E_{\tau'}$: This follows from C2.

$E_{\tau\cdot\tau'} \supseteq E_\tau \cap E_{\tau'}$: Assume that $uE_\tau v$ and $uE_{\tau'} v$. By C3, there is a foundational bridge E_σ between u and v. So for some $w_1, \ldots, w_n \in N$, $uE_\sigma w_1 E_\sigma \ldots E_\sigma w_n E_\sigma v$, which implies that $uE_\sigma v$ by transitivity of

E_σ. Together with C2, it implies that $\uparrow \sigma = \{\delta \in T \mid uE_\delta v\}$. From our assumption it follows that $\sigma \leq \tau$ and $\sigma \leq \tau'$. So $\sigma \leq \tau \cdot \tau'$, which implies that $uE_{\tau\cdot\tau'}v$.

$E_{\sigma+\sigma'} \subseteq (E_\sigma \cup E_{\sigma'})^*$: This follows from S1.

$E_{\sigma+\sigma'} \supseteq (E_\sigma \cup E_{\sigma'})^*$: Assume that $u(E_\sigma \cup E_{\sigma'})^* v$. By C2, $u(E_{\sigma+\sigma'})^* v$. By transitivity of $E_{\sigma+\sigma'}$, it follows that $uE_{\sigma+\sigma'}v$.

(E_τ is an equivalence relation): obvious. □

Lemma A.0.11 (Existence Lemma for perfect network)**.** *Let $\mathscr{P}_\mathscr{N}$ be a perfect network generated from a good network $\mathscr{N}$. For all formulas $\langle\tau\rangle\psi$ in $\neg\mathrm{FL}(\Sigma)$ and $s \in N$, $\langle\tau\rangle\psi \in \nu(s)$ iff there is a $s' \in N$ such that $sE_\tau s'$ and $\psi \in \nu(s')$.*

Proof. The proof is similar to that of Lemma A.0.8. □

Next, we identify those defects we may encounter during our construction of a perfect network, after which we show that they can all be repaired.

Definition A.12. Let $\mathscr{N} = (N, \{R_\tau\}_{\tau\in T}, \nu)$ be a network. An S1-defect of $\mathscr{N}$ consists of a pair of nodes $(s,s') \in R_{\sigma+\sigma'}$ for which no path from s to s' exists such that each step is R_σ or $R_{\sigma'}$. An S2-defect consists of a node s and a formula $\langle\tau\rangle\psi \in \nu(s)$ for which there is no $s' \in N$ such that $sR_\tau s'$ and $\psi \in \nu(s')$.

Definition A.13. Let $\mathscr{N}_0 = (N_0, \{R^0_\tau\}_{\tau\in T}, \nu_0)$ and $\mathscr{N}_1 = (N_1, \{R^1_\tau\}_{\tau\in T}, \nu_1)$ be two networks. We say that $\mathscr{N}_1$ extends $\mathscr{N}_0$ (notation $\mathscr{N}_1 \rhd \mathscr{N}_0$) if $\mathfrak{F}_{\mathscr{N}_0}$ is a subframe of $\mathfrak{F}_{\mathscr{N}_1}$ and ν_0 agrees with ν_1 on N_0.

Lemma A.0.12 (Repair lemma)**.** *For any defect of a finite, coherent network $\mathscr{N}$ there is a finite, coherent $\mathscr{N}' \rhd \mathscr{N}$ lacking this defect.*

Proof. We first deal with the case when the defect is S2 type. Given the defect pair $(s, \langle\tau\rangle\psi)$, choose a new point which is not in N and let B be an atom such that $\nu(s)S^\Sigma_\tau B$ in the FCM for Σ (such a B exists by the Existence Lemma for FCM). Define $\mathscr{N}' = (N', \{R'_\tau\}_{\tau\in T}, \nu')$ as follows:

- $N' := N \cup \{s'\}$;
- $R'_\sigma := R_\sigma$ for all $\sigma \not\geq \tau$ and $R'_\sigma := R_\sigma \cup \{(s,s'),(s',s)\}$ for all $\sigma \geq \tau$;
- $\nu' = \nu \cup \{(s', B)\}$.

We first prove that $\mathscr{N}'$ satisfies (C3). If there is a path from t to s' in $\mathscr{N}'$, the path must pass through s. Because $\uparrow \tau = \{\delta \mid sR_\delta s'\}$ and there is $\sigma \in T$ such that $\uparrow \sigma = \{\delta \in T \mid$ there is a δ-path from t to s in $\mathscr{N}'\}$, by Proposition A.4.(3), it follows that $\uparrow (\sigma + \tau) = \{\delta \in T \mid$ there is a δ-path from t to s' in $\mathscr{N}'\}$.

The fact that everytime we add a pair into a relation R_τ we also add the pair into R_σ for $\sigma \geq \tau$ makes sure that (C2) still holds in $\mathscr{N}'$. The reason for (C1) holding in $\mathscr{N}'$ is that all pairs we add into a relation R_τ are based on the relation S^Σ_τ between their lables and S^Σ_τ is symmetric. (C0) is guaranteed by the way we add pairs into each relation.

Next, we deal with the case when the defect is S1 type. Given the defect triple $(u, \sigma+\sigma', v)$, because $\mathscr{N}$ is coherent, we know that there is a foundational bridge σ_0 between u and v such that $\sigma + \sigma' \in \uparrow \sigma_0$ and $\sigma_0 \not\leq \sigma$ and $\sigma_0 \not\leq \sigma'$.

We will construct a path $s_0 R_{\sigma_1} s_1 R_{\sigma_2} \ldots R_{\sigma_n} s_n$ so that $s_0 = u$, $s_n = v$, $\sigma_{2i+1} = \sigma$ and $\sigma_{2i} = \sigma'$, each s_i is labeled by an atom B_i and $B_0 S^\Sigma_{\sigma_1} B_1 S^\Sigma_{\sigma_2} \ldots S^\Sigma_{\sigma_n} B_n$ holds in the FCM for Σ where $n > 3$ is even.

Since $\nu(u)S^\Sigma_{\sigma+\sigma'}\nu(v)$ in FCM and $S^\Sigma_{\sigma+\sigma'} \subseteq (S^\Sigma_\sigma \cup S^\Sigma_{\sigma'})^*$, so there must be a path from $\nu(u)$ to $\nu(v)$ in the FCM each step of which is S^Σ_σ or $S^\Sigma_{\sigma'}$. We take such a path, from which we can construct a walk from $\nu(u)$ to $\nu(v)$, $B_0 S^\Sigma_{\sigma_1} B_1 S^\Sigma_{\sigma_2} \ldots S^\Sigma_{\sigma_n} B_n$ where $\nu(u) = B_0$ $\nu(v) = B_n$, $n > 3$, $\sigma_{2i+1} = \sigma$, $\sigma_{2i} = \sigma'$ and n is even by repeating some nodes in the path using the reflexivity of the relations. Define $\mathscr{N}' = (N', \{R'_\tau\}_{\tau\in T}, \nu')$ as follows:

- $N' := N \cup \{s_1, s_2 \ldots, s_{n-1}\}$;
- let $s_0 = u$ and $s_n = v$:
  - $R'_\tau := R_\tau$ for all $\tau \in T$ satisfying $\tau \not\geq \sigma$ and $\tau \not\geq \sigma'$; and
  - $R'_\tau := R_\tau \cup \{(s_i, s_{i+1}), (s_{i+1}, s_i) \mid n > i \geq 0 \text{ is even}\}$ for all $\tau \geq \sigma$; and;
  - $R'_\tau := R_\tau \cup \{(s_i, s_{i+1}), (s_{i+1}, s_i) \mid n > i \geq 0 \text{ is odd}\}$ for all $\tau \geq \sigma'$;
- $\nu' = \nu \cup \{(s_i, B_i) \mid 1 \leq i \leq n-1\}$.

Conditions (C0), (C1) and (C2) of $\mathscr{N}'$ are proved in a similar way to the previous case.

For (C3), given a path from s to s'. Consider three cases.

In the first case, the starting point s is in $\{s_1, s_2 \ldots, s_{n-1}\}$, say $s = s_i$, and the ending point s' is not in $\{s_1, s_2 \ldots, s_{n-1}\}$.

One of the following cases must hold:

- there is a path from v to s' in $\mathscr{N}$ which does not go via u;
- there is a path from u to s' in $\mathscr{N}$ which does not go via v.

Without loss of generality, we assume that there is a path from v to s' in $\mathscr{N}$ which does not go via u. It follows that there is $\delta \in T$ such that $\uparrow \delta = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } v \text{ to } s' \text{ in } \mathscr{N}\}$.

Next, we consider all paths from s_i to v in $\mathscr{N}'$. All paths from s_i to v either goes via u or does not go via u. Let

$$u^+ = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ via } u \text{ in } \mathscr{N}'\}$$

and

$$u^- = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ not via } u \text{ in } \mathscr{N}'\}.$$

So $\{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ in } \mathscr{N}'\} = u^+ \cup u^-$.

Because there is only one path from s_i to v in $\mathscr{N}'$ which does not go via u, that is,

$$s_i R'_{\sigma_{i+1}} s_{i+1} R'_{\sigma_{i+2}} \ldots R'_\tau s_{n-1} R'_{\sigma_n} v \ .$$

It follows that u^- equals $\uparrow (\sigma + \sigma')$ or $\uparrow \sigma'$, depending on whether $i < n-1$ or $i = n-1$. There is only one path from s_i to u in $\mathscr{N}'$ which does not go via v in $\mathscr{N}'$, that is, $uR'_{\sigma_1} s_1 R'_{\sigma_2} \ldots R'_\tau s_{i-1} R'_{\sigma_i} s_i$. It follows that $\{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } u \text{ not via } v \text{ in } \mathscr{N}'\}$ equals $\uparrow (\sigma + \sigma')$ or $\uparrow \sigma$, depending on whether $i > 1$ or $i = 1$.

When $i = n-1$, $\uparrow (\sigma + \sigma') = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } u \text{ not via } v \text{ in } \mathscr{N}'\}$. Together with the fact that R_{σ_0} is a foundational bridge from u to v in $\mathscr{N}$, it follows that $u^+ = \uparrow (\sigma + \sigma' + \sigma_0)$. Since $\sigma_0 \leq \sigma + \sigma'$, $u^+ = \uparrow (\sigma + \sigma')$. Combining with the facts that u^- equals $\uparrow \sigma'$ and $\sigma' \leq \sigma + \sigma'$, it follows that $\uparrow \sigma' = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ in } \mathscr{N}'\}$, that is, the foundational bridge from s_i to v in $\mathscr{N}'$ is $R'_{\sigma'}$.

When $1 < i < n-1$, it is still the case that $u^+ = \uparrow (\sigma + \sigma')$ using a similar argument in the case $i + 1 = n$. But u^- equals $\uparrow (\sigma + \sigma')$ in this case. Thus it follows that

$$\uparrow (\sigma + \sigma') = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ in } \mathscr{N}'\}$$

that is, the foundational bridge from s_i to v in $\mathscr{N}'$ is $R'_{\sigma+\sigma'}$.

When $i = 1$, $\uparrow \sigma = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } u \text{ not via } v \text{ in } \mathscr{N}'\}$. Together with the fact that R_{σ_0} is the foundational bridge from u to v in $\mathscr{N}$, it follows that $u^+ = \uparrow (\sigma_0 + \sigma)$. Since u^- equals $\uparrow (\sigma + \sigma')$ in this case and $\sigma_0 + \sigma \leq \sigma + \sigma'$, it follows that

$$\uparrow (\sigma_0 + \sigma) = \{\tau \in T \mid \text{ there is a } \tau\text{-path from } s_i \text{ to } v \text{ in } \mathscr{N}'\}$$

that is, the foundational bridge from s_i to v in $\mathscr{N}'$ is $R'_{\sigma_0+\sigma}$.

Therefore, the foundational bridge from s_i to s in $\mathscr{N}'$ is either $R'_{\sigma'+\delta}$, $R'_{\sigma+\sigma'+\delta}$ or $R'_{\sigma_0+\sigma+\delta}$.

The second case where the ending point s' is in $\{s_1, s_2 \ldots, s_{n-1}\}$ but the starting point s is not in $\{s_1, s_2 \ldots, s_{n-1}\}$. This case is symmetric to the first case, so it follows by C0.

In the third case, s and s' are both in $\{s_1, s_2 \ldots, s_{n-1}\}$, say $s = s_i$ and $s' = s_j$ and $i \leq j$. The proof is similar to the one we use in the first case for showing that there is a foundational bridge between s and v in $\mathscr{N}'$. In this case, there are two ways to go from s_i to s_j. The first one is $s_i R_{\sigma_i+1} \ldots R_{\sigma_j} s_j$ which does not go via u or v. It is easy to see that the set of all possible paths via this way is either $\uparrow(\sigma+\sigma')$, $\uparrow \sigma$ or $\uparrow \sigma'$. The other one goes via u and v. The set of all possible paths is $\uparrow(\sigma+\sigma')$. Therefore, the foundational bridge from s_i to s_j in $\mathscr{N}'$ is either $R'_{\sigma+\sigma'}$, R'_{σ} or $R'_{\sigma'}$.

In the last case, neither s nor s' is in $\{s_1, s_2 \ldots, s_{n-1}\}$. If there is no path from s to s' in $\mathscr{N}$ going via both u and v, the foundational bridge from s to s' in $\mathscr{N}$ is the foundational bridge from s to s' in $\mathscr{N}'$.

Assume that there are paths from s to s' in $\mathscr{N}$ going via both u and v. There is $\delta \in T$ such that $\uparrow \delta = \{\tau \in T \mid$ there is a τ-path from s to s' in $\mathscr{N}\}$.

Obviously, $\uparrow \delta \subseteq \{\tau \in T \mid$ there is a τ-path from s_i to s' in $\mathscr{N}'\}$.

For the other direction, take any τ-path from s to s' in $\mathscr{N}'$, $sR'_\tau w_1 R'_\tau \ldots R'_\tau w_m R'_\tau s'$. If the path goes via $s_1, \ldots, s_{n-1}$, then $\sigma \leq \tau$ and $\sigma' \leq \tau$. So $\sigma + \sigma' \leq \tau$ and we can delete $s_1, \ldots, s_{m-1}$ from the path and connect u and v by a τ-path in $\mathscr{N}$, becauase $\sigma_0 \leq \sigma + \sigma' \leq \tau$. This implies that there is a path in $\mathscr{N}$: $sR_\tau w_1 R_\tau \ldots u R_\tau \ldots R_\tau v \ldots R_\tau w_m R_\tau s'$ So $\tau \in \uparrow \delta$.

Therefore, $\uparrow \delta = \{\tau \in T \mid$ there is a τ-path from s to s' in $\mathscr{N}'\}$. □

Lemma A.0.13. *For an atom $A \in At(\Sigma)$, there is a good network where one of its nodes is labelled by A.*

Proof. The proof is similar to that of Theorem 4.65 in [5].

Choose some set $S = \{s^i \mid i \in \omega\}$ and enumerate the set of potential defects (that is, the union of the sets $S \times T \times T \times S$ and $S \times T \times \mathscr{L}$). Given an atom $A \in At(\Sigma)$, let $\mathscr{N}_0$ be the network $(\{s^0\}, \emptyset, (s^0, A))$. Trivially, $\mathscr{N}_0$ is a finite, coherent network.

Let $n \geq 0$ and suppose $\mathscr{N}_n$ is a finite, coherent network. If there is no defect in $\mathscr{N}_n$, let $\mathscr{N}_m = \mathscr{N}_n$ for all $m > n$. If there are defects in $\mathscr{N}_n$, let D be the defect of $\mathscr{N}_n$ that is minimal in our enumeration. Form $\mathscr{N}_{n+1}$ by repairing the defect D as described in the proof of the Repair Lemma. Observe that D will not be a defect of any network extending $\mathscr{N}_{n+1}$.

Let $\mathscr{N} = (N, \{R_\tau\}_{\tau \in T}, \nu)$ be given by $N = \bigcup_{n \in \omega} N_n$, $\quad R_\tau = \bigcup_{n \in \omega} R^n_\tau$ for all $\tau \in T$, $\quad \nu = \bigcup_{n \in \omega} \nu_n$. It is a coherent network since our repair preserves all conditions of coherency. Moreover, $\mathscr{N}$ is saturated. Suppose otherwise, that is,there are some defects in $\mathscr{N}$. From these defects, take the minimal one in the enumeration, say D_k. It must be a defect in some approximation of $\mathscr{N}$, say $\mathscr{N}_n$. Although D_k is not necessarily the minimal defect in $\mathscr{N}_n$, there are finite defects before D_k in $\mathscr{N}_n$ to be repaired afterwards. After finite many steps of repair, D_k becomes the minimal defect to be repaired. So D_k is repaired at some stage of the process of repair, contradiction. □

A.6 Proving the completeness of $S5\mathsf{LCDK}$

For any finite consistent set of formulas Σ, extend it to an atom A over Σ and construct a good network $\mathscr{N}$ according to Lemma A.0.13. Construct the perfect network $\mathscr{P}_\mathscr{N}$ based on the good network $\mathscr{N}$. Let $\mathfrak{F}_\mathscr{P}$ be the frame underlying the perfect network $\mathscr{P}$. Finally, prove the truth lemma for the induced model $\mathfrak{I}_\mathscr{P} = (\mathfrak{F}_\mathscr{P}, V_\mathscr{N})$ using the Existence Lemma for perfect network (Lemma A.0.11). Because the frame underlying a perfect network is S5-regular, we have proved the completeness of $S5\mathsf{LCDK}$ with respect to the class of S5-regular frames.

A.7 Proving the completeness of $S4\mathsf{LCDK}$

In the case of S4, Lemma A.0.4 should be changed to

Lemma A.0.14. *The canonical relations are reflexive.*

Proposition A.2 should be changed to

Proposition A.7. *$AS^{\Sigma}_{\tau}B$ implies that $B_\tau \subseteq A_\tau$*

Lemma A.0.8 makes use of Proposition A.2. Its proof in the S4 case should be changed accordingly.

The condition C0 in the definition of coherent networks should be removed. The proofs in Section A.5 which are based on the condition C0 need to be modified accordingly. In the proof of the repair lemma, when constructing new networks, we keep the forward pairs (s,s'), (s_i,s_{i+1}) and get rid of the backward pairs (s',s), (s_{i+1},s_i). The proof that our repair of the two types of defects preserves all conditions of coherency becomes simpler, because the paths from s_1 to s_{n-1} in $\mathscr{N}'$ are all one-way paths. For S1 defect, the original proof still works. For S2 defect, we only need to consider the first, the third and the fourth cases we consider in the original proof. For the first case, the only way to reach s' from s_i is via v and all paths from s_i to v go only via s_k where $i \leq k < n$. For the third case, all path from s_i to s_j go only via s_k where $i \leq k \leq j$. For the fourth case, the orignal proof still works.

Acknowledgements I'd like to thank Alexandru Baltag for his suggestions and comments on several early versions of this paper, which help clarify some critical components of the logic. Many thanks to the three anonymous reviewers for their useful suggestions on improving the presentation of the paper.

References

[1] Alexandru Baltag & Johan van Benthem (2021): *A Simple Logic of Functional Dependence*. *Journal of Philosophical Logic* 50, pp. 939–1005, doi:10.1007/s10992-020-09588-z.

[2] 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.

[3] Alexandru Baltag & Sonja Smets (2020): *Learning What Others Know*. In Elvira Albert & Laura Kovacs, editors: *LPAR23. LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning*, *EPiC Series in Computing* 73, EasyChair, pp. 90–119, doi:10.29007/plm4. Available at `/publications/paper/V8Jp`.

[4] Johan van Benthem & Ştefan Minică (2012): *Toward a Dynamic Logic of Questions*. *Journal of Philosophical Logic* 41, pp. 633–669, doi:10.1007/s10992-012-9233-7.

[5] Patrick Blackburn, Maarten de Rijke & Yde Venema (2001): *Modal Logic*. *Cambridge Tracts in Theoretical Computer Science* 53, Cambridge University Press, doi:10.1017/CBO9781107050884. Available at `https://doi.org/10.1017/CBO9781107050884`.

[6] Zoé Christoff, Norbert Gratzl & Olivier Roy (2022): *Priority Merge and Intersection Modalities*. *The Review of Symbolic Logic* 15(1), pp. 165–96, doi:10.1017/S1755020321000058.

[7] Hans van Ditmarsch, Wieber van der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. *Synthese Library* 337, Springer Dordrecht, doi:10.1007/978-1-4020-5839-4.

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

[9] Jeroen Groenendijk & Martin Stokhof (1984): *Studies on the Semantic of Questions and the Pragmatics of Answers*. Phd dissertation, University of Amsterdam.

[10] Øystein Ore (1942): *Theory of Equivalence Relations*. *Duke Mathematical Journal* 9, pp. 573–627, doi:10.1215/S0012-7094-42-00942-6.

Who is Afraid of Minimal Revision?

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

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

Nina Gierasimczuk
Technical University of Denmark
nigi@dtu.dk

Rineke Verbrugge
University of Groningen
l.c.verbrugge@rug.nl

The principle of minimal change in belief revision theory requires that, when accepting new information, one keeps one's belief state as close to the initial belief state as possible. This is precisely what the method known as minimal revision does. However, unlike less conservative belief revision methods, minimal revision falls short in learning power: It cannot learn everything that can be learned by other learning methods. We begin by showing that, despite this limitation, minimal revision is still a successful learning method in a wide range of situations. Firstly, it can learn any problem that is finitely identifiable. Secondly, it can learn with positive and negative data, as long as one considers finitely many possibilities. We then characterize the prior plausibility assignments (over finitely many possibilities) that enable one to learn via minimal revision, and do the same for conditioning and lexicographic upgrade. Finally, we show that not all of our results still hold when learning from possibly erroneous information.

1 Introduction

Many human activities aim to find the truth. Consider, for instance, a scientist deciding which of two possible theories, T_1 or T_2, is the correct one. At the start of their inquiry, the scientist might have some *prior plausibility ordering* that reflects their preference among the two. Over time, new experimental results and observations become available to the scientist, who revises their assessment accordingly using some *revision method*. How should the scientist revise their prior assessment? Some inspiration about the structure of that dynamics can be drawn from belief revision theory. In particular, one natural way to revise one's beliefs is to accept the newly learned information but change the overall belief state as little as possible. This *minimal revision* method (also known as "conservative upgrade") was first introduced in [13], where it was shown that the method is consistent with the basic AGM principles of (iterated) belief revision [1], that it naturally accommodates the Ramsey test for conditionals [31], and that, while continually updating the propositional beliefs, it preserves as many as possible of the extended, conditional beliefs.

Despite its natural appeal, minimal revision falls short in its *learning power*, i.e., in its ability to successfully lead an agent to identify the true state of the world among a set of possibilities. In particular, using tools from formal learning theory, [23] and [17, 4, 6] show that, unlike the more radical belief revision methods of *conditioning* (also known as "update") and *lexicographic revision* (also known as "radical upgrade"), it does not excel as a learning method, because it suffers from memory limitations [17] and has a tendency to fall into cycles [7]. Given these negative results, other revision methods appear *in general* preferable to minimal revision to track the truth. However, the much simpler minimal revision could still be the right method when considering *particular classes* of learning scenarios. Building on the framework introduced in [17, 4, 6], we show that this is indeed the case.

After recalling the framework from [6] in Section 2, we show in Section 3 that: Minimal revision can successfully learn any problem that is finitely identifiable (Section 3.1); It can learn on positive and negative data if the possibilities considered are finite (Section 3.2); Only some prior plausibility

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 301–320, doi:10.4204/EPTCS.437.25

© E. Baccini, Z. Christoff, N. Gierasimczuk, R. Verbrugge
This work is licensed under the
Creative Commons Attribution License.

assessments can lead minimal revision to success, and we characterize the class of such plausibility assessments when one considers finitely many possibilities (Section 3.3); Not all our results are preserved when learning by minimal revision with possibly erroneous observations (Section 3.4). In Section 4, we consider conditioning and lexicographic revision and characterize the prior plausibility assessments which allow one to successfully learn via these methods. We conclude in Section 5.

2 Background and Preliminary Definitions

We consider a discrete-time iterative process ordered like the natural numbers: At each time-step, an agent receives an observation from the world, on the basis of which they revise their plausibility order and make a conjecture about which state is the actual one. The aim of the agent is to correctly identify the actual state of the world. Below, we recall the framework developed in [17, 4, 6].

Definition 1 ([6, Def. 1])**.** *An **epistemic space S** is a pair $\langle S, O\rangle$, where S is a set of possible worlds (at most countable), and $O \subseteq \mathscr{P}(S)$ is a set of observables (at most countable). For every $s \in S$, O_s denotes the set of observables $p \in O$ such that $s \in p$. For any two worlds $s,t \in S$, if $O_s = O_t$ then $s = t$ (no two worlds satisfy exactly the same set of observables). An epistemic space is said to be finite if $|S| < |\mathbb{N}|$.*

The set of states represents the possibilities entertained by an agent, e.g., theories a scientist considers possible. The set of observables represents the observations the agent might come across. Figure 1 provides an example of an epistemic space (on the left). To represent an agent's doxastic state, we enrich the space with a plausibility relation to obtain a plausibility space (Figure 1, right) [21, 13, 8].

Definition 2 ([6, Def. 2])**.** *A **plausibility space B** is the tuple $\langle \mathbf{S}, \preceq\rangle$ where $\mathbf{S} = \langle S, O\rangle$ is an epistemic space and $\preceq$ is a total preorder, i.e., a total, transitive and reflexive relation over S.*

For $s,t \in S$, $s \preceq t$ means that "s is at least as plausible as t". We write $s \prec t$ if $s \preceq t$ and $t \not\preceq s$; $s \simeq t$ if $s \preceq t$ and $t \preceq s$. Since in Section 4 we will also consider preorders that are not necessarily total, we define additional notation: $s \sim t$ if $s \preceq t$ or $t \preceq s$; $s \not\sim t$ if not $s \sim t$.

The assumption that the preorder be total is common in the dynamic epistemic logic literature and is rooted in Grove's semantics [21], inspired in turn by Lewis' sphere system for counterfactuals [26]. Plausibility spaces allow for a simple representation of belief as truth in the most plausible worlds [8]: An agent believes a proposition p iff p is true in all states that the agent considers most plausible.

Definition 3. *Let a preorder $\preceq$ (not necessarily total) over a set of elements S be given. An element $s \in S$ is a **$\preceq$-minimal element** of S iff for all $t \in S$ such that $t \sim s$, $s \preceq t$. We denote by $min_{\preceq}X$ the set of $\preceq$-minimal elements of the set $X \subseteq S$. For every $p \in O$, we define the preorders $\preceq_p := \preceq \cap (p \times p)$ and $\preceq_{\bar{p}} := \preceq \cap (S \setminus p \times S \setminus p)$.*

The observations that an agent receives (in a step-wise manner) form an infinite stream:

Definition 4 ([6, Def. 4, 5])**.** *A **data sequence** σ is a finite sequence of elements of O; a **data stream** $\vec{O}$ is an infinite sequence of elements in O. Given a data sequence $\sigma = (O_0, O_1, \ldots, O_n)$, we denote by $set(\sigma)$ the set of elements of O that occur in σ; given a data stream $\vec{O} = (O_0, O_1, \ldots)$, we let $set(\vec{O})$ be the set of elements of O that occur in $\vec{O}$. Given $n \in \mathbb{N}$, we denote by O_n the n^{th} observation in $\vec{O}$. $\vec{O}[k]$ is the initial segment of $\vec{O}$ of length k. $\sigma * \vec{O}$ is the concatenation of the sequence σ with the data stream $\vec{O}$. The empty sequence is denoted by ε.*

We consider streams containing all and only observations that are true in the actual world.

Definition 5 ([6, Defs. 6, 7])**.** *A **data sequence** σ is: **sound with respect to a state** s iff $set(\sigma) \subseteq O_s$; **complete with respect to a state** s iff $O_s \subseteq set(\sigma)$. A **data stream** $\vec{O}$ is: **sound with respect to a state** s iff $set(\vec{O}) \subseteq O_s$; **complete with respect to a state** s iff $O_s \subseteq set(\vec{O})$.*

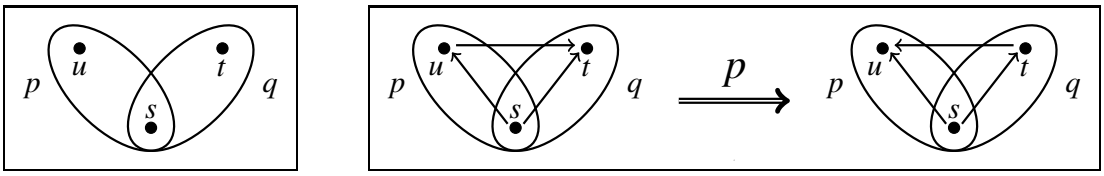


Figure 1: On the left: An epistemic space with $S = \{u,s,t\}$, $O = \{p,q\}$. On the right: a plausibility order on the same space with $t \prec u \prec s$ and the revised order $u \prec t \prec s$ after observing p (represented by the p-labelled arrow). This is an example of (one-step) *minimal revision* (Def. 12). An arrow from one world to another indicates that the latter is more plausible than the former. We omit reflexive arrows.

At each time-step, a learner uses some belief revision method to revise their plausibility order on the basis of the new observation received. Formally:

Definition 6 ([6, Def. 10])**.** *A **one-step belief revision** is a function R_1 taking a plausibility space $\boldsymbol{B}$ and an observable $p \in O$ to output the plausibility space $R_1(\boldsymbol{B}, p)$. A one-step belief revision function R_1 induces an **iterated belief revision function** R in the following way: $R(\boldsymbol{B}, \varepsilon) = \boldsymbol{B}$, $R(\boldsymbol{B}, \sigma * p) = R_1(R(\boldsymbol{B}, \sigma), p)$.*

Any belief-revision method can be used to define a corresponding learning method.

Definition 7 ([6, Def. 8])**.** *A **learning method** L is a function that takes as input an epistemic space $\boldsymbol{S} = \langle S, O\rangle$ and a data sequence σ and outputs a conjecture $L(\boldsymbol{S}, \sigma) \subseteq S$.*

Definition 8 ([6, Def. 11])**.** *A **plausibility assignment PLAUS** is a function that takes as input an epistemic space $\boldsymbol{S} = \langle S, O\rangle$ and assigns to it a plausibility order $\preceq$, converting it into a plausibility space $\boldsymbol{PLAUS}(S) = \langle \boldsymbol{S}, \preceq\rangle$.*

Definition 9 ([6, Def. 12])**.** *Let a belief-revision method R be given. Given any epistemic space $\boldsymbol{S}$ and plausibility assignment **PLAUS**, a belief-revision method R induces a **canonical learning method** in the following way: $L_R^{PLAUS}(\boldsymbol{S}, \sigma) := min\ R(PLAUS(\boldsymbol{S}), \sigma)$.*

Given an epistemic space **S**, a plausibility assignment **PLAUS**, a one-step belief-revision method R_1, and $\vec{O}[n]$, we denote with $\preceq^{R_1}_{\vec{O}[n]}$ the plausibility order in the updated epistemic space $R(\mathbf{PLAUS}(S), \vec{O}[n])$.

As the central criterion of successful learning we take *identifiability in the limit*, which originally comes from the studies of inductive grammar inference [20], and is often combined with computational restrictions on learners and the classes they learn. In this paper, we abstract away from such constraints and focus on purely structural considerations.

Definition 10 ([6, Def. 18], originally in [20])**.** *Let an epistemic space $\boldsymbol{S} = \langle S, O\rangle$ be given. A **state** $s \in S$ is **identified in the limit by a learning method** L iff for any sound and complete data stream $\vec{O}$ for s, there is an $n \in \mathbb{N}$ such that for all $k \geq n$, $L(\boldsymbol{S}, \vec{O}[k]) = \{s\}$. The **epistemic space S is identified in the limit by learning method** L iff all $s \in S$ are identified in the limit by learning method L. The **epistemic space S** is **identifiable in the limit (learnable)** iff it is identified in the limit by some learning method.*

Instead of such general, unrestricted learning functions, we will focus on more constructive ones, namely, those guided by some underlying belief revision methods.

Definition 11 ([6, Def. 19])**.** *An **epistemic space S is identifiable in the limit by the belief revision method** R if there exists a prior plausibility assignment $PLAUS$ such that L_R^{PLAUS} identifies S in the limit. A **belief revision method** R is **universal on a class** C of epistemic spaces if every space in C is identifiable by the belief revision method. A **belief revision method** R is **universal (tout court)** if it is universal in the class of all learnable spaces.*

3 Learning with Minimal Revision

What is the least disruptive way of revising plausibility spaces? Such a minimal policy was first discussed in [30] and then generalised to total preorders in [13]. Upon receiving p, minimal revision moves all the most plausible p-worlds in front of all other worlds, leaving the rest of the plausibility order unchanged. Minimal revision is *conservative* in the sense that it leaves as much as possible of the old plausibility ordering unchanged [13, 17].[1] In our setting, it is defined in the following way.

Definition 12 (Minimal Revision, [8])**.** *Let $\boldsymbol{B} = \langle S, O, \preceq \rangle$ be a plausibility space and $p \in O$. The one-step revision function $\boldsymbol{mini}(\boldsymbol{B}, p)$ generates the plausibility space $\boldsymbol{B'} = \langle S, O, \preceq' \rangle$, where $\preceq'$ is such that if $s \in min_{\preceq} p$ and $t \notin min_{\preceq} p$, then $s \preceq' t$ and $t \not\preceq' s$; $s \preceq' t$ iff $s \preceq t$, otherwise.*

Minimal revision suffers from a number of shortcomings. For instance, it can fail to stabilize when revising with true higher-order information [7], and it fails to learn the least complex problems in the learning setting of [23]. With respect to the framework adopted in this paper, it has been shown that minimal revision is not universal, in the sense that not all spaces that are learnable can be learned via minimal revision [17, 4, 6]. Consider the space in Figure 1 from [6]. First, note that the space is learnable by the cruder *conditioning* method (Definition 19): At each time-step, eliminate all worlds at which the new information is false. In this case, any initial plausibility order would do, as long as $u \prec s$ and $t \prec s$. Yet, there is no plausibility assignment allowing identification of this space in the limit via minimal revision: In order to identify either u or t, both of them must be strictly more plausible than s; but then s is not identified, since no observation will ever suffice to make s strictly more plausible than u and t.

We are now ready to state our main research question: Even if minimal revision cannot identify all learnable epistemic spaces, are there interesting classes of epistemic spaces that it can identify? In the sections below, we give a positive answer to this question and consider some notable classes of spaces in which minimal revision is universal. All omitted proofs are included in the Appendix.

3.1 Minimal Revision is Universal on the Class of Finitely Identifiable Spaces

The success criterion of identifiability in the limit in Definition 10 is often contrasted with a stronger criterion of conclusive learning (see [19, 14]), adapted from finite identifiability as proposed in [16, 24, 22]. Finite identifiability was characterized in [27], studied, e.g., in [25, 32], and applied to problems in AI, e.g., in [11, 12, 29]. It requires not only that the learner *converges* to a correct conjecture, but also that they know with certainty that the conjecture is true. The latter condition of certainty can be rendered as the requirement that for any stream the learner is a once-defined function, i.e., throughout the learning process it only has one-shot at the correct guess [19]. In this section, we will show that **mini** can identify in the limit the class of epistemic spaces that can be finitely identified, i.e., those spaces for which there exists a learner that can identify the actual world with a single guess.

Definition 13 ([27])**.** *Let an epistemic space $\boldsymbol{S} = \langle S, O \rangle$ be given. A **state** $s \in S$ is **finitely identified** by a learning method L iff for any sound and complete data stream $\vec{O}$ for s, when inductively given $\vec{O}$, L outputs at some point a single conjecture s. The **epistemic space S** is **finitely identified** by learning method L iff all $s \in S$ are finitely identified by learning method L. The **epistemic space S** is **finitely identifiable** (**conclusively learnable**) just in case it is finitely identified by some learning method.*

In [27], in a computable setting, it has been shown that this class of spaces is characterized by the existence of the following map:

[1] For a systematic analysis of the learning-theoretic properties of the canonical learning methods induced by minimal revision and other belief revision methods, see [17].

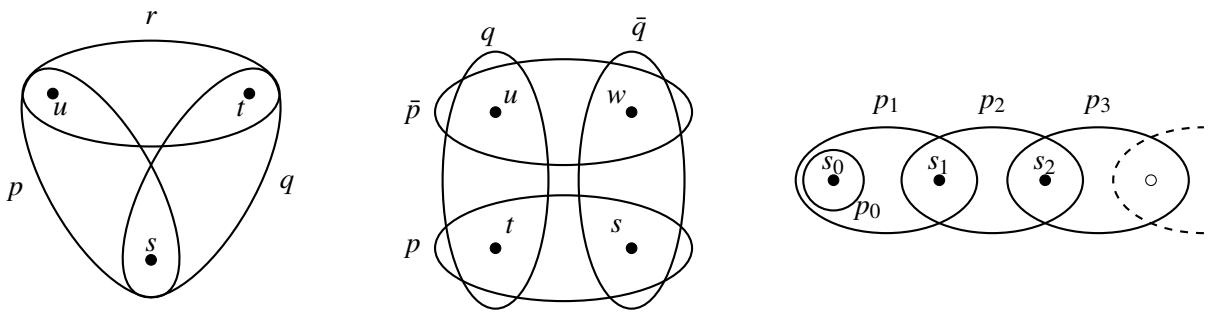


Figure 2: Example of three epistemic spaces that are finitely identifiable. On the left: $S = \{s,u,t\}$, $O = \{p,q,r\}$. At the center: $S = \{s,u,t,w\}$, $O = \{p,\bar{p},q,\bar{q}\}$. On the right: $S = \{s_i | i \in N\}$, $O = \{p_i : i \in \mathbb{N}\}$, and for all s_i, $O_i = \{p_j \in O : j = i \text{ or } j = i+1\}$.

Definition 14 ([18, Def. 10.8], originally in [27])**.** *Let $\boldsymbol{S} = \langle S, O\rangle$ be an epistemic space, and let $s \in S$ be given. A **definite finite tell-tale map** is a total map assigning to each $s \in S$ a **definite finite tell-tale set** $D_s \subseteq O$ such that: (i) D_s is finite; (ii) $D_s \subseteq O_s$; (iii) for any $t \in S$, if $D_s \subseteq O_t$, then $s = t$.*

Informally, the existence of a definite finite tell-tale map on a space means that each world is uniquely identified by some finite conjunction of observables. Fig. 2 shows examples of spaces with definite finite tell-tale maps. The characterisation below adapts to our setting the original theorem in [27].

Proposition 1. *An epistemic space **S** is finitely identifiable iff it has a definite finite tell-tale map.*

We show now that **mini** can learn the class of finitely identifiable spaces, i.e., it can identify this class in the limit. We do this by showing that **mini** can learn any space that has a definite finite tell-tale map. We start with the following Lemma:

Lemma 1. *Let an epistemic space $\boldsymbol{S} = \langle S, O\rangle$ be given. Let **PLAUS** be the plausibility assignment that assigns to **S** the preorder $\preceq$ such that for all $s,t \in S$, $s \simeq t$. Then, for any state s and for any sound data-stream $\vec{O}$ for s, $s \in L^{\boldsymbol{PLAUS}}_{\boldsymbol{mini}}(\boldsymbol{S}, \vec{O}[n])$ for all $n \in \mathbb{N}$.*

Proof. We prove by induction on n that $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[n]}}$. [Base Case: $n = 0$] The sequence $\vec{O}[0]$ is empty, and as such the plausibility model we consider is the initial plausibility model. Since for all $t \in S$, $s \preceq t$, we have that s is a minimal element of $\preceq$. [Inductive Step] Assume as inductive hypothesis that for all $m \leq n$, it is the case that $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}}$. Consider now $n+1$, and consider the observable p occurring in the position $n+1$ in the stream $\vec{O}$, i.e., the observable after the sequence $\vec{O}[n]$. By assumption, $\vec{O}$ is sound with respect to s, and thus $s \in p$. In addition, by inductive hypothesis we have that $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[n]}} p$. Hence, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[n+1]}}$ by Def. 12. Therefore, in any sound and complete data-stream $\vec{O}$ for all $n \in \mathbb{N}$, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[n]}}$, and thus $s \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(\mathbf{S}, \vec{O}[n])$. □

Theorem 1. *Minimal revision is universal on the class of spaces that are finitely identifiable.*

Proof. Let **S** be an epistemic space that is finitely identifiable. Then, by Prop. 1, **S** has a definite finite tell-tale map. Let the plausibility order $\preceq$ be given such that for all $t,s \in S$, $s \simeq t$. Assume towards a contradiction that the learner induced by **mini** and $\preceq$ does not identify the space S, i.e., there is a state $s \in S$ and a sound and complete data-stream $\vec{O}$ for s such that there is no $n \in \mathbb{N}$ such that for all $k \geq n$, $L^{\mathbf{PLAUS}}_{\mathbf{mini}}(\mathbf{S}, \vec{O}[k]) = \{s\}$. Consider such a data-stream. Since $\vec{O}$ is sound and complete with respect to s,

and since there exists a definite finite tell-tale set D_s for s, there is some $n \in \mathbb{N}$ such that $D_s \subseteq set(\vec{O}[n])$. Consider such a sequence $\vec{O}[n]$. Since s is not identifiable in the limit, there exists a $k \geq n$, such that $L^{\mathbf{PLAUS}}_{\mathbf{mini}}(\mathbf{S}, \vec{O}[k]) \neq \{s\}$. By Lemma 1, it must be the case that $s \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(\mathbf{S}, \vec{O}[k])$. Therefore, there must exist a $t \neq s$ such that $t \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(\mathbf{S}, \vec{O}[k])$. It follows that $t \in \bigcap set(\vec{O}[k])$ (otherwise, it would have become strictly less plausible than s for some $k' \leq k$). Since $D_s \subseteq set(\vec{O}[k])$, we have $t \in \bigcap D_s$. By Def. 14, we have $t = s$. Contradiction. □

The theorem above states that any space that is finitely identifiable can be learned via **mini** by using an initially 'unbiased' plausibility order, which considers all possibilities equally plausible. This plausibility assignment together with **mini** induces a method that patiently waits for enough data to dismiss all but one possibility. Note, however, that starting from a prior that ranks all states as equally plausible is not necessary for learning finitely identifiable spaces via **mini**, i.e., there are finitely learnable spaces that can be learned via **mini** and a plausibility assignment where for at least two states s,t, $s \prec t$. For instance, the space on the right in Figure 2 can be identified by ranking the states in a way such that $s_i \preceq s_j$ iff $i \leq j$.

Let us also note that **mini** can identify spaces that are not finitely identifiable and thus the class of finitely identifiable epistemic spaces is properly included in the class of spaces on which **mini** is universal. For instance, the epistemic space with $S = \{s,t\}$, $O = \{p,q\}$, $O_s = \{p,q\}$ and $O_t = \{q\}$ is a space identifiable via **mini** using the plausibility order $t \prec s$. This space is not finitely identifiable since $O_t \subseteq O_s$ and therefore $t \in S$ has no definite finite tell-tale set.

3.2 Minimal Revision is Universal on Positive and Negative Data on Finite Spaces

In the inductive inference literature, it is common to distinguish between learning from only positive information and from positive and negative information. This distinction is given a lot of attention, because the latter setting is much more informative, and so it leads to different learning powers of the same learning methods (see, e.g., [3]). In our case, learning from positive and negative information means that for every observable its complement can also be observed. In logic parlance, this is expressed as closure under negation (i.e., if p can be observed, so can $\neg p$). We will therefore speak of epistemic spaces that are closed under negation [6]. Negation-closed spaces are a subclass of the class of *strongly separated* spaces [5].

Definition 15. *An epistemic space $\mathbf{S} = \langle S, O \rangle$ is:* ***strongly separated*** *when for every $s,t \in S$, $O_s \not\subseteq O_t$;* ***negation-closed*** *iff for every $p \in O$, there is $\bar{p} \in O$ such that $p = S \setminus \bar{p}$.*

Figure 2 shows an example of a strongly separated space that is not negation-closed (left) and an example of a negation-closed space (center). From [17], we know that **mini** is not universal on positive and negative data on the class of all negation-closed spaces, and thus the same holds for the larger class of strongly separated spaces. Nevertheless, we can show that **mini** is in fact universal on the class of *finite* strongly separated spaces and thus on the class of *finite* negation-closed spaces.

Lemma 2. *Let $\mathbf{S}$ be a finite strongly separated epistemic space. Then $\mathbf{S}$ has a definite finite tell-tale map.*

Proof. Consider a strongly separated epistemic space $\mathbf{S}$. Take an arbitrary$s \in S$. Consider the set $D^s := \bigcup_{s' \in S \setminus \{s\}} (O_s \setminus O_{s'})$. Since the space is strongly separated, for all $s' \in S$ such that $s' \neq s$, $O_s \setminus O_{s'} \neq \emptyset$. Thus for all $s' \neq s$, $s' \notin \bigcap D^s$. Since the space is finite, D^s is finite, and $D^s \subseteq O_s$ by construction. By Def. 14, D^s is a definite finite tell-tale set for s. Since s was arbitrary, S has a definite finite tell-tale map. □

Theorem 2. *Minimal revision is universal on the class of finite strongly separated epistemic spaces.*

Proof. By Lemma 2 and Prop. 1, the class of finite strongly separated epistemic spaces is a subset of the class of finitely identifiable spaces. By Th. 1, **mini** is universal on the class of finitely identifiable spaces, and thus on the class of finite strongly separated spaces. □

Corollary 1. *Minimal revision is universal on positive and negative data on the class of finite negation-closed epistemic spaces.*

Therefore, one can safely learn from positive and negative data using **mini** when only finitely many states are possible. Let us note here that this result has been foreshadowed by Boutilier in his seminal work [13], where minimal revision is applied to finitely-grounded knowledge sets over propositional logic (which is closed on negation).

3.3 What Priors for Minimal Revision?

We have considered above some classes of spaces where **mini** is universal. However, in those and all other classes on which **mini** is universal, only some plausibility orders will allow learning via **mini**. We call these orders *appropriate*:

Definition 16. *Let **S** be an epistemic space. A **plausibility order** $\preceq$ is **appropriate for learning S via the belief revision method** R if the canonical learning method induced by R and the plausibility assignment assigning $\preceq$ to S identifies **S** in the limit.*

Figure 3 shows a plausibility order that is not appropriate for learning a space via **mini**: The state s (in red) is not identified with the (sound and complete) data-stream in which q occurs once first, followed by an infinite stream of ps. With this stream, the state u will forever remain equi-plausible to s, and thus s will never be the uniquely minimal element of the space.

Below, we give a characterisation of the class of appropriate plausibility orders for learning via **mini** on *finite* epistemic spaces. Our focus on finite spaces is justified by the fact that they are often sufficient for a variety of applications in belief revision theory, knowledge representation, and dynamic epistemic logic (where possible worlds are identified with valuations over a finite set of propositions, see, e.g., [15]). Furthermore, the full characterisation of the class of spaces for which **mini** is universal is still unknown, so we cannot resort to knowledge about the underlying structure of the epistemic space in our analysis.

Let us remark that our focus here is different from that in [6], where the focus is on whether a belief revision method is universal in the sense that for any learnable space one can construct *some* plausibility order that guarantees successful learning via that method. The interest in [6] therefore lies in the existence of a suitable plausibility order on every learnable space. In contrast, we aim at characterising for **mini** (and later in Section 4 also for conditioning and lexicographic upgrade) the class of *all* prior plausibility orders that are suited for learning every space that is learnable via the method under consideration. Of course, an appropriate order for a method can only exist on spaces that can be learned via that method, and thus [6] provides the limits of applicability of our analysis. Let us start with some preliminary lemmas.

Lemma 3. *If a plausibility order $\preceq$ is appropriate to learn an epistemic space **S** via **mini**, then for all $s \in S$, there exists a $p \in O_s$ such that $s \in min_{\preceq} p$.*

Proof. Assume that $\preceq$ is appropriate to learn **S** via **mini**. Assume towards a contradiction that there exists an s such that for all $p \in O_s$, $s \notin min_{\preceq} p$. Consider now a sound and complete data stream such that it enumerates all the elements in O_s. We show by induction on the length n of any finite segment of $\vec{O}[n]$

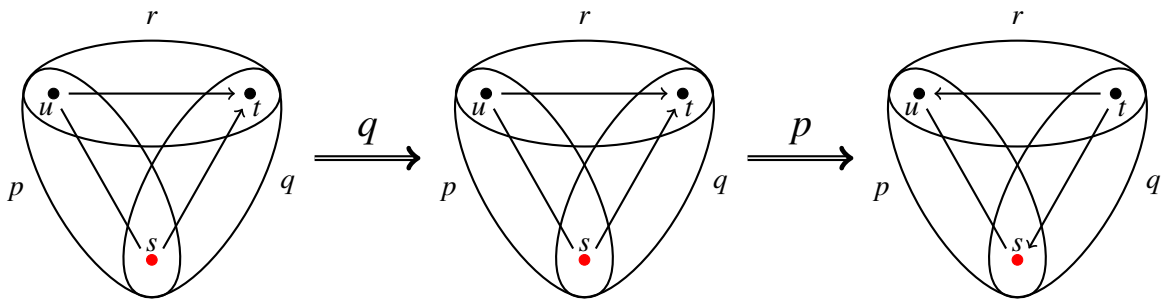


Figure 3: Example of a plausibility order that is not appropriate to learn a space via **mini**. The epistemic space is given by $S = \{s,u,t\}$ and $O = \{p,q,r\}$; the prior plausibility order (leftmost space) is $t \prec s$, $t \prec u$, $u \simeq s$. The labelled arrows represent two revision steps, first when observing q, and then p.

that for all $p \in O_s$ there exists a t such that $t \prec^{\mathbf{mini}}_{\vec{O}[n]} s$ and $t \in p$. [Base case $n = 0$] Follows directly from the assumption. [Inductive step: $n = m+1$] Assume that for $O[m]$, for all $p \in O_s$ there is a t such that $t \in p$ and $t \prec^{\mathbf{mini}}_{O[m]} s$. Consider now $p \in O_s$ that occurs in $\vec{O}$ after $O[m]$. By inductive hypothesis, it follows that there is a $t \in S$ such that $t \prec^{\mathbf{mini}}_{O[m]} s$ and $t \in p$. Hence $t \prec^{\mathbf{mini}}_{O[m+1]} s$. Hence $s \notin min_{\preceq^{\mathbf{mini}}_{\vec{O}[m+1]}}$. □

The above lemma expresses that an appropriate plausibility order guarantees that in every state it is possible to make an observation that 'favors' that state.

Lemma 4. *Suppose that a plausibility order $\preceq$ is appropriate to learn an epistemic space $\boldsymbol{S}$ via **mini**. For all $s \in S$, for all $t \in S$ such that $s \simeq t$ and $t \neq s$, if $|\{p \in O_s : t \notin p\}| < |\mathbb{N}|$, then there exists a $p \in O_s$ such that $s \in min_{\preceq} p$ and $t \notin p$.*

Proof. Towards a contradiction, assume that there are $s,t \in S$ such that: $s \simeq t$, $t \neq s$; $|\{p \in O_s : t \notin p\}| < |\mathbb{N}|$; for all $p \in O_s$ if $s \in min_{\preceq} p$ then $t \in p$. Let $\vec{O}$ be a stream and $n \in \mathbb{N}$ be such that $set(\vec{O}[n]) = \{p \in O_s : t \notin p\}$, which exists since $|\{p \in O_s : t \notin p\}| < |\mathbb{N}|$. Denote with $\preceq^{\mathbf{mini}}_{\vec{O}[n]}$ the order in $\mathbf{mini}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. Since by assumption for all $p \in O_s$ if $t \notin p$ then $s \notin min_{\preceq} p$, and since by construction for all $p \in set(\vec{O}[n])$, $t \notin p$, we have that $s \notin min_{\preceq} p$ for all $p \in set(\vec{O}[n])$. It follows that $s \simeq^{\mathbf{mini}}_{\vec{O}[n]} t$ (otherwise one of them would be a minimal element of some $p \in \vec{O}[n]$). Consider now any extension of the sequence $\vec{O}[n]$ that enumerates all remaining elements of O_s, i.e. all elements in $O_s \setminus \{p \in O_s : t \notin p\}$. By construction all $p \in set(\vec{O})$ occurring after the initial sequence $\vec{O}[n]$ are such that $t \in p$. Since $t \simeq^{\mathbf{mini}}_{\vec{O}[n]} s$ and $\vec{O}$ is sound, it follows that for all $m \geq n$, $t \simeq^{\mathbf{mini}}_{\vec{O}[m]} s$ by Def. 12. Thus, there is no m such that $min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}} = \{s\}$. Hence, $\preceq$ is not appropriate for learning S via **mini**. □

The lemma 4 ensures that an appropriate plausibility order cannot be deceived by sound and complete data streams such as the one illustrated in Figure 3. The problem there is that a distinct state from s, namely u, is equi-plausible to s and minimal in all observables in which s is minimal. Lemma 4 excludes such configurations of points in finite epistemic spaces.

Lemma 5. *Let $\boldsymbol{S} = \langle S, O\rangle$ be an epistemic space, and let $\preceq$ be a plausibility order over S. Consider a state s and assume that there exists a $p \in O_s$ such that $s \in min_{\preceq} p$. Then for any finite sequence $\vec{O}[n]$ of a sound data stream $\vec{O}$ for s, it holds that $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[n]}} p$.*

Proof. We prove this by induction on n. [Base case: $n = 0$] Follows directly from the assumption. [Inductive step] Assume that for all $m < n$, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}} p$. Consider now the p' occurring in $\vec{O}$ after the initial segment $\vec{O}[m]$. Assume towards a contradiction that $s \notin min_{\preceq^{\mathbf{mini}}_{\vec{O}[m+1]}} p$. Hence there is $t \in p$ such that $t \prec^{\mathbf{mini}}_{\vec{O}[m+1]} s$. Since $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}} p$ by inductive hypothesis, then $t \not\prec^{\mathbf{mini}}_{\vec{O}[m]} s$, hence $s \preceq^{\mathbf{mini}}_{\vec{O}[m]} t$, since $\preceq$ is total. Since $s \notin min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}} p'$ (otherwise it would be a minimal element in p after updating on p'), there is $t' \prec^{\mathbf{mini}}_{\vec{O}[m]} s$, and $t \in p'$. By transitivity, $t' \prec^{\mathbf{mini}}_{\vec{O}[m]} t$. Hence, $t \notin min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}} p'$. Thus by Def. 12, $s \preceq^{\mathbf{mini}}_{\vec{O}[m+1]} t$, which contradicts the assumption that $t \prec^{\mathbf{mini}}_{\vec{O}[m+1]} s$. □

To characterise the plausibility orders appropriate for minimal revision, we make use of the concept of a *finite tell-tale map*. We have already seen a concept of this kind in Definition 14. The concept was first introduced by Dana Angluin in her seminal work [2]. Broadly speaking, a finite tell-tale map assigns to each possible state a special finite set of observables (a finite tell-tale), which guides learning and guarantees that the state can be identified. We consider a specific variant of tell-tale maps that satisfies a number of conditions regarding not only the available observables but also the plausibility order of the states.

Definition 17. *Let $\mathbf{S} = \langle S, O\rangle$ be a finite epistemic space and $\preceq$ be a plausibility order over S be given. A **finite space mini tell-tale map** for $\preceq$ and $\mathbf{S}$ is a total map assigning to each state $s \in S$ a **finite space mini tell-tale set** $F_s \subseteq O$ such that: (i) $F_s \subseteq O_s$; (ii) there is a $p \in F_s$ such that $s \in min_{\preceq} p$ and for all $t \neq s$, $t \simeq s$ such that $t \in min_{\preceq} p$, there exists $q \in F_s$ such that $s \in min_{\preceq} q$ and $t \notin min_{\preceq} q$.*

Theorem 3. *Let $\preceq$ be a plausibility order on a finite epistemic space $\mathbf{S} = \langle S, O\rangle$. The plausibility order $\preceq$ is appropriate to learn $\mathbf{S}$ via **mini** iff there exists a finite space mini tell-tale map for S and $\preceq$.*

Proof. [⇒] Let $\mathbf{S} = \langle S, O\rangle$ be finite, and let $\preceq$ be appropriate for learning $\mathbf{S}$ via **mini**. Since $O_s \subseteq O_s$, it trivially satisfies (i) from Def. 17. Assume towards a contradiction that for all $p \in O_s$, either $p \notin min_{\preceq} p$ or there exists a $t \neq s$ such that $t \in min_{\preceq} p$ and there is no $q \in O_s$ such that $s \in min_{\preceq} q$ and $t \notin min_{\preceq} q$. By Lemma 3 we know that there is a $p \in O_s$ such that $s \in min_{\preceq} p$. By assumption, it follows that there exists a $t \in min_{\preceq} p$ such that there is no $q \neq p \in O_s$ such that $s \in min_{\preceq} q$ and $t \notin min_{\preceq} q$. By Lemma 4, it follows that the set $|\{p \in O_s : t \notin p\}| = |\mathbb{N}|$. Contradiction.

[⇐] Let $\mathbf{S} = \langle S, O\rangle$ be a finite epistemic space, and let $\preceq$ be given. Assume there exists a finite space mini tell-tale map for $\preceq$ and S. Let $s \in S$ be arbitrary and F_s be its corresponding finite space mini tell-tale set. Since S is finite, also O_s is a finite space mini tell-tale set. Consider an arbitrary sound and complete stream $\vec{O}$ for s. Since O_s is finite and $\vec{O}$ is complete, there exists an $n \in \mathbb{N}$ such that $O_s \subseteq set(\vec{O}[n])$. Since O_s is a finite space mini tell-tale set for s, by Def. 17 there exists a $p \in O_s$ such that $s \in min_{\preceq} p$. Let $\vec{O}[m]$ be the sequence where the last element is the first occurrence of a $p \in O_s$ such that $s \in min_{\preceq} p$ in $\vec{O}$. By Lemma 5, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m-1]}} p$. By Def. 12, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}}$, and thus $s \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(S, O[m])$. If there is no $t \neq s$ such that $t \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(S, O[m])$, then for all $m' \geq m$, $L^{\mathbf{PLAUS}}_{\mathbf{mini}}(S, O[m]) = \{s\}$ (since $\vec{O}$ is sound with respect to s will forever be the uniquely minimal element for all $p \in O_s$). If there is a $t \neq s$ such that $t \in L^{\mathbf{PLAUS}}_{\mathbf{mini}}(S, O[m])$, then $t \simeq^{\mathbf{mini}}_{\vec{O}[m]} s$. This implies that $s \simeq t$, and $t \in min_{\preceq} p$. Since by assumption $\mathbf{S}$ has a finite space mini tell-tale, there is q such that $s \in min_{\preceq} q$ and $t \notin min_{\preceq} q$, thus $t \notin q$. Since p was the first occurrence of a $p \in O_s$ such that $s \in min_{\preceq} p$, since $q \in O_s$ and $s \in min_{\preceq} q$, there is $m' > m$ such that q is the last observable occurring in the initial segment $\vec{O}[m']$ of $\vec{O}$, since $\vec{O}$ is complete. Take the smallest such s'. Since $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}}$, and $\vec{O}$ is sound with respect to s, $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m'-1]}}$ and also $s \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m'-1]}} q$, and since $t \notin q$, $s \prec^{\mathbf{mini}}_{\vec{O}[m']} t$ by Def. 12. Since t was arbitrary, for all $t \in S$ such that $t \neq s$ and $t \simeq^{\mathbf{mini}}_{\vec{O}[m]}$, there

is $k > m$ such that $s \prec^{\mathbf{mini}}_{\vec{O}[k]} t$. Since the space is finite, there can only be finitely many $t \neq s$ such that $t \in min_{\preceq^{\mathbf{mini}}_{\vec{O}[m]}}$, and thus there is k' such that for all such $t \in S$, $s \prec^{\mathbf{mini}}_{\vec{O}[k']} t$, and $min_{\preceq^{\mathbf{mini}}_{\vec{O}[k']}} = \{s\}$. By Def. 12 and the soundness of $\vec{O}$, for all $k'' \geq k$, $L^{\mathbf{PLAUS}}_{\mathbf{mini}}(S, O[k'']) = \{s\}$. □

3.4 No Guarantees for Minimal Revision when Learning with Errors

Theorem 3 characterises the appropriate priors when learning via **mini** under the assumption that no erroneous observation can occur (i.e., the data streams are sound). Does the characterisation still hold when considering learning scenarios in which errors can occur? Here, we consider fair streams [6], i.e., data streams that can contain only finitely many errors, all of which will eventually be corrected.

Definition 18 ([6, Def. 27])**.** *Let $\mathbb{S} = \langle S, O\rangle$ be a negation-closed epistemic space. A **data stream** $\vec{O}$ from O is **fair with respect to the state** s precisely in case: (i) $\vec{O}$ is complete with respect to s; (ii) there is an $n \in \mathbb{N}$ such that for all $k \geq n$, $s \in O_k$; (iii) for every $i \in \mathbb{N}$ such that $s \notin O_i$, there is a $k > i$ such that $O_k = \bar{O}_i$.*[2]

When considering fair data streams, not only Theorem 3 is no longer an adequate characterisation of the appropriate plausibility orders to learn via **mini**, but also **mini** is not anymore a universal method on the class of negation-closed finite epistemic spaces.[3]

Proposition 2. *The following two hold: (i) There exist plausibility orders appropriate for learning some negation-closed space via **mini** on sound and complete data streams that are not appropriate for learning the same space via **mini** on fair data streams; (ii) Minimal revision is not universal on fair data-streams on the class of all finite negation-closed epistemic spaces.*

Proof. [Point (i)] Consider the epistemic space at the center of Fig. 2. The space is negation-closed and the plausibility order in which for all $s,t \in S$, $t \simeq s$ is appropriate to identify the space via **mini** on sound and complete data-streams by the proof of Th. 1. Consider the state s and consider the data stream starting with the sequence $q, \bar{p}, p$ followed by an infinite stream of $\bar{q}$. This stream is fair and prevents **mini** from ever identifying s, since w will always be a minimal element with s after the initial segment $q, \bar{p}, p$. [Point (ii)] Consider again the epistemic space at the center of Fig. 2. The only appropriate plausibility order to learn that space via **mini** on sound and complete data streams is such that $s,t \in S, t \simeq s$. This order is not appropriate for learning the space via **mini** on fair data streams as shown above for (i). □

We leave to future work a characterisation of the plausibility assignments that are appropriate for minimal revision on fair data streams (and of the class of spaces on which they exist).

4 What Priors are Appropriate for Stronger Methods?

We turn to considering how one ought to assign priors when employing more radical belief revision methods than minimal revision. In particular, we consider *conditioning* and *lexicographic upgrade*.

Definition 19 (Conditioning [8])**.** *Let $\mathbb{B} = \langle S, O, \preceq\rangle$ be a plausibility space, and let $p \in O$. The one-step revision function **cond**$(\mathbb{B}, p)$ generates the plausibility space $\mathbb{B}' = \langle S', O', \preceq'\rangle$, where $S' = S \cap p$, $O' = \{q \in \mathscr{P}(S') : \exists q' \in O \text{ such that } q = q' \setminus \bar{p}\}$, and $\preceq' = \preceq_p$.*

[2] All definitions in Section 2 and Definition 16 can be straightforwardly adapted for the case of fair data-streams.

[3] **mini**'s sensitivity to errors was highlighted in [28] (Proposition 3.3), where it was applied to cases of cognitive bias, in which some observations are (systematically) misinterpreted by the revising agent.

Definition 20 (Lexicographic Upgrade [8])**.** *Let $\mathbf{B} = \langle S, O, \preceq \rangle$ be a plausibility space, and $p \in O$. The one-step revision function $\mathbf{lex}(\mathbf{B}, p)$ generates the plausibility space $\mathbf{B}' = \langle S, O, \preceq' \rangle$, where $s \preceq' t$ if and only if: $s \preceq_p t$, or $s \preceq_{\bar{p}} t$, or $s \in p$ and $t \notin p$.*

When revising with **cond** upon receiving some observation p, an agent simply eliminates all the worlds that do not satisfy p, and restricts the plausibility order to the worlds that do satisfy p. On the other hand, lexicographic upgrade does not eliminate worlds, but rather moves all the p-worlds in front of the $\neg p$-worlds ([30, 9]).

For the case of conditioning and lexicographic revision, we give a fully general characterisation of appropriate plausibility orders. In particular, we consider spaces that are not necessarily finite, as well as preorders that are not necessarily total.[4] As done for **mini**, we start by defining a tell-tale map variant.

Definition 21. *Let $\mathbf{S} = \langle S, O \rangle$ be an epistemic space and $\preceq$ be a preorder over S be given. A **generalised conditioning tell-tale map** for $\preceq$ and $\mathbf{S}$ is total map assigning to each state $s \in S$ a set $F_s \subseteq O$ such that: (i) F_s is finite; (ii) $F_s \subseteq O_s$; (iii) for all $t \in S$, if $t \sim s$, $F_s \subseteq O_t$ and $s \neq t$, then $s \prec t$; (iv) for all finite $F'_s \supseteq F_s \subseteq O_s$, for all $t \in S$, if $t \not\sim s$ and $F'_s \subseteq O_t$, then there exists a $v \in S$ such that $v \prec t$ and $F'_s \subseteq O_v$.*

One can then prove that the existence of a generalised conditioning tell-tale characterises appropriate preorders for conditioning as well as lexicographic upgrade:

Theorem 4. *Let $\preceq$ be a preorder over an epistemic space $\mathbf{S}$. The following are equivalent: (i) There exists a conditioning tell-tale map for $\mathbf{S}$ and $\preceq$; (ii) $\preceq$ is appropriate to learn $\mathbf{S}$ via **cond**; (iii) $\preceq$ is appropriate to learn $\mathbf{S}$ via **lex**.*

It is easy to see that the equivalence between (i) and (ii) in Theorem 4 fails when considering fair streams (Definition 18): Since **cond** irreparably eliminates the actual state as soon as an error occurs, no preorder is appropriate to learn via **cond** on fair streams. Finally, since any lexicographic upgrade on a finite initial segment of a fair data stream can be shown to be equivalent to the lexicographic upgrade obtained by removing the corrected errors from the sequence [6], the equivalence between (i) and (iii) in Theorem 4 still holds, and thus the existence of a generalised conditioning tell-tale map also characterises the appropriate preorders for learning via **lex** on fair streams in negation-closed spaces.

5 Conclusion

We showed that minimal revision is universal when used to learn finitely identifiable spaces (Theorem 1), and that it is consequently universal on finite strongly separated spaces (Theorem 2). Thus, it can always be used to learn from positive and negative data on finite spaces (Corollary 1) unless erroneous observations possibly occur (Proposition 2). In addition, we characterised the prior plausibility orders that are appropriate for learning finite spaces via minimal revision (Theorem 3) as well as the preorders appropriate for conditioning and lexicographic upgrade (Theorem 4).

Beyond the scope of this paper, several questions are left for future research. First, concerning minimal revision: What is the class of epistemic spaces on which the method is universal? What priors are appropriate for learning via minimal revision on arbitrary spaces, and possibly with wrong information? Second, beyond minimal revision, what priors are appropriate for other revision methods? Finally, we plan to consider appropriate priors from the viewpoint of doxastic attitudes, in the spirit of [14]: What type of doxastic attitudes correspond to holding an appropriate prior?

[4] Definition 16 can be straightforwardly adapted to the case of arbitrary preorders.

Acknowledgements

Rineke Verbrugge acknowledges support from the project "Hybrid Intelligence: Augmenting Human Intellect", a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (grant number 024.004.022). Zoé Christoff acknowledges support from the project "Democracy on Social Networks" (VENI project number Vl.Veni.201F.032) financed by the Netherlands Organisation for Scientific Research (NWO). Edoardo Baccini acknowledges support from the Evert Willem Beth Foundation of the Royal Netherlands Academy of Arts and Sciences (KNAW) (grant number KNAW WF/743 – 13). Edoardo Baccini would also like to thank the "Reasoning, Rationality and Science" research group at the Ruhr-University Bochum for insightful discussion on an earlier version of this manuscript. All authors thank three anonymous referees for helpful comments.

References

[1] Carlos E. Alchourrón, Peter Gärdenfors & David Makinson (1985): *On the Logic of Theory Change: Partial Meet Contraction and Revision Functions*. *Journal of Symbolic Logic* 50(2), pp. 510–530, doi:10.2307/2274239.

[2] Dana Angluin (1980): *Inductive inference of formal languages from positive data*. *Information and Control* 45(2), pp. 117–135, doi:10.1016/S0019-9958(80)90285-5.

[3] Dana Angluin & Carl H. Smith (1983): *Inductive inference: Theory and methods*. *ACM Comput. Surv.* 15(3), pp. 237–269, doi:10.1145/356914.356918.

[4] Alexandru Baltag, Nina Gierasimczuk & Sonja Smets (2011): *Belief revision as a truth-tracking process*. In: *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK XIII, ACM, New York, NY, USA, pp. 187–190, doi:10.1145/2000378.2000400.

[5] Alexandru Baltag, Nina Gierasimczuk & Sonja Smets (2016): *On the solvability of inductive problems: A study in epistemic topology*. In Ram Ramanujam, editor: *Electronic Proceedings in Theoretical Computer Science*, 215, Open Publishing Association, p. 81–98, doi:10.4204/eptcs.215.7.

[6] Alexandru Baltag, Nina Gierasimczuk & Sonja Smets (2019): *Truth-tracking by belief revision*. *Studia Logica* 107, pp. 917–947, doi:10.1007/s11225-018-9812-x.

[7] Alexandru Baltag & Sonja Smets (2011): *Keep changing your beliefs, aiming for the truth*. *Erkenntnis* 75, pp. 255–270, doi:10.1007/s10670-011-9294-y.

[8] Alexandru Baltag & Sonja Smets (2016): *A qualitative theory of dynamic interactive belief revision*. In Horacio Arló-Costa, Vincent F. Hendricks & Johan van Benthem, editors: *Readings in Formal Epistemology: Sourcebook*, Springer International Publishing, Cham, pp. 813–858, doi:10.1007/978-3-319-20451-2_39.

[9] Johan 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.

[10] Lenore Blum & Manuel Blum (1975): *Toward a mathematical theory of inductive inference*. *Information and Control* 28(2), pp. 125–155, doi:10.1016/S0019-9958(75)90261-2.

[11] Thomas Bolander & Nina Gierasimczuk (2015): *Learning actions models: Qualitative approach*. In Wiebe van der Hoek, Wesley H. Holliday & Wen-Fang Wang, editors: *Logic, Rationality, and Interaction - 5th International Workshop, LORI 2015 Taipei, Taiwan, October 28-31, 2015, Proceedings*, *Lecture Notes in Computer Science* 9394, Springer, pp. 40–52, doi:10.1007/978-3-662-48561-3_4.

[12] Thomas Bolander & Nina Gierasimczuk (2018): *Learning to act: Qualitative learning of deterministic action models*. *Journal of Logic and Computation* 28(2), pp. 337–365, doi:10.1093/logcom/exx036.

[13] Craig Boutilier (1993): *Revision sequences and nested conditionals*. In: *Proceedings of the 13th International Joint Conference on Artifical Intelligence - Volume 1*, IJCAI'93, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 519–525, doi:10.5555/1624025.1624098.

[14] Cédric Dégremont & Nina Gierasimczuk (2011): *Finite identification from the viewpoint of epistemic update*. *Information and Computation* 209(3), pp. 383–396, doi:10.1016/j.ic.2010.08.002.

[15] Hans van Ditmarsch, Wiebe van der Hoek & Barteld Kooi (2007): *Dynamic Epistemic Logic*. Synthese Library Series, Springer, Cham, doi:10.1007/978-1-4020-5839-4.

[16] R. V. Freivald & Rolf Wiehagen (1979): *Inductive inference with additional information*. *J. Inf. Process. Cybern.* 15, pp. 179–185. Available at `https://api.semanticscholar.org/CorpusID:598084`.

[17] Nina Gierasimczuk (2010): *Knowing one's limits: Logical Analysis of Inductive Inference*. Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, The Netherlands. Available at `https://eprints.illc.uva.nl/id/eprint/2094/`. ILLC Dissertation Series DS-2010-11.

[18] Nina Gierasimczuk, Vincent F Hendricks & Dick de Jongh (2014): *Logic and learning*. In Alexandru Baltag & Sonja Smets, editors: *Johan van Benthem on Logic and Information Dynamics*, Springer, pp. 267–288, doi:10.1007/978-3-319-06025-5_10.

[19] Nina Gierasimczuk & Dick de Jongh (2012): *On the complexity of conclusive update*. *The Computer Journal* 56(3), pp. 365–377, doi:10.1093/comjnl/bxs059.

[20] E Mark Gold (1967): *Language identification in the limit*. *Information and Control* 10(5), pp. 447–474, doi:10.1016/S0019-9958(67)91165-5.

[21] Adam Grove (1988): *Two modellings for theory change*. *Journal of Philosophical Logic* 17(2), pp. 157–170, doi:10.1007/bf00247909.

[22] Klaus P. Jantke & Hans-Rainer Beick (1981): *Combining postulates of naturalness in inductive inference*. *J. Inf. Process. Cybern.* 17(8/9), pp. 465–484.

[23] Kevin Kelly (1999): *Iterated belief revision, reliability, and inductive amnesia*. *Erkenntnis* 50(1), pp. 7–53, doi:10.1023/A:1005444112348.

[24] R Klette & R Wiehagen (1980): *Research in the theory of inductive inference by GDR mathematicians-A survey*. *Information Sciences* 22, pp. 149–169, doi:10.1016/S0020-0255(80)80006-5.

[25] Steffen Lange & Thomas Zeugmann (1992): *Types of monotonic language learning and their characterization*. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, Association for Computing Machinery, New York, NY, USA, p. 377–390, doi:10.1145/130385.130427.

[26] David Lewis (1973): *Counterfactuals*. Library of Philosophy and Logic, Blackwell, Oxford, doi:10.2307/2273738.

[27] Yasuhito Mukouchi (1992): *Characterization of finite identification*. In: *International Workshop on Analogical and Inductive Inference*, Springer, pp. 260–267, doi:10.1007/3-540-56004-1_18.

[28] Panagiotis Papadamos & Nina Gierasimczuk (2023): *Cognitive bias and belief revision*. In Rineke Verbrugge, editor: *Proceedings Nineteenth conference on Theoretical Aspects of Rationality and Knowledge, TARK 2023, Oxford, United Kingdom, 28-30th June 2023*, *EPTCS* 379, pp. 441–454, doi:10.4204/EPTCS.379.34.

[29] Joseph Singleton & Richard Booth (2024): *Truth-tracking with non-expert information sources*. *J. Artif. Intell. Res.* 81, pp. 619–641, doi:10.1613/JAIR.1.15273.

[30] Wolfgang Spohn (1988): *Ordinal conditional functions: A dynamic theory of epistemic states*. In William L. Harper & Brian 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.

[31] Robert Stalnaker (1968): *A theory of conditionals*. In Nicholas Rescher, editor: *Studies in Logical Theory*, Blackwell, pp. 98–112, doi:10.1007/978-94-009-9117-0_2.

[32] Ana Lucía Vargas-Sandoval (2020): *On the Path to the Truth*. Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, The Netherlands. Available at `https://eprints.illc.uva.nl/id/eprint/2176/`. ILLC Dissertation Series DS-2020-07.

Appendix

Proof of Proposition 1

Proposition 1. *An epistemic space* $\mathbf{S}$ *is finitely identifiable if and only if there exists a definite finite tell-tale map for* $\mathbf{S}$.

Proof. [$\Rightarrow$] Consider an epistemic space $\mathbf{S} = \langle S, O\rangle$ and assume that it is finitely identifiable. Therefore there exists a learner L that finitely identifies every state $s \in S$ on any sound and complete data stream for s. Consider an arbitrary state $s \in S$ and consider an arbitrary sound and complete data stream $\vec{O}$ for s. By Def. 13, it follows that there exists an $n \in \mathbb{N}$ such that $L(S, \vec{O}[n]) = \{s\}$. Define $D_s := set(\vec{O}[n])$. Assume towards a contradiction that D_s is not a definite finite tell-tale set for s. It follows that there exists a $t \in S$ such that $t \neq s$ and $D_s \subseteq O_t$. Consider now a data stream for $\vec{O}_t$ for t such that $\vec{O}_t[n] = \vec{O}[n]$. It follows that $L(\mathbf{S}, \vec{O}_t[n]) = \{s\}$. Thus, L does not finitely identify t contrary to our initial assumption.

[$\Leftarrow$] Consider an epistemic space $\mathbf{S} = \langle S, O\rangle$ and assume there exists a definite finite tell-tale map for $\mathbf{S}$. Let an enumeration $s_0, s_1, ..., s_i, ...$ of the elements of S be given. Define L in the following way: $L(\mathbf{S}, \vec{O}[k]) = \{s_i\}$, where $i \leq k$ is the smallest i such that $D_{s_i} \subseteq set(O[k])$, if such s_i exists and no guess has been made in the past; else, L makes no guess. Consider an arbitrary state s_j and a sound and complete data stream $\vec{O}$ for s_j. Since $\vec{O}$ is complete and D_{s_j} is finite, there exists an $n \in \mathbb{N}$ such that $D_{s_j} \subseteq set(\vec{O}[n])$. Consider the smallest $n' \leq n$ for which $D_{s_j} \subseteq set(\vec{O}[n])$ and $j \leq n'$. Such an n' exists. Consider indeed the least n for which $D_{s_j} \subseteq set(\vec{O}[n])$: If $n \geq j$, then $n' := n$; If $n < j$, set $n' := j$ since $D_{s_j} \subseteq set(\vec{O}[n]) \subseteq set(\vec{O}[j])$. We now need to show that L has not made guesses for any $k < n'$. Assume towards a contradiction that there is $k < n'$, such that $L(\mathbf{S}, \vec{O}[k]) = \{s_{j'}\}$. Since n' is the smallest $n \in \mathbb{N}$ for which $D_{s_j} \subseteq \vec{O}[n]$ and $j \leq n'$, it follows that $D_{s_j} \not\subseteq set(\vec{O}[k])$. Thus $s_{j'} \neq s_j$. By definition of L, it follows that $D_{s_{j'}} \subseteq set(\vec{O})[k] \subseteq O_{s_j}$. But then $D_{s_{j'}}$ is not a definite finite tell-tale for $s_{j'}$, which contradicts our assumption. Since no previous guess has been made by L before observing the sequence $\vec{O}[n']$, $L(\mathbf{S}, \vec{O}[n']) = \{s_j\}$ (if $L(\mathbf{S}, \vec{O}[n']) \neq \{s_j\}$, then there would be $s_i \neq s_j$ with $i < j$ and such that $D_{s_i} \subseteq set(\vec{O}[n']) \subseteq O_{s_j}$, and thus D_{s_i} would not be a definite finite tell-tale set for s_i contrary to our assumptions). By definition of L, no later guess will be output by L. Thus, L finitely identifies s_j. Since s_j was arbitrary, L finitely identifies S. □

Proof of Theorem 4

Theorem 4. *Let* $\mathbf{S}$ *be an epistemic space and let* $\preceq$ *be a preorder over* $\mathbf{S}$. *The following are equivalent: (i) There exists a conditioning tell-tale map for* $\mathbf{S}$ *and* $\preceq$*; (ii)* $\preceq$ *is appropriate to learn* $\mathbf{S}$ *via* ***cond****; (iii)* $\preceq$ *is appropriate to learn* $\mathbf{S}$ *via* ***lex****.*

We prove Theorem 4 by proving the following two propositions:

Proposition 3. *Let* $\preceq$ *be a preorder on an epistemic space* $\mathbf{S} = \langle S, O\rangle$. *The preorder* $\preceq$ *is appropriate to learn* $\mathbf{S}$ *via* ***cond*** *if and only if there exists a generalised conditioning tell-tale map for S and* $\preceq$.

Proposition 4. *Let* $\preceq$ *be a preorder on an epistemic space* $\mathbf{S} = \langle S, O\rangle$. *The preorder* $\preceq$ *is appropriate to learn* $\mathbf{S}$ *via* ***lex*** *if and only if there exists a generalised conditioning tell-tale map for S and* $\preceq$.

Proposition 3 proves that (i) and (ii) in Theorem 4 are equivalent. Proposition 4 proves that (i) and (iii) in Theorem 4 are equivalent.

To prove both propositions, we rely on the following definition and lemma.

Definition 22 (Locking sequence [6, Def. 22], originally in [10])**.** *Let an epistemic space $\boldsymbol{S} = \langle S, O\rangle$, a possible world $s \in S$, a learning method L, and a data sequence σ be given. Sequence σ is called a locking sequence for s if σ is sound with respect to s and for each data sequence τ with $s \in \bigcap set(\tau)$, $L(\boldsymbol{S}, \sigma * \tau) = L(\boldsymbol{S}, \sigma) = \{s\}$.*

Lemma 6 ([6, Lemma 1])**.** *If learning method L identifies possible world s in the limit, then there exists a locking sequence for s and L.*

Proof of Proposition 3

Proposition 3. *Let $\preceq$ be a preorder on an epistemic space $\boldsymbol{S} = \langle S, O\rangle$. The preorder $\preceq$ is appropriate to learn $\boldsymbol{S}$ via **cond** if and only if there exists a generalised conditioning tell-tale map for S and $\preceq$.*

Proof. [$\Rightarrow$] Let a preorder $\preceq$ and an epistemic space $\mathbf{S}$ be given. Assume that the preorder $\preceq$ is appropriate to learn $\mathbf{S}$ via $\preceq$ and **cond**. Let **PLAUS** be the plausibility assignment assigning the preorder $\preceq$ to $\mathbf{S}$. By Def. 16, the learner $L^{\mathbf{PLAUS}}_{\mathbf{cond}}$ identifies the space S in the limit. By Lemma 6, for each state $s \in S$, there exists a locking sequence σ_s for $L^{\mathbf{PLAUS}}_{\mathbf{cond}}$ and s.

Consider an arbitrary s and define $F_s := set(\sigma_s)$. F_s is finite, since σ_s is finite, and thus satisfies point (i) in Def. 21. Furthermore, $F_s \subseteq O_s$ since $set(\sigma_s) \subseteq O_s$, and thus satisfies point (ii) in Def. 21.

We now prove that F_s satisfies (iii) of Def. 21. Assume for contradiction that there is a $t \in S$ such that $t \sim s$, $F_s \subseteq O_t$ and $t \neq s$, and $s \not\preceq t$. Then, $t \simeq s$, or $t \prec s$. [Case: $t \prec s$] Consider now a sound and complete data stream $\vec{O}$ such that there exists n such that $\vec{O}[n] = \sigma_s$. By Def. 19, and the fact that $F_s \subseteq O_t$, t is in the updated space after $\vec{O}[n]$, just like s. Consider the preorder $\preceq^{\mathbf{cond}}_{\vec{O}[n]}$, i.e., the preorder in the updated space $\mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. Since $t \prec s$, we have that $t \prec^{\mathbf{cond}}_{\vec{O}[n]} s$. Thus $s \notin L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n])$, since $s \notin min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$. But by assumption, σ_s is a locking sequence for s and $L^{\mathbf{PLAUS}}_{\mathbf{cond}}$, and thus $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n]) = \{s\}$. [Case: $t \simeq s$] Consider now a sound and complete data stream $\vec{O}$ such that there exists n such that $\vec{O}[n] = \sigma_s$. By Def. 19, and the fact that $F_s \subseteq O_t$, t is in the updated space after $\vec{O}[n]$, just like s. Since $t \simeq s$, we have that $t \simeq^{\mathbf{cond}}_{\vec{O}[n]} s$. By assumption, σ_s is a locking sequence for s and $L^{\mathbf{PLAUS}}_{\mathbf{cond}}$, and thus $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n]) = \{s\}$, and thus $s \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$. But since $t \simeq^{\mathbf{cond}}_{\vec{O}[n]} s$, then $t \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$. Hence $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n]) \neq \{s\}$.

We now prove that F_s satisfies (iv) of Def. 21. Assume by contradiction that there exists a set $F'_s \supseteq F_s \subseteq O_s$ such that there exists a $t \in S$ such that $t \not\sim s$, $F'_s \subseteq O_t$, and there is no v such that $v \prec r$ and $F'_s \subseteq O_v$. Consider now a sound and complete data-stream $\vec{O}$ such that for some $n \in N$, $\vec{O}[n] = \sigma_s$, and for some $k \geq n$, $set(\vec{O}[k]) = F'_s$. Since $\vec{O}[n]$ is a locking sequence for s, $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n]) = \{s\}$ and for all sequences τ for s, $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n] * \tau) = \{s\}$. Hence also $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[k]) = \{s\}$. By assumption, we know that there is a $t \in S$ such that $t \not\sim s$ and $set(\vec{O}[k]) \subseteq O_t$. Therefore, such a t is in the updated model $\mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[k])$, and moreover $t \not\sim^{\mathbf{cond}}_{\vec{O}[k]} s$. By assumption, for such a t, there is no v such that $v \prec r$ and $F'_s \subseteq O_v$. Hence, for all $v \in S$, either $v \not\sim t$ or $F'_s \not\subseteq O_v$. Consider such a v. If $F'_s \not\subseteq O_v$, then v has been eliminated. Therefore, for all $v \sim t$, either v has been eliminated or t is at least as good as v in the updated model, i.e., $t \preceq^{\mathbf{cond}}_{\vec{O}[k]} v$. Therefore, if $v \sim^{\mathbf{cond}}_{\vec{O}[k]} t$, then $t \preceq^{\mathbf{cond}}_{\vec{O}[k]} v$. Thus, in the model updated on the sequence $\vec{O}[k]$, we have $t \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[k]}}$, and thus $t \in L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[k])$. Contradiction.

[$\Leftarrow$] Let a preorder $\preceq$ and an epistemic space $\mathbf{S}$ be given such that for each s there exists a set F_s as in Def. 21. Let the plausibility assignment **PLAUS** be given that assigns the preorder $\preceq$ to the space $\mathbf{S}$. Consider now an arbitrary world s, and a sound and complete data stream $\vec{O}$ for s. Since F_s is finite, there exists an $n \in \mathbb{N}$ such that $F_s \subseteq set(\vec{O}[n])$. We now show that $L^{\mathbf{PLAUS}}_{\mathbf{cond}}$ identifies s in the limit, by showing

that s is the unique minimal element in $\mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$, and that it is the unique minimal element for any $k > n$. Consider $t \in S$ distinct from s in the model obtained after updating on $\vec{O}[n]$. By definition of **cond**, it must be the case that $set(\vec{O}[n]) \subseteq O_t$ and thus $F_s \subseteq O_t$. Consider the preorder $\preceq^{\mathbf{cond}}_{\vec{O}[n]}$ in the updated space $\mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. We consider two cases.

[Case 1: $t \sim^{\mathbf{cond}}_{\vec{O}[n]} s$] It follows that $t \sim s$. Since $F_s \subseteq set(O[n])$, and since $set(O[n]) \subseteq O_t$ (since otherwise t would have been eliminated), $F_s \subseteq O_t$. By Def. 21 it follows that $s \prec t$, and thus $s \prec^{\mathbf{cond}}_{\vec{O}[n]} t$. Since t is arbitrary, it follows that for all $t \sim^{\mathbf{cond}}_{\vec{O}[n]} s$, $s \prec^{\mathbf{cond}}_{\vec{O}[n]} t$, and thus $s \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$.

[Case 2: $t \not\sim^{\mathbf{cond}}_{\vec{O}[n]} s$] Consider a $t \not\sim^{\mathbf{cond}}_{\vec{O}[n]} s$. Since $t \not\sim^{\mathbf{cond}}_{\vec{O}[n]} s$, and since $set(\vec{O}[n]) \subseteq O_t$, we have that $F_s \subseteq set(\vec{O}[n])$. By Def. 21 there exists a $v \in S$ such that $v \prec t$ and $set(\vec{O}[n]) \subseteq O_v$. It follows that $v \prec^{\mathbf{cond}}_{\vec{O}[n]} t$, and thus $t \notin min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$. Since t was arbitrary, there is no $\in S$ such that $t \not\sim^{\mathbf{cond}}_{\vec{O}[n]} s$ and $t \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}}$.

It follows that, $min_{\preceq^{\mathbf{cond}}_{\vec{O}[n]}} \mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n]) = \{s\}$. Thus $L^{\mathbf{PLAUS}}_{\mathbf{cond}}(\mathbf{S}, \vec{O}[n]) = \{s\}$.

We now show that no new minimal element is created. Consider an arbitrary $k > n$, and assume by contradiction that some $t \neq s$ is a minimal element after updating on the sequence $\vec{O}[k]$. Note that $s \in min_{\preceq^{\mathbf{cond}}_{\vec{O}[k]}}$, since $\vec{O}$ is sound with respect to s, and **cond** does not change the relative order of the worlds. Also, $F_s \subseteq set(\vec{O}[n]) \subseteq set(\vec{O}[k])$. Since $set(\vec{O}[k]) \subseteq O_t$, we have $F_s \subseteq O_t$. Either $t \sim^{\mathbf{cond}}_{\vec{O}[k]} s$ or $t \not\sim^{\mathbf{cond}}_{\vec{O}[k]} s$. In the first case, $t \sim s$. Thus, by assumption, $s \prec t$, and by definition of conditioning $s \prec^{\mathbf{cond}}_{\vec{O}[k]} t$, which contradicts the assumption that t is minimal. Therefore, it must be the case that $t \not\sim^{\mathbf{cond}}_{\vec{O}[k]} s$. This implies that $t \not\sim s$. But since $F_s \subseteq set(\vec{O}[k])$, there exists a $v \in S$ s.t. $v \prec t$ and $set(\vec{O}[k]) \subseteq O_v$. It follows that, $v \prec^{\mathbf{cond}}_{\vec{O}[k]} t$ and therefore t is not a minimal element of $\mathbf{cond}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[k])$. Contradiction. □

Proof of Proposition 4

Proposition 4. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. The preorder $\preceq$ is appropriate to learn $\mathbf{S}$ via **lex** if and only if there exists a general conditioning tell-tale map for S and $\preceq$.*

To prove the proposition above we need a number of facts about **lex**.

Fact 1. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. If $s \sim t$, then $s \sim^{\mathbf{lex}}_{p} t$, for any $p \in O$.*

Proof. Assume that $s \sim t$. Then $s \preceq t$ or $t \preceq s$. Without loss of generality, assume that $s \preceq t$. There are four cases. [Case 1: $s \in p$ and $t \in p$] Since $s \in p$, $t \in p$ and $s \preceq t$ by assumption, we have that $s \preceq^{\mathbf{lex}}_{p} t$ by Def. 20. Hence $s \sim^{\mathbf{lex}}_{p} t$. [Case 2: $s \notin p$ and $t \notin p$] Since $s \notin p$, $t \notin p$ and $s \preceq t$ by assumption, we have that $s \preceq^{\mathbf{lex}}_{p} t$ by Def. 20. Hence $s \sim^{\mathbf{lex}}_{p} t$. [Case 3: $s \in p$ and $t \notin p$] If $s \in p$ and $t \notin p$, then $s \preceq^{\mathbf{lex}}_{p} t$ by Def. 20. Hence, $s \sim^{\mathbf{lex}}_{p} t$. [Case 4: $s \notin p$ and $t \in p$] If $s \notin p$ and $t \in p$, then $s \preceq^{\mathbf{lex}}_{p} t$ by Def. 20. Hence, $s \sim^{\mathbf{lex}}_{p} t$. □

Fact 2. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. Consider $s,t \in S$ such that $s \not\sim t$, and consider a sound data stream $\vec{O}$ for s. Then, for any $n \in \mathbb{N}$, if $set(\vec{O}[n]) \subseteq O_t$, then $s \not\sim^{\mathbf{lex}}_{\vec{O}[n]} t$.*

Proof. By induction on n. [Base Case: $n = 0$] Given that $\vec{O}[0]$ is empty, the updated space is the same as the initial space, thus $t \not\sim s$. [Inductive Step] Assume that for all $n < m$, $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$ in the corresponding update $\mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. Therefore, $t \not\sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$. Hence, $t \not\preceq^{\mathbf{lex}}_{\vec{O}[m-1]} s$ and $s \not\preceq^{\mathbf{lex}}_{\vec{O}[m-1]} t$. Assume $set(\vec{O}[m]) \subseteq O_t$, then $t, s \in p$, where $p \in O_s$ is the proposition occurring in $\vec{O}$ after the sequence $\vec{O}[m-1]$.

By Def. 20, since $s,t \in p$, we have that $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ if and only if $s \preceq^{\mathbf{lex}}_{\vec{O}[m-1]} t$, and $t \preceq^{\mathbf{lex}}_{\vec{O}[m]} s$ if and only if $t \preceq^{\mathbf{lex}}_{\vec{O}[m-1]} s$. Using the inductive hypothesis, we then conclude that $s \not\preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ and $t \not\preceq^{\mathbf{lex}}_{\vec{O}[m]} s$. Hence, $s \not\sim^{\mathbf{lex}}_{\vec{O}[m]} t$. □

Fact 3. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. Consider $s,t \in S$ such that $s \prec t$, and consider a sound data stream $\vec{O}$ for s. We have that for any $n \in \mathbb{N}$, $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$.*

Proof. Proof by induction on n. [Base Case: $n = 0$] Given that $\vec{O}[0]$ is empty, the updated space is the same as the initial space, thus $s \prec t$. [Inductive Step] Assume that for all $n < m$, $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$ in the corresponding update $\mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. Therefore, $s \prec^{\mathbf{lex}}_{\vec{O}[m-1]} t$. Since the stream is sound with respect to s, we have $s \in p$, where $p \in O$ occurs in the stream $\vec{O}$ after the sequence $\vec{O}[m-1]$. Either $t \in p$, or $t \notin p$. If $t \in p$, then $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ and $t \not\preceq^{\mathbf{lex}}_{\vec{O}[m]} s$ by Def. 20; hence, $s \prec^{\mathbf{lex}}_{\vec{O}[m]} t$. If $t \notin p$, then again $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ and $t \not\preceq^{\mathbf{lex}}_{\vec{O}[m]} s$ by Def. 20, and thus $s \prec^{\mathbf{lex}}_{\vec{O}[m]} t$. □

Fact 4. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. If there is an $s \in S$ such that $s \in min_{\preceq}$, then for any sound data stream $\vec{O}$ for s, for any $n \in \mathbb{N}$, $s \in min_{\preceq^{\mathbf{lex}}_{\vec{O}[n]}} \mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$.*

Proof. By induction on n. [Base case] Follows from the fact that $s \in min_{\preceq}$. [Inductive step] Assume that for all $n < m$, $s \in min_{\preceq^{\mathbf{lex}}_{\vec{O}[n]}}$, where $\preceq^{\mathbf{lex}}_{\vec{O}[n]}$ is the preorder in the space $\mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. This is also true for $n = m-1$. Therefore, $s \in min_{\preceq^{\mathbf{lex}}_{\vec{O}[m-1]}}$; hence, for all $t \sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$, $s \preceq^{\mathbf{lex}}_{\vec{O}[m-1]} t$. Now consider $t \sim^{\mathbf{lex}}_{\vec{O}[m]} s$. There are two possibilities: either $t \sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$ or not, i.e., $t \not\sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$. [Case 1: $t \sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$] Since $\vec{O}$ is sound with respect to s, we have $s \in p$, where $p \in O$ is the observable occurring in $\vec{O}$ after the initial segment $\vec{O}[m-1]$. Either $t \notin p$ or $t \in p$. If $t \notin p$, then $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ by Def. 20; if $t \in p$, then since $s \preceq^{\mathbf{lex}}_{\vec{O}[m-1]} t$ by inductive hypothesis, we have $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ by Def. 20. [Case 2: $t \not\sim^{\mathbf{lex}}_{\vec{O}[m-1]} s$] Since $\vec{O}$ is sound with respect to s, we have $s \in p$, where $p \in O$ is the observable occurring in $\vec{O}$ after the initial segment $\vec{O}[m-1]$. Either $t \notin p$ or $t \in p$. If $t \notin p$, then $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$ by Def. 20; if $t \in p$, then $s \not\sim^{\mathbf{lex}}_{\vec{O}[m]} t$ by Def. 20. Hence, for all $t \in S$, if $t \sim^{\mathbf{lex}}_{\vec{O}[m]} s$, then $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$. Hence $s \in min_{\preceq^{\mathbf{lex}}_{\vec{O}[m]}}$. □

We can now show that the existence of a general conditioning tell-tale map (Definition 21) is necessary (Lemma 7, right to left direction of Proposition 4) and sufficient (Lemma 8, left to right direction of Proposition 4) for a preorder to be appropriate to learn via **lex**.

Lemma 7. *Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$. If $\preceq$ is appropriate to learn $\mathbf{S}$ via **lex**, then there exists a conditioning tell-tale map for S and $\preceq$.*

Proof. Let $\preceq$ be a preorder on an epistemic space $\mathbf{S} = \langle S, O\rangle$ that is identifiable via lexicographic upgrade. Let the plausibility assignment **PLAUS** be given that assigns $\preceq$ to $\mathbf{S}$. Assume that $L^{\mathbf{PLAUS}}_{\mathbf{lex}}$ identifies $\mathbf{S}$ in the limit. By Lemma 6, for every state $s \in S$, there exists a locking sequence σ_s. Consider an arbitrary state s. Let $F_s := set(\sigma_s)$. We now show that F_s is a conditioning tell-tale for s. F_s is finite, since σ_s is finite, and thus satisfies point (i) of Def. 21. $F_s \subseteq O_s$, since σ_s is sound with respect to s. Thus, F_s satisfies point (ii) of Def. 21. We now show that F_s satisfies (iii) in Def. 21. Assume by contradiction that there is some $t \sim s$ such that $F_s \subseteq O_t$, $s \neq t$ and $s \not\prec t$. Since $s \not\prec t$, either $t \prec s$ or $t \simeq s$. [Case 1: $t \prec s$] Consider a sound and complete data stream $\vec{O}$ for which there is an $n \in \mathbb{N}$ such that $\vec{O}[n] = \sigma_s$.

Since $t \prec s$, $F_s \subseteq O_t$ and $\vec{O}[n]$ is sound with respect to s, we have that $t \prec^{\mathbf{lex}}_{\vec{O}[n]} s$ by Fact 3. Hence, $s \notin min_{\preceq^{\mathbf{lex}}_{\vec{O}[n]}} \mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$, and thus $s \notin L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[n])$, which contradicts the fact that σ_s is a locking sequence for s. [Case 2: $t \simeq s$] Similar as the case above: s is a minimal element if and only if t is a minimal element. And hence we contradict the assumption that σ_s is a locking sequence.

We now show that F_s satisfies (iv) of Def. 21. Assume in order to derive a contradiction that there exists a finite set $F'_s \supseteq F_s \subseteq O_s$ such that some $t \in S$ with $t \not\sim s$, and $F'_s \subseteq O_t$ has no v such that $v \prec t$ and $F'_s \subseteq O_v$. Consider a data stream $\vec{O}$ such that there is an $n \in \mathbb{N}$ such that $\vec{O}[n] = \sigma_s$, and for which there is a $k \geq n$ such that $set(\vec{O}[k]) = F'_s$. Since $\vec{O}[n]$ is a locking sequence for s, $L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k]) = \{s\}$. Consider now $t \in S$ as above: for all $v \in S$, either $v \not\prec t$ or $F'_s \not\subseteq O_v$. We now show that t is a minimal element of $\preceq^{\mathbf{lex}}_{\vec{O}[k]}$ in the updated plausibility space $\mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[k])$. Consider $v \sim^{\mathbf{lex}}_{\vec{O}[k]} t$: either $v \preceq^{\mathbf{lex}}_{\vec{O}[k]} t$ or $t \preceq^{\mathbf{lex}}_{\vec{O}[k]} v$. If $v \preceq^{\mathbf{lex}}_{\vec{O}[k]} t$, it must be the case that $F'_s \subseteq O_v$, otherwise v would have become strictly less plausible than t at some point and would have stayed so, since $F'_s \subseteq O_t$. But by assumption, it follows that $v \not\prec t$, and hence, either $v \not\sim t$ or, $t \preceq v$. If $v \not\sim t$, then $v \not\sim^{\mathbf{lex}}_{\vec{O}[k]} t$, since both $F'_s \subseteq O_t$ and $F'_s \subseteq O_v$. But $v \sim^{\mathbf{lex}}_{\vec{O}[k]} t$, and thus contradiction. Therefore, it must be the case that $t \preceq v$. Since both t, v satisfy all the propositions occurring in $\vec{O}[k]$, we have that $t \preceq^{\mathbf{lex}}_{\vec{O}[k]} v$. In all cases, for all $v \sim^{\mathbf{lex}}_{\vec{O}[k]} t$, we have $t \preceq^{\mathbf{lex}}_{\vec{O}[k]} v$, which implies that $t \in min_{\preceq^{\mathbf{lex}}_{\vec{O}[k]}}$, and thus $t \in L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k])$. But since σ_s is a locking sequence and $s \in \bigcap set(\vec{O}[k])$, we may conclude $L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k]) = \{s\}$. Contradiction. □

Lemma 8. *Let $\preceq$ be a preorder on an epistemic space $\boldsymbol{S} = \langle S, O \rangle$. If there exists a conditioning tell-tale map for S and $\preceq$, then $\preceq$ is appropriate to learn $\boldsymbol{S}$ via **lex**.*

Proof. Let a space $\mathbf{S}$ be given and let $\mathbf{PLAUS}$ be a plausibility assignment assigning a preorder $\preceq$ to $\mathbf{S}$ such that there exists a conditioning tell-tale map for S and $\preceq$.

Consider a state $s \in S$ and a sound and complete data stream $\vec{O}$ for s. Consider a set F_s for s as in Def. 21. Since the data stream is complete and F_s is finite, there exists an $n \in \mathbb{N}$ such that $F_s \subseteq set(\vec{O}[n])$.

We now show that $L^{\mathbf{PLAUS}}_{\mathbf{lex}}$ outputs s after updating on $\vec{O}[n]$, and forever after that. Consider a $t \neq s$ in $\mathbf{lex}(\mathbf{PLAUS}(\mathbf{S}), \vec{O}[n])$. Either $t \sim^{\mathbf{lex}}_{\vec{O}[n]} s$ or $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$.

[Case 1: $t \sim^{\mathbf{lex}}_{\vec{O}[n]} s$] Either $t \not\sim s$ or $t \sim s$. If $t \not\sim s$, then by Fact 2, we have $set(\vec{O}[n]) \not\subseteq O_t$, and thus there exists an $m \leq n$ such that $s \preceq^{\mathbf{lex}}_{\vec{O}[m]} t$, and $t \not\preceq^{\mathbf{lex}}_{\vec{O}[m]} s$ by Def. 20. Since $\vec{O}$ is sound with respect to s, we have that $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$ by Fact 3. If $t \sim s$, then there are two possibilities: $F_s \subseteq O_t$ or $F_s \not\subseteq O_t$. If $F_s \subseteq O_t$, then $s \prec t$ by assumption, and by Fact 3, $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$. If $F_s \not\subseteq O_t$, then there exists a $p \in set(\vec{O}[n])$ such that $t \notin set(\vec{O}[n])$. Therefore, $s \prec^{\mathbf{lex}}_{\vec{O}[m]} t$ for some $m \leq n$. By Fact 3, $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$.

[Case 2: $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$] Since $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$, we have that $t \not\sim s$ by Fact 1. It must be the case that $set(\vec{O}[n]) \subseteq O_t$, otherwise, $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$. By assumption, we therefore know that there exists a $v \prec t$ such that $set(\vec{O}[n]) \subseteq O_v$. Since both $set(\vec{O}[n]) \subseteq O_v$ and $set(\vec{O}[n]) \subseteq O_t$, we have that $v \prec^{\mathbf{lex}}_{\vec{O}[n]} t$. Therefore, $t \notin min_{\preceq^{\mathbf{lex}}_{\vec{O}[n]}}$.

We therefore know that for all $t \sim s$, $s \preceq^{\mathbf{lex}}_{\vec{O}[n]} t$, and there is no $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$ such that for all $v \sim^{\mathbf{lex}}_{\vec{O}[n]} t$, $t \preceq^{\mathbf{lex}}_{\vec{O}[n]} v$. Therefore $min_{\preceq^{\mathbf{lex}}_{\vec{O}[n]}} = \{s\}$, and thus $L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[n]) = \{s\}$.

We now need to show that for all $k \geq n$, $L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k]) = \{s\}$. Assume that there is some k for which $L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k]) \neq \{s\}$. By Fact 4, we know that $s \in L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k])$. Therefore, there must exist

a t such that $t \neq s$ and $t \in L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k])$.

[Case 1: $t \sim^{\mathbf{lex}}_{\vec{O}[n]} s$] Since s is the uniquely minimal element after the sequence $\vec{O}[n]$, it must be the case that $s \prec^{\mathbf{lex}}_{\vec{O}[n]} t$. Since $\vec{O}$ is sound, by Fact 3, $s \prec^{\mathbf{lex}}_{\vec{O}[k]} t$. Contradiction. [Case 2: $t \not\sim^{\mathbf{lex}}_{\vec{O}[n]} s$] Then $t \not\sim s$, and $F_s \subseteq set(\vec{O}[n]) \subseteq set(\vec{O}[k]) \subseteq O_t$, otherwise, t would have become less plausible than s. But then, by assumption, there exists a v such that $set(\vec{O}[k]) \subseteq O_v$ and $v \prec t$. Since $set(\vec{O}[k]) \subseteq O_v$, and $v \prec t$, we have $v \prec^{\mathbf{lex}}_{\vec{O}[k]} t$ (since they always satisfy the same propositions, their relative plausibility never changes). Thus $t \notin min_{\prec^{\mathbf{lex}}_{\vec{O}[k]}}$, hence $t \notin L^{\mathbf{PLAUS}}_{\mathbf{lex}}(\mathbf{S}, \vec{O}[k])$. Contradiction. □

Are Large Random Graphs Always Safe to Hide?

Sourav Chakraborty
Indian Statistical Institute
sourav@isical.ac.in

Sujata Ghosh
Indian Statistical Institute
sujata@isichennai.res.in

Smiha Samanta
Indian Statistical Institute
smi1995ha@gmail.com

We discuss winning possibilities of players in various variants of cops and robber game played on large random graphs, a testbed for various kinds of network queries, search problems in particular. We explore the use of logic frameworks to investigate such results; in particular, we show that whenever a winning condition for either player can be expressed as a certain kind of formula in first-order logic, that player almost always wins. In the process, we obtain more insight into the logic-game connection from the zero-one law perspective.

1 Introduction

We begin with describing the cops and robber game [6] in detail, which is the focus of the current study. In the cops and robber game, we have a finite (≥ 1) number, say k, of cops looking for a robber on a game arena, provided by a graph structure. The players occupy vertices of the graph, and they move along the edges to adjacent vertices. We assume that the underlying graph is reflexive and, accordingly, staying at her vertex can also be considered as a move for the player. This is a turn-based game with the cops and the robber occupying certain vertices and moving alternately, and in their turn, the cops move simultaneously, where some may stay put on their vertices. The game is played in a countable sequence of rounds. The cops win if after some finite number of rounds, at least one of the cops can occupy the same vertex as the robber. The robber wins if he can avoid the cops in an infinite run. For this work, as is the traditional case, players are assumed to have perfect information in these games.

The cops and robber games that are played on graphs are proposed in [17, 16, 1], and since then they have been extensively studied in various forms from distinct viewpoints. From the algorithmic and combinatorial perspectives, the central point of the study is the *cop number* of the game, that is, the minimum number of cops required to win a game. For a detailed summary of these studies, see [6]. In the study of robotic systems, search missions in a pursuit-evasion environment, in particular adversarial search problems, have been a primary object for investigations [7, 12]. In recent years, logicians have studied the hide and seek game, which shares a similar pursuit-evasion structure, with a focus on modeling the dynamic interaction between players [3, 13].

These games, played on graphs, occur in many natural contexts where the underlying graphs are generated by some stochastic processes; for example, see [20]. In [15], sabotage games (introduced in [2]), played between a traveler and a demon, are investigated to check whether they are biased towards the traveler (who travels along the edges of the graph to reach her goal vertex) or the saboteur (who removes edges from the graph to prevent the traveler from reaching her goal), and it is found that the game is indeed biased towards the traveler in large random graphs. In this work, we investigate the cops and robber game to check whether the game conditions favor the cops or the robber when the size of the underlying graph structure (number of vertices) becomes larger and larger. Since we do not have any control over the occurrence of edges in the graphs, we consider random graphs [9, 11] with constant and varying edge probabilities. Thus, we deal with cops and robber games played on random graphs.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 321–333, doi:10.4204/EPTCS.437.26

© Sourav Chakraborty, Sujata Ghosh & Smiha Samanta
This work is licensed under the
Creative Commons Attribution License.

We should note here that from the algorithmic perspectives (checking the cop number of the game), these games have already been studied with respect to random graphs (e.g., see [4]). Our goal here is to highlight how the notion of winning in these games is skewed towards one of the players when we consider these graphs as the corresponding game arenas. These results can be used for further developments with respect to the design and expansion of relevant networks. The general interest in studies featuring large networks has increased manifold, and the present study may help us deal with pursuit-evasion problems in randomly generated networks. Our intuition tells us that the larger the graph size is, the easier it should be for the robber to hide inside. We will see below whether the formal results agree with our intuitions. In addition to the traditional cops and robber game, we also study various variants of the game played on random graphs to explore how the bias that we discussed earlier, varies in terms of these variants.

The proofs that we have here are simple applications of various known results relating random graphs and first-order logic. Our main goal is to showcase the elegant connection between first-order logic, probability theory and graph games. We believe that this connection is worth-pursuing over various kinds of random models and intend to provide a glimpse of this rich field of study. Our proof techniques rely on the same for showing zero-one laws in first-order logic over graphs [10], and we first give a brief primer on those results.

2 On zero-one laws in first-order logic

The main idea behind the zero-one law, satisfied by some logic, is the feature that various properties that are expressible in the logic under consideration can be either *almost surely true* or *almost surely false*. In the following, we show that the first-order theory of graphs, whose vocabulary consists only of a binary relation symbol, satisfies the zero-one law. Let us first rigorously define what we mean by 'almost surely', following [14].

Let GR_n denote the set of all graphs with n vertices given by $\{0,1,\ldots,n-1\}$. Then, the number of possible graphs on these n vertices is given by: $|GR_n| = 2^{\binom{n}{2}}$.

Definition 1. *Let φ denote some first-order formula expressing some property of the graphs, say* P. *We define:*

$$\mu_n(\varphi) = \frac{|\{G \in GR_n : G \text{ satisfies } \varphi\}|}{|GR_n|}.$$

Thus $\mu_n(\varphi)$ is the probability that a randomly chosen graph on the set of nodes $\{0,1,\ldots,n-1\}$ satisfies the property expressed by φ. We now deine the asymptotic probability of φ as:

$$\mu(\varphi) := \lim_{n\to\infty} \mu_n(\varphi).$$

In the same way, without going into the underlying logic per se, we can define $\mu_n(\mathsf{P})$ and $\mu(\mathsf{P})$. Let us now establish the following well-known result from first-order logic. We show this for the first-order language of graphs, $\mathscr{L}_G$. To begin with, we have the following definition of *almost surely* true or false formulas.

Definition 2. *A formula $\varphi \in \mathscr{L}_G$ is said to be* almost surely true *if $\mu(\varphi) = 1$, and is said to be* almost surely false *if $\mu(\varphi) = 0$.*

Accordingly, we have the zero-one law for the first-order logic of graphs. This is a well-established result; however, for a better understanding of the later results, we provide a proof idea below.

Theorem 1 ([10])**.** *Any sentence in first-order logic with only a binary relation symbol in the vocabulary is either almost surely true or almost surely false.*

Proof. The main steps of the proof involve the following: (i) defining a set of axioms termed as *extension axioms* regarding how graphs with n vertices can be extended to graphs with $n+1$ vertices, and showing them to be surely true; (ii) constructing a countable model, RG, say, of these axioms, unique upto isomorphism, and showing that for any sentence φ in the given language, either $\mathsf{RG} \vDash \varphi$ or, $\mathsf{RG} \vDash \neg\varphi$; and finally, (iii) showing that for any sentence φ with $\mathsf{RG} \vDash \varphi$, φ is almost surely true. Thus we have that any sentence φ is either almost surely true or almost surely false.

Extension axioms: The extension axioms simply state the fact that for any subgraph of size n, one can get a subgraph of size $n+1$ and one can make the new vertex adjacent or non-adjacent to the pre-existing vertices in whatever way one wants. In case the new vertex is adjacent to m pre-existing vertices, it can be expressed as follows:

$$EA_{m,n} : \forall x_1, x_2 \ldots x_n \Big(\bigwedge_{1 \le i < j \le n} \neg(x_i = x_j) \rightarrow \exists z \Big(\bigwedge_{1 \le i \le n} \neg(z = x_i) \wedge \bigwedge_{1 \le i \le m} E(z, x_i) \wedge \bigwedge_{m < i \le n} \neg E(z, x_i) \Big) \Big)$$

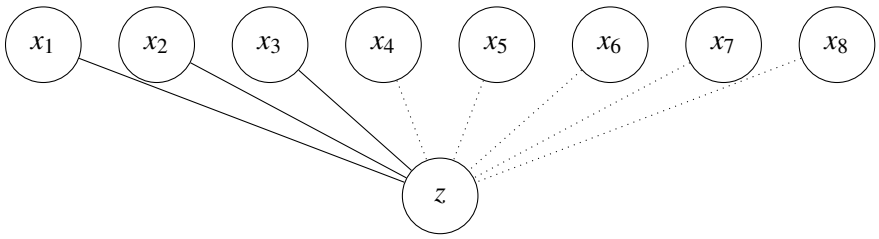


Figure 1: $EA_{3,8}$: for all possible choice of vertices $x_1, x_2, \ldots x_8$ we can always find a vertex z that is adjacent to only 3 of those 8 vertices.

Extension axioms are almost surely true: We consider $\mu_k(EA_{m,n})$, for $k > n$, and concentrate on $\mu_k(\neg EA_{m,n})$. Our goal is to show that $\lim_{k\to\infty} \mu_k(\neg EA_{m,n}) = 0$, which would then imply $\lim_{k\to\infty} \mu_k(EA_{m,n}) = 1$. For $EA_{m,n}$ to be false in a k-size graph structure, there must be an m-size set of vertices, say A and an $n-m$-size set of vertices, say B, such that no vertex in the complement of $A \cup B$ is adjacent to the vertices in A and not adjacent to the vertices in B. For vertices in $A \cup B$, the possible choices together with the choices for their adjacency are given by $\binom{k}{m} \times \binom{k-m}{n-m} \times 2^{\binom{n}{2}}$. For the vertices outside $A \cup B$, the choice of adjacency is given by $2^{\binom{k-n}{2}}$. Finally, for any vertex outside A∪B, it can be adjacent to one in A∪B in any way except for the desired one, and hence adjacency can be chosen in $2^n - 1$ ways. Thus we have:

$$\mu_k(\neg EA_{m,n}) \le \frac{\binom{k}{m} \times \binom{k-m}{n-m} \times 2^{\binom{n}{2}} \times 2^{\binom{k-n}{2}} \times (2^n - 1)^{k-n}}{2^{\binom{k}{2}}} \le \frac{k^n \times (2^n - 1)^{k-n}}{2^{n \times (k-n)}} = k^n \times (1 - \frac{1}{2^n})^{k-n}$$

Thus, $\lim_{k\to\infty} \mu_k(\neg EA_{m,n}) = 0$, and hence $\lim_{k\to\infty} \mu_k(EA_{m,n}) = 1$, that is, $\mu(EA_{m,n}) = 1$ that is, $EA_{m,n}$ is almost surely true for all m,n $\in \mathbb{N}$.

Extension axioms have a unique countable model, say RG**:** Let us consider RG to be a countable random graph, which has the set of natural numbers as its vertices and for any pair (i, j) of vertices with

$i < j$, we assume them to be adjacent with probability 1/2. We claim that $EA_{m,n}$ holds in RG. If $EA_{m,n}$ does not hold at RG, then there are n vertices such that any other vertex can be related to them in all but one way (given by $EA_{m,n}$). The probability that one vertex cannot be related in the desired way is $(1-\frac{1}{2^n})$. Thus, the probability that no vertex is related in the desired way is given by $\lim_{k\to\infty}(1-\frac{1}{2^n})^k = 0$. Hence, $EA_{m,n}$ holds in RG, and this completes the proof of our claim. That such a countable random graph is unique up to isomorphism can be shown by an inductive argument.

For any sentence φ either $\mathsf{RG} \models \varphi$ or, $\mathsf{RG} \models \neg\varphi$: Let us suppose that there is a sentence φ such that $\mathsf{RG} \not\models \varphi$ and $\mathsf{RG} \not\models \neg\varphi$. Then we can consider a model for all the extension axioms with φ, say G', and another one for all the extension axioms with $\neg\varphi$, say G''. But then, $\mathsf{G}' \cong \mathsf{G}'' \cong \mathsf{RG}$, as both G' and G'' satisfy all the extension axioms. This implies that $\mathsf{RG} \models \varphi$ and $\mathsf{RG} \models \neg\varphi$, a contradiction.

For any sentence φ with $\mathsf{RG} \models \varphi$, φ is almost surely true : As $\mathsf{RG} \models \varphi$, by compactness, there is a finite subset of extension axioms, *Fin*, say, such that $Fin \models \varphi$. Now, observe that $EA_{m',n'} \models EA_{m,n}$ if $\min\{m', n'-m'\} \geq \max\{m, n-m\}$. Thus, if we take $k = \max_{EA_{m,n}\in Fin}\{\max\{m,n-m\}\}$, for any $EA_{m,n} \in Fin$, $EA_{k,2k} \models EA_{m,n}$ as $k \geq \max\{m,n-m\}$. Choosing k accordingly, and using the fact that $Fin \models \varphi$, we have that $EA_{k,2k} \models \varphi$. Then, $\mu(\varphi) \geq \mu(EA_{m,n}) = 1$, that is, $\mu(\varphi) = 1$. Thus, for any sentence φ such that $\mathsf{RG} \models \varphi$, φ is almost surely true.

□

3 Hiding on random graphs with constant edge probability

Let us now use the proof discussed above to shed light on winning or losing in cops and robber games played on random graphs, when the number of vertices of the graph tends to infinity. As mentioned in the introduction, it would be natural to assume that *the bigger the graph is, easier it is for the robber to hide in there*. In what follows, we provide a formal seal to this intuition that we have and also explore cases where the results may not agree with such instinctive way of thinking.

We assume here that for any edge of the graph, the probability that the edge appears in the graph is p, where $0 < p < 1$. We analyze what happens for the graph $G(N,p)$, where N is the number of vertices and p is the constant probability of an edge appearing that does not depend on N. We showed above that any FOL definable graph property is true in almost all graphs if it is true in the random graph $G(N,p)$ with $N \to \infty$ and p constant. We considered $p = 1/2$ in the result proved in the previous section, but the same result is applied for any p with $0 < p < 1$. Let us first look at the game between a single cop and a robber played on random graphs $G(N,p)$.

Theorem 2. *In a random graph $G(N,p)$ where p is the constant probability of an edge appearing in the graph and $N \to \infty$, the robber almost always has a winning strategy.*

Proof. Let us first define a neighborhood $\mathsf{N}(u)$ of a vertex u in a graph by: $\mathsf{N}(u) := \{v : E(u,v)\}$, with E denoting the binary edge relation on a graph. For a single cop and a robber, if x is the position of the cop and y is the position of the robber, then the cop can win in the next move iff $\mathsf{N}(y) \subseteq \mathsf{N}(x)$. Similarly, the robber can prevent a win for the cop iff $\exists z$ such that $z \in \mathsf{N}(y)$ and $z \notin \mathsf{N}(x)$. Let us consider the extension axioms now.

$$EA_{m,n} : \forall x_1, x_2 \ldots x_n \Big(\bigwedge_{1\leq i<j\leq n} \neg(x_i = x_j) \to \exists z \Big(\bigwedge_{1\leq i\leq n} \neg(z = x_i) \wedge \bigwedge_{1\leq i\leq m} E(z,x_i) \wedge \bigwedge_{m<i\leq n} \neg E(z,x_i) \Big) \Big)$$

It can be shown as earlier that the extension axioms hold in the random graph $G(N,p)$, as $N \to \infty$. If we take any n vertices, the probability that a vertex is adjacent or not adjacent to these vertices in the

desired way is given by $p^m \times (1-p)^{n-m}$. Thus, the probability of the corresponding extension axiom not being true is given by $\lim_{N\to\infty}(1-(p^m \times (1-p)^{n-m}))^N = 0$. Now, we claim that for any positions of the cop and the robber, the robber can avoid the cop in the next move, that is, for any two vertices x and y, there is a vertex z, adjacent to y but not adjacent to x. This statement is equivalent to the following sentence in first-order language:

$$\forall x \forall y \exists z(\neg(x=y) \rightarrow E(y,z) \wedge \neg E(x,z))$$

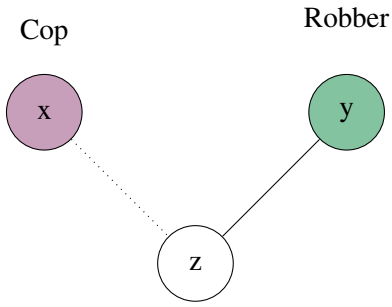


Figure 2: The robber at y has a move to a vertex z avoiding the cop at x.

We have that

$$\forall x \forall y \exists z(\neg(x=y) \rightarrow \neg(z=x) \wedge \neg(z=y) \wedge E(y,z) \wedge \neg E(x,z))$$

is an extension axiom, and it holds in the random graph. This implies that at any point, there is always a vertex z, such that the robber has a move from its current position y to the position z, in which case he can avoid the cop whose current position is x, as x is not adjacent to z. Moreover, as extension axioms are almost surely true, we can say that robber has a winning strategy in almost all possible graphs. □

We now show that it does not help if you have a million cops on the lookout for a single robber when the game arena gets large enough, that is, even in the case for $k > 1$ cops and a single robber, the robber would still be able to avoid the cops.

Theorem 3. *In a random graph $G(N,p)$ where p is the constant probability of an edge appearing in the graph and $N \to \infty$, the robber almost always has a winning strategy, playing against $k > 1$ cops.*

Proof. We can show that the robber has a winning strategy by considering the following extension axiom:

$$\forall x_1 \forall x_2 ... \forall x_k \forall y \exists z((\bigwedge_{1\le i<j\le k} \neg(x_i = x_j) \wedge \bigwedge_{i=1}^{k} \neg(x_i = y)) \rightarrow \bigwedge_{i=1}^{k} \neg(x_i = z) \wedge \neg(y=z) \wedge \bigwedge_{i=1}^{k} \neg E(x_i,z) \wedge E(y,z))$$

Arguing in a similar manner, we can say that the robber has a winning strategy on a random graph, as at any point there is a vertex z, such that the robber with its current position y, can move to z, and the cops whose current positions are $x_1, \ldots, x_k$ cannot reach z, as z is not adjacent to any one of them. □

Combining the results above, we can make the following statement regarding cops and robber game played on random graphs.

Corollary 1. *The robber almost always has a winning strategy in cops and robber game played on random graphs with $k \geq 1$ cops and a single robber.*

We have dealt with the general cops and robber games played on random graphs where edges occur with a constant probability p. Now we move on to certain variants of the game played on large random game graphs to investigate the possibility of cops catching the robber in these variants.

3.1 Cops and robber games with traps and roadblocks

Among the variants, we start with a lesser known one. The main idea is that of *traps* [8, 6] that can be placed on any vertex by a cop. If the robber ever occupies the same vertex, then he is caught. The cops can set those traps in some vertices when they arrive and can also remove when one of them arrives again at that vertex and reuse that trap. The robber gets caught if he meets either any of the cops or reaches any of the traps. In some sense, cops may have more power when we consider m cops with n traps and a single robber. However, that is not always the case: if we consider the complete bipartite graph $K_{3,3}$, 2 cops can win the corresponding game, whereas 1 cop having the ability to put 1 trap cannot win the game. Evidently, even with all these traps, the cops cannot win as we see below.

Theorem 4. *In a random graph $G(N,p)$ where p is the constant probability of an edge appearing in the graph and $N \to \infty$, the robber almost always has a winning strategy, playing against m cops equipped to put n traps.*

Proof. Let $x_1, x_2, ..., x_m$ be the positions of the cops and $t_1, t_2, ..., t_k$ be the positions of the traps at some points where $k \leq n$, and y be the position of the robber. Consider the following extension axiom:

$$\forall x_1 ... \forall x_m \forall t_1 ... \forall t_k \forall y \exists z \Big(\Big(\bigwedge_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}} (\neg(x_i = t_j) \wedge \neg(y = x_i) \wedge \neg(y = t_j)) \wedge \bigwedge_{\substack{i \neq j \\ i,j=1}}^{m} \neg(x_i = x_j) \wedge \bigwedge_{\substack{i \neq j \\ i,j=1}}^{k} \neg(t_i = t_j) \Big) \rightarrow$$

$$\Big(\bigwedge_{i=1}^{m} \neg(z = x_i) \wedge \bigwedge_{i=1}^{k} \neg(z = t_i) \wedge \neg(y = z) \wedge \bigwedge_{i=1}^{m} \neg E(z, x_i) \wedge E(y, z) \Big) \Big)$$

If there is any such z, then robber can move to z, as they are adjacent. And he can avoid all cops as no cop position is adjacent to z, and he can also avoid traps as there is no trap position is equal to z. Since extension axioms hold in random graph, we have that the robber can avoid crops and traps at any point in that graph, and thus the robber has a winning strategy. □

We can also consider a similar variant where a finite number of *road-blocks* can be placed on the edges by cops [6]. The cops can move over those blocked edges but the robber cannot. These road-blocks can only be put on edges when a cop is at one of the endpoints of the edge. As in the case of traps, the cops get empowered once again, but still cannot win. The results follow in the same way as in the case of the games with traps, as placing a road-block has the same effect as placing a trap on one of the end points between which the road-block is put.

3.2 Cops and Robbers with different edge set

In the next variant of the game that we consider, the cops and the robber get different edge sets to move – the cops get the complement of the robber's edge set. Without loss of generality we will consider a

single cop here. We consider the given graph to be the edge set of the robber, and its complementary set provides the cop's edges to move.

Theorem 5. *In a random graph $G(N,p)$ where p is the constant probability of an edge appearing in the graph and $N \to \infty$, where cops and robber move along complementary edge sets, the robber almost always has a winning strategy.*

Proof. The proof is similar to the one cop and one robber game, the only subtle difference is in the extension axiom:

$$\forall x \forall y \exists z (\neg(x = y) \to E(y,z) \wedge E(x,z))$$

The statement says that for any cop position x and any robber position y, there exists a move from y to z, where x cannot reach, as the cop and the robber move in complementary edge sets, and E corresponds to robber's moves in particular. As earlier, we can say that robber has a winning strategy in almost all such games. Note the difference in the axiom from the one cop - one robber case. □

3.3 Cops and robber game with tandem-cops

The final variant that we will consider is a game played by tandem-cops. In this case, two cops move together *in tandem*, that is, they can either be at the same vertex or at vertices adjacent to each other, and this condition continues to hold as they move along the edges of the graph. Such a pair of cops constitute the notion of tandem-cops. Thus, when one of the tandem-cops moves to a vertex, the other one may to move to the same vertex or to one that is adjacent to the former. If we consider two cops, say C_1 and C_2, without loss of generality, we can have C_1 moving first to a vertex x'_1, say, which is adjacent to his previous position. Then the other cop C_2 can move to x'_1 or any other vertex x'_2, say, that is adjacent to x'_1.

A game is *tandem-win* if one pair of tandem-cops suffices to capture the robber. We note that the class of games where the cops win is a proper subset of the class of games where the tandem-cops win. If we consider the cycle with 4 vertices, C_4, one can easily check that it is a tandem-win, whereas, in the case of one cop, the robber has a winning strategy by taking his position at the vertex diagonally opposite to that of the cop. Consider the *Petersen Graph* given below, where at least three cops are needed to beat the robber. Note that the graph has diameter two; that is, any two vertices are at a distance of at most two units. As tandem-cops can cover two units of distance in a single move, the game on this graph is tandem-win. In the following, we check whether tandem-cops provide any advantage over the robber in random graphs. Finally, they do!

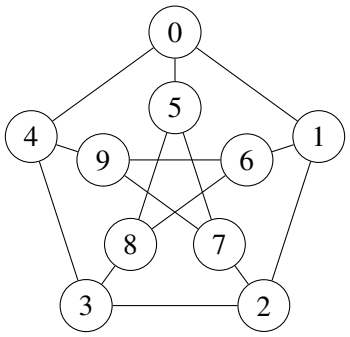


Figure 3: Petersen Graph

Theorem 6. *In a random graph $G(N,p)$ where p is the constant probability of an edge appearing in the graph and $N \to \infty$, cops almost always have a winning strategy whenever there is a pair of tandem-cops playing.*

Proof. Suppose x_1 and x_2 are the positions of the tandem-cops. Then there is an edge between x_1 and x_2. Let y be the position of the robber. We consider the following extension axiom:

$$\forall x_1 \forall x_2 \forall y \exists z((\neg(x_1 = x_2) \wedge \bigwedge_{i=1,2} \neg(x_i = y)) \to (\bigwedge_{i=1,2} \neg(x_i = z) \wedge \neg(y = z) \wedge E(x_1,z) \wedge E(y,z)))$$

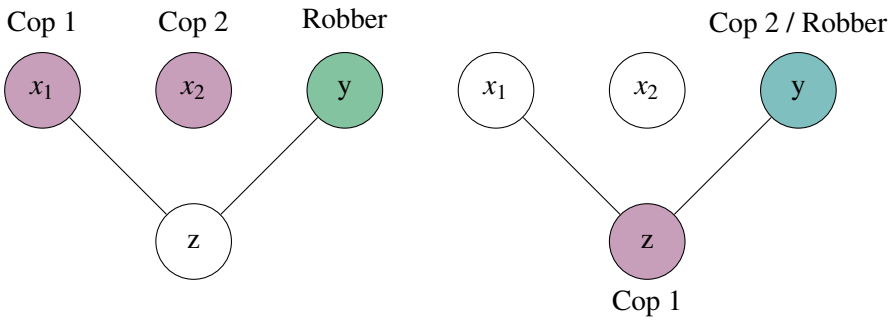


Figure 4: The tandem-cops can catch the robber in a single move.

If there is a vertex z adjacent to both x_1 and y, we can move one of the tandem-cops, at x_1 say, to z and as y is adjacent to z, we can move the other tandem-cop to y and catch the robber. And as we know, extension axioms are true in random graph $G(N,p)$ with $N \to \infty$ and p constant, we can conclude that the cops have a winning strategy in the corresponding game. Moreover, as extension axioms are true in almost all possible graphs, cops have a winning strategy in almost all cops and robber games in presence of tandem-cops. □

A natural question to ask: what is so special about tandem-cops? In whatever ways these cops move, they stay in two adjacent vertices. When one of the cops at x_1, say, move to a neighboring vertex, say x_1', the other cop can move to any neighbor of x_1', say x_2'. This way it is guaranteed that the cops can cover twice the distance, as the distance between x_1 and x_2' is 2 units. In a random graph $G(N,p)$ where $0 < p < 1$ is the constant edge probability, the distance between any two vertices is at most 2 almost surely – once again, an application of the extension axioms, which gives the cops an advantage. We should mention here that in a connected graph, a cop who can move two steps at a time always has a winning strategy against the robber, as he can continue reducing the distance between them. Using the fact that a large random graph is *almost surely* connected, we can get the same result: a large random graph is *almost surely* a tandem-win graph.

To finish this discussion, we note that we can similarly deal with other variants of the cops and robber game present in the literature, and show the ubiquitous applications of the extension axioms with respect to various studies on random graphs. Let us now move on to the case of random graphs with varying edge probability, which has its own nuances.

4 Hiding on random graphs with varying edge probability

In the previous section, we have considered random graphs $G(N,p)$, where $0 < p < 1$ is constant. Now, we will focus on the graphs where the edge-assigning probability depends on the graph size N, given by $p(N)$, say, a function of N. We assume here that for any edge of a graph of size N, the probability that the edge appears in the graph is $p(N)$, and we analyze what happens for the cops and robber game played over the random graph $G(N,p(N))$, as $N \to \infty$. We first look at the game between a single cop and a robber played on random graphs $G(N,p(N))$, and explore the values for $p(N)$ for which the cop or the robber has a winning strategy as $N \to \infty$. We have already seen that for the constant function $p(n) = p$ where $0 < p < 1$, the probability that the robber has a winning strategy tends to 1 as $N \to \infty$.

In what follows, we consider bounds over the edge-probability function $p(N)$ explore their effects on the winning strategies of players in the game. To this end we define the threshold functions [5] for properties in random graphs.

Definition 3 (Threshold function)**.** *Let $G(N,p(N))$ be a random graph with varying edge probability and P be a graph property. We call a function $f(N)$ threshold function for P if:*

1. *whenever $p(N) \ll f(N)$, $\mu_G(\mathsf{P}) = 0$.*
2. *whenever $f(N) \ll p(N)$, $\mu_G(\mathsf{P}) = 1$.*

where, $a(n) \ll b(n)$ means that $\lim_{n\to\infty} \frac{a(n)}{b(n)} = 0$

We give below a sufficient condition for the existence of threshold functions. Before doing that, let us first introduce the notion of edge-monotone property.

Definition 4 (Edge-monotone property)**.** *A graph property P is said to be edge-monotone increasing (decreasing) if for any graph G satisfying P, a graph G' obtained from G by adding (deleting) an edge would also satisfy P.*

Theorem 7 ([5])**.** *Any edge-monotone graph property always admits a threshold function.*

Thus, if we can show that the property of a player winning the one cop and one robber game is an edge-monotone property, that would ensure the existence of threshold functions. However, that is not the case.

Proposition 1. *The property of a player winning the game of one cop and one robber played on a graph is neither edge-monotone increasing nor edge-monotone decreasing.*

Proof. Without loss of generality, consider the case of the robber. Let us call the property of the robber having a winning strategy as Win_{R}. Consider the graph C_4, a cycle with 4 vertices. In this graph, the robber can always evade the cop by choosing to stay at a vertex that is diagonally opposite to that of cop. Thus, C_4 satisfies Win_{R}. Now, if we add a diagonal edge, we will get a diamond D_4, where if the cop chooses to stay at any of the vertices with degree 3, she can catch the robber in the next move. Thus, D_4 does not satisfy Win_{R}. Similarly, if we remove any edge from C_4, we will get a four-vertex path P_4. In P_4, the cop has a winning strategy. Hence, the property of the robber winning the game is neither edge-monotone increasing nor edge-monotone decreasing. Similarly, the property of the cop winning the game is neither edge-monotone increasing nor edge-monotone decreasing. □

Let us now get back to the extension axioms to see whether those ideas can shed some light on the threshold functions. As we have shown previously in Theorem 2, $EA_{1,2} := \forall x \forall y \exists z (\neg(x = y) \to E(y,z) \wedge \neg E(x,z))$ is almost surely true, and thus robber has a winning strategy in almost all possible one cop and

one robber games played on random graphs with constant edge probability. Now, with the varying edge probability $p(N)$, the probability that $EA_{1,2}$ is satisfied is given by $1-(1-p(N)(1-p(N)))^N$. Thus, if $\lim_{N\to\infty}(1-p(N)(1-p(N)))^N = 0$, then the robber has a winning strategy in $G(N,p(N))$ almost surely and also in $G(N,(1-p(N)))$. Thus, for the property $EA_{1,2}$, we cannot have a threshold function, and so we cannot move forward as in Section 3. Even though the extension axioms may not work for getting threshold functions we have the following result which serves our purpose to some extent.

Theorem 8 ([18])**.** *Any sentence in the first-order theory of graphs is either almost surely true or almost surely false in random graphs $G(N,p(N))$, with $p(N)$ satisfying any one of the following conditions.*

1. $p(N) \ll N^{-2}$
2. *for some integer k,* $N^{-1-\frac{1}{k}} \ll p(N) \ll N^{-1-\frac{1}{k-1}}$
3. $N^{-1-\varepsilon} \ll p(N) \ll n^{-1}$ *for all* $\varepsilon > 0$
4. $N^{-1} \ll p(N) \ll N^{-1}\log N$

The proof idea is quite similar to that of Theorem 1, the only difference being that the set of axioms that is considered in each of the above cases is somewhat more involved. For a detailed proof, see [18]. Let us now apply this result to find the threshold functions corresponding to the existence of winning strategies in the single cop and single robber game. We use relevant axioms used in the proof of the theorem above.

Theorem 9. *In a one cop and one robber game played on a random graph $G(N,p(N))$ where $p(N)$ is the varying edge probability and $N \to \infty$, the robber almost always has a winning strategy, whenever any one of the four conditions of Theorem 8 holds.*

Proof. We prove the result as follows: (1) For $p(N) \ll N^{-2}$, consider the formula $\forall x \forall y \neg(E(x,y))$ which is almost surely true. Evidently, the robber has a winning strategy. (2) For $N^{-1-\frac{1}{k}} \ll p(N) \ll N^{-1-\frac{1}{k-1}}$, consider the formula $\varphi_{t,r}$ saying that: there are at least r components t, t being a tree on at most k points. For example, if we take t_1 as the tree with one vertex, then the formula $\varphi_{t_1,2}$ is given by $\exists x_1 \exists x_2 \forall y(((y = x_1) \vee \neg(E(y,x_1))) \wedge ((y = x_2) \vee \neg(E(y,x_2))))$. Now, $\varphi_{t,r}$ is almost surely true for any r and any such tree t, in particular $\varphi_{t,2}$. This implies that the robber has a winning strategy as there is more than one component, almost surely. (3) For $N^{-1-\varepsilon} \ll p(N) \ll N^{-1}$ for all $\varepsilon > 0$, the proof is similar to the case above, based on the corresponding axioms discussed in [18]. (4) For $N^{-1} \ll p(N) \ll N^{-1}\log N$, the proof is once again similar.

□

Along similar lines we have the following result for the case of $k > 1$ cops and 1 robber.

Theorem 10. *In a $k > 1$ cops and one robber game played on a random graph $G(N,p(N))$ where $p(N)$ is the varying edge probability and $N \to \infty$, the robber almost always has a winning strategy, whenever each of the four conditions of Theorem 8 holds.*

Proof. The proofs for all the cases are as above with necessary modifications with respect to the number k. As an example, for the second case, we need to consider $k+1$ components instead of 2 components that we did for the one-cop case. That is, we need to consider $\varphi_{t,k+1}$, and the proof would go through as earlier. To give an example as earlier, if t_1 denotes the tree with one vertex, $\varphi_{t_1,k+1}$ is given by: $\exists x_1 \exists x_2 ... \exists x_k \exists x_{k+1} \forall y(\bigwedge_{i=1}^{k+1}((y = x_i) \vee \neg(E(y,x_i))))$. □

4.1 Other variants

Similar results can be proved for the *cops and robber game with traps* and *cops and robber games with roadblocks* played on random graphs $G(N,p(N))$ with varying edge probability and $N \to \infty$, and we leave it to the readers. The case of games with *different edge sets* played on random graphs with varying edge probabilities is more involved as we see below.

For the complementary edge sets, we basically consider a graph and its edge-complement graph. Without loss of generality, we assume that the robber is traveling through the original graph, and the cop is traveling through the complement. Evidently, the property Win_{R} is monotone increasing, which implies that it would always have a threshold function. As in the constant probability case, the robber wins whenever the underlying graph satisfies the extension axiom: $\forall x \forall y \exists z(\neg(x=y) \to E(y,z) \wedge E(x,z))$. Our aim is now to compute the threshold function. To this end, let us first introduce the following concepts that are discussed in [19].

Definition 5 (Rooted graph)**.** *A graph $H=(V_H,E_H)$ can be considered as a rooted graph (H,R), where $R \subseteq V_H$ is called the root.*

Definition 6 (Extension statement)**.** *A graph G is said to satisfy Extension statement $(Ext(H,R))$ where $R=\{x_1,\ldots,x_r\}$ and $V_H=\{x_1,\ldots,x_r,y_1,\ldots,y_s\}$, if for all choices of $\{x'_1,\ldots,x'_r\} \subseteq V_G$, we can find $\{y'_1,\ldots,y'_s\} \subseteq V_G$ such that whenever $(x_i,y_j) \in E_H$, $(x'_i,y'_j) \in E_G$ and whenever $(y_i,y_j) \in E_H$, $(y'_i,y'_j) \in E_G$.*

We note that a graph G satisfying the formula $\forall x \forall y \exists z(E(x,z) \wedge E(z,y))$ is the same as G satisfying the extension statement $Ext(H,R)$ where $R=\{x,y\}$, $V_H=\{x,y,z\}$ and $E_H=\{(x,z),(y,z)\}$. And as discussed in Theorem 5, whenever a graph satisfies $Ext(H,R)$, the robber has a winning strategy.

Some more notions from [19] are introduced for the sake of completion: (1) **Subextension:** We call a rooted graph (S,R) a subextension of a rooted graph (H,R) if S is an induced subgraph of H. It is proper if V_S is a proper subset of V_H. (2) **Density of a rooted graph:** Density of a rooted graph $dens(H,R)$ is given by $\frac{|E_H \setminus E_R|}{|V_H|}$. (3) **Strictly balanced:** A rooted graph (H,R) is called strictly balanced if for all subextensions (S,R), $dens(S,R) < dens(H,R)$. (4) **Maximal average degree:** A maximal average degree of a rooted graph (H,R) is given by $mad(H,R) := max\{dens(S,R) | (S,R)$ is a subextension of $(H,R)\}$ (5) **Primal subextension:** We call a subextension (S,R) of (H,R) a primal subextension if $dens(S,R) = mad(H,R)$. (6) **Grounded subextension:** We call a subextension grounded if there is at least one edge between a root and a non-root vertex. All these concepts are used in the following result that provides us with the required threshold function.

Theorem 11 ([19])**.** *For a rooted graph (H,R), the following holds: (i) If no primal subextension of (H,R) is grounded, then $f(n)=n^{-\frac{1}{mad(H,R)}}$ is a theshold function for the property $Ext(H,R)$. (ii) If there are grounded primal subextensions and s is the smallest value of $|E_S \setminus E_R|$ over all such subextensions S, then $f(n)=n^{-\frac{1}{mad(H,R)}}(\log n)^{\frac{1}{s}}$ is a threshold function for $Ext(H,R)$.*

We now have the following threshold for the cops and robber game with complementary edge sets.

Theorem 12. *In a cops and robber game with complementary edge sets played on a random graph $G(N,p(N))$ where $p(N)$ is the varying edge probability and $N \to \infty$, the robber has a winning strategy almost surely, whenever $N^{-\frac{3}{2}}(\log N)^{\frac{1}{2}} \ll p(N)$. Similarly, the cop has a winning strategy almost surely, whenever $p(N) \ll N^{-\frac{3}{2}}(\log N)^{\frac{1}{2}}$.*

Proof. We prove below the result for the robber. In this case, we have the extension axiom $\forall x \forall y \exists z(E(x,z) \wedge E(z,y))$, and corresponding to the axiom, we have the extension statement $Ext(H,R)$ where $R=\{x,y\}$, $V_H=\{x,y,z\}$ and $E_H=\{(x,z),(y,z)\}$. Note that in the extension statements, for all choices of root vertices, we get a

choice of z such that it is incident to the required edges. Now, (H,R) has only one subextension, which is (H,R) itself. It is primal as well as grounded as there is an edge between a root vertex and a non-root vertex, (e.g., (x,z) and (y,z)). Now, $mad(H,R) = \frac{|E_H \setminus E_R|}{|V_H|} = \frac{2}{3}$ and $s = |E_H \setminus E_R| = 2$. As there is a primal subextension which is grounded, using the previous theorem, we get the threshold function of $Ext(H,R) = N^{-\frac{3}{2}}(\log N)^{\frac{1}{2}}$. And, as a graph satisfying $Ext(H,R)$ implies that the robber has a winning strategy, $N^{-\frac{3}{2}}(\log N)^{\frac{1}{2}} \ll p(N)$ implies that the robber has a winning strategy in $G(N,p(N))$ almost surely. The case for the cop can be shown similarly, with respect to the extension statement given by $\neg\varphi'$, where φ' is given by $\exists x\forall y\forall z(\neg E(y,z) \vee \neg E(x,z))$. □

In the proof above, we have extension statements for the extension axioms but not the earlier ones, e.g., the cases of $k \geq 1$ cops and robber, as these statements do not preserve the property of non-adjacency of two vertices (note that, $\neg\varphi'$ above is also all about adjacency).

Finally, we look into the game of *tandem-cops* and robber. The tandem-cops have a winning strategy if the graph satisfies the extension axiom: $\forall x_1\forall x_2\forall y\exists z((\neg(x_1 = x_2) \wedge \bigwedge_{i=1,2} \neg(x_i = y)) \rightarrow (\bigwedge_{i=1,2} \neg(x_i = z) \wedge \neg(y = z) \wedge E(x_1,z) \wedge E(y,z)))$. This does not contain any non-adjacency condition. Thus, following the same extension statement idea, we have:

Theorem 13. *In a tandem-cops and robber game played on a random graph $G(N,p(N))$ where $p(N)$ is the varying edge probability and $N \to \infty$, the cops have a winning strategy almost surely, whenever $N^{-\frac{3}{2}}(\log N)^{\frac{1}{2}} \ll p(N)$.*

5 Concluding remarks

In conclusion, we can say that in accordance with our intuition, in almost all cases, it is possible for the robber to hide in large random graphs. The cops have a better chance in capturing the robber if they move in a synchronous way so that the distance between them always remains less than two units. In fact, the tandem-cops, who can move two units at a time, move twice the speed of the robber. If the winning condition is edge-monotone, we get a proper threshold function. The question remains for the non-monotone cases, which we would like to work on in the future. That all these results can be shown as an application of first order theory of random graphs gives us the added benefit of exploring the inherent connection between logic, probability, games and combinatorics.

Acknowledgements

The authors thank Johan van Benthem for a discussion on the relevant results of sabotage games, which initiated the work on this paper. They also thank the TARK 2025 reviewers for their detailed comments which helped to improve the paper.

References

[1] M. Aigner & M. Fromme (1984): *A game of cops and robbers*. Discrete Applied Mathematics 8(1), pp. 1–12, doi:10.1016/0166-218X(84)90073-8.

[2] J. van Benthem (2005): *An Essay on Sabotage and Obstruction*. In D. Hutter & W. Stephan, editors: *Mechanizing Mathematical Reasoning: Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday*, Springer, pp. 268–276, doi:10.1007/978-3-540-32254-2_16.

[3] J. van Benthem & F. Liu (2020): *Graph Games and Logic Design*. In F. Liu, H. Ono & J. Yu, editors: *Knowledge, Proof and Dynamics*, Logic in Asia: Studia Logica Library, Springer, pp. 125–146, doi:10.1007/978-981-15-2221-5_7.

[4] B. Bollobás, G. Kun & I. Leader (2013): *Cops and robbers in a random graph*. *Journal of Combinatorial Theory, Series B* 103(2), pp. 226–236, doi:10.1016/j.jctb.2012.10.002.

[5] B. Bollobás & A.G. Thomason (1987): *Threshold functions*. *Combinatorica* 7, pp. 35–38, doi:10.1007/BF02579198.

[6] A. Bonato & R. J. Nowakowski (2011): *The Game of Cops and Robbers on Graphs*. *The Student Mathematical Library* 61, AMS, Providence, doi:10.1090/stml/061.

[7] T.H. Chung, G.A. Hollinger & V. Isler (2011): *Search and pursuit-evasion in mobile robotics*. *Autonomous Robots* 31(4), pp. 299–316, doi:10.1007/s10514-011-9241-4.

[8] N. Clarke & R. Nowakowski (2001): *Cops, robber and traps*. *Utilitas Mathematica* 60, pp. 91–98.

[9] Paul Erdős & Alfréd Rényi (1959): *On Random Graphs*. *Publicationes Mathematicae Debrecen* 6, pp. 290–297, doi:10.5486/PMD.1959.6.3-4.12.

[10] Ronald Fagin (1976): *Probabilities on Finite Models*. *The Journal of Symbolic Logic* 41, pp. 50–58, doi:10.2307/2272945.

[11] E.N. Gilbert (1959): *Random Graphs*. *The Annals of Mathematical Statistics* 30(4), pp. 1141–1144, doi:10.1214/aoms/1177706098.

[12] M.W. Hasan & L.G. Ibrahim (2024): *A pursuit-evasion game robot controller design based on a neural network with an improved optimization algorithm*. *Results in Control and Optimization* 17, p. 100503, doi:10.1016/j.rico.2024.100503.

[13] D. Li, S. Ghosh, F. Liu & Y. Tu (2023): *A simple logic of the hide and seek game*. *Studia Logica* 111, pp. 821–853, doi:10.1007/s11225-023-10039-4.

[14] Leonid Libkin (2010): *Elements of Finite Model Theory*. Springer, doi:10.1007/978-3-662-07003-1.

[15] K. Mierzewski (2025): *When random graphs are safe for travel:a note on the sabotage game*. In J. van Benthem & F. Liu, editors: *Graph Games and Logic Design: Recent Developments and Future Directions*, Springer, doi:10.1007/978-981-15-2221-5_7.

[16] R. Nowakowski & P. Winkler (1983): *Vertex-to-vertex pursuit in a graph*. *Discrete Mathematics* 43(2), pp. 235–239, doi:10.1016/0012-365X(83)90160-7.

[17] A. Quilliot (1978): *Jeux et pointes fixes sur les graphes*. Ph.D. thesis, Université de Paris VI.

[18] S. Shelah & J.H. Spencer (1988): *Zero-one laws for sparse random graphs*. *Journal of the American Mathematical Society* 1, pp. 97–115, doi:10.1090/S0894-0347-1988-0924703-8.

[19] J. Spencer (1990): *Threshold functions for extension statements*. *J. Comb. Theory Ser. A* 53(2), p. 286–305, doi:10.1016/0097-3165(90)90061-Z.

[20] M. Stojaković & T. Szabó (2006): *Positional games on random graphs*. *Random Structures & Algorithms* 26, doi:10.1002/rsa.20059.

Prudent Rationalizability and the Best Rationalization Principle

Nicodemo De Vito
Department of Economics and Statistics
University of Salerno
Salerno, Italy
ndevito@unisa.it

We study cautious reasoning in finite sequential games played by agents with perfect recall. Our contribution lies in formulating a definition of *prudent rationalizability* (Heifetz et al. 2021, *BEJTE*) as an iterative reduction procedure of beliefs. To this end, we represent the players' beliefs by systems of conditional *non-standard* probability measures. The key novelty is the notion of *c-strong belief*, a non-standard, "cautious" version of strong belief (Battigalli and Siniscalchi 2002, *JET*). Our formulation of prudent rationalizability embodies a "best rationalization principle" similar to the one that underlies the solution concept of strong rationalizability. The main results show the equivalence between the proposed definition with the one originally put forth by Heifetz et al. (2021) in terms of conditional beliefs represented by standard probabilities. In particular, it is shown that prudent rationalizability can be algorithmically characterized by iterated admissibility. Finally, our formulation can be extended to sequential games with unawareness.

Keywords: Cautious reasoning, conditional non-standard probability systems, cautious belief, iterated admissibility, sequential games, rationalizability, non-standard analysis.

1 Introduction

Prudent rationalizability is a solution concept for games where some moves are sequential, henceforth *sequential games*. It was introduced by Heifetz et al. (2021) for the analysis of cautious reasoning in the framework of sequential games with *unawareness* (Heifetz et al., 2013), i.e., games which allow for asymmetric awareness of the players concerning their actions. As argued by the authors, studying cautious reasoning is natural in such a context since, as the play unfolds, players frequently become aware of actions they (or the co-players) could have chosen. With this, they put forward the notion of prudent rationalizability as a "cautious" analogue of *strong rationalizability*, aka extensive-form rationalizability (Pearce 1984, Battigalli 1997), a solution concept that captures forward-induction thinking. In the context of limited awareness, prudent rationalizability proved to be important in applications of interest—see Meier and Schipper (2024) and the relevant references therein.

The aim of our work is to address an issue left open by Heifetz et al. (2021, Section 6). Specifically, their definition of prudent rationalizability is in terms of an iterative reduction procedure of strategies. The open question is whether a definition via a reduction procedure for conditional beliefs—in the style of strong rationalizability—is feasible. Such an alternative definition would be desirable because, as we will argue below, it would shed light on the possible epistemic foundations for prudent rationalizability.

To address the above question, here we study sequential games played by agents with perfect recall, without assuming unawareness. However, the definitions and results can be extended to the case of games with unawareness. We will show that our formulation of prudent rationalizability provides a positive resolution to the open issue in Heifetz et al. (2021). We will also show that such a formulation

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 334–349, doi:10.4204/EPTCS.437.27

© N. De Vito
This work is licensed under the
Creative Commons Attribution License.

embodies a *best rationalization principle* similar to the one that underlies the solution concept of strong rationalizability (Battigalli 1996).

We base our analysis on two key notions which pertain to the representation of players' beliefs. First, we represent the players' beliefs by *conditional non-standard probability systems* (CNPSs): as the play unfolds, each players holds conditional beliefs about the co-players' strategies which are updated or revised according to the rules of conditional probability. More importantly, such (conditional) beliefs are represented by probability measures whose range is a subset of the hyperreal field ${}^*\mathbb{R}$, a non-Archimedean field that includes the real numbers but also has *infinitesimals* (i.e., numbers that are positive but smaller than than any positive number). This representation of beliefs allows to formalize the idea that a player is both cautious (no strategy of the co-players is ruled out) and, at the same time, he/she deems an event infinitely more likely than another.

The other relevant notion is *c-strong belief*, a "cautious" version of strong belief (Battigalli and Siniscalchi 2002). We say that a player c-strongly believes an event E if he/she *cautiously believes* it at the beginning of the game; as the play unfolds, he/she continues to do so as long as E is not falsified by the evidence. The concept of cautious belief was introduced by Catonini and De Vito (2024) for their epistemic analysis of iterated weak dominance in finite static games. Loosely speaking, a player cautiously believes an event E if he/she deems every co-players' strategy in E infinitely more likely than the strategies in the complement not-E. As shown by Catonini and De Vito (2024), this notion of belief embodies a form of caution relative to the event.

We illustrate our approach with two results. First, we prove (Theorem 1 in Section 6) that our formulation of prudent rationalizability can be given an algorithmic characterization in terms of iterated admissibility, i.e., the set of strategies surviving the maximal iterated deletion of weakly dominated strategies. Thus, although the conceptual definition of prudent rationalizability depends on the extensive-form structure of the game, it has a "strategic-form" flavor from an operational standpoint (i.e., in terms of finding out the set of strategies). Theorem 1 is, in fact, a version—for the case of games without unawareness—of Theorem 3 in Meier and Schipper (2024), which is stated under the definition of prudent rationalizability given by Heifetz et al. (2021). Our second result (Theorem 2 in Section 6) shows the equivalence between our definition of prudent rationalizability and the one put forth by Heifetz et al. (2021). As already mentioned, this equivalence extends to sequential games with unawareness.

The present work is part of a research program concerning the epistemic foundations of prudent rationalizability. A paper in progress by the author of this work shows that, by means of hypermethods—mainly, the *transfer principle* of non-standard analysis—a "canonical" epistemic model for hierarchies of CNPSs exists. With this, it should be possible to formally provide epistemic conditions for prudent rationalizability: the natural starting point is Theorem 3 in Catonini and De Vito (2024), which provides an epistemic foundation for *lexicographic rationalizability* (Stahl 1995)—a version of prudent rationalizability for the case of static games—under the assumption called "transparency of cautiousness."[1]

Organization. Section 2 introduces the required mathematical notation and auxiliary notions from non-standard analysis. Section 3 introduces the game-theoretic framework and formalizes the players'

[1]The following issue is quite subtle, and can be properly understood by a reader who is familiar with Heiftez at al. (2021, Sections 4.2-4.3). As argued by the authors, there is a tension between caution (or "prudence") and the logic of the rationalization underlying forward-induction thinking. A strategy can be incautious (or "imprudent") but strongly rationalizable. Heifetz et al. (2021, Section 4.2) argue that "...the definition of prudent rationalizability resolves this tension unequivocally in favor of the prudence consideration." In our view, the implicit assumption underlying the above statement is that caution is "transparent" in the terminology of Catonini and De Vito (2024) or Battigalli and De Vito (2021): whenever a history reveals that a player's strategy is inconsistent with prudent rationalizability, such a strategy is always interpreted by the co-players as cautious ("prudent"), but it is not strongly rationalizable. Put differently, caution is given **doxastic** (or "**epistemic**") **priority** at histories inconsistent with the predicted path of play (cf. Battigalli and De Vito 2021, Section 6).

beliefs. Section 4 defines c-strong belief. Section 5 introduces the proposed definition of prudent rationalizability. Section 6 contains the results. All proofs are in the Appendix.

2 Preliminaries

This section reviews some basic concepts from non-standard analysis pertaining to the elements of ${}^*\mathbb{R}$, the hyperreal line—see, e.g., Goldblatt (1998) for an excellent introduction. Here, we restrict attention to the set of non-negative hyperreal numbers less than or equal to 1, denoted by ${}^*[0,1]$. Specifically, ${}^*[0,1]$ is a non-standard extension of the set of real numbers $[0,1]$, and it contains the latter. In particular, ${}^*[0,1]$ also contains positive infinitesimal numbers, i.e., numbers strictly greater than zero but smaller than any positive real number. The existence of such a non-standard extension of $[0,1]$ follows from the ultrapower construction of the hyperreal field—see Luxemburg (1962).

Formally, a number $x \in {}^*\mathbb{R}$ is an **infinitesimal** if $|x| < 1/n$ for each natural number $n \in \mathbb{N}$. The unique infinitesimal real number is 0. For any $x, y \in {}^*[0,1]$, say that x and y are **infinitely close**, and write $x \simeq y$, if $|y-x|$ is an infinitesimal. The set of positive infinitesimals is denoted by $\text{Hal}^+(0)$, i.e., the set of all $x \in {}^*[0,1]$ which are infinitely close to 0.[2] The **standard part** of $x \in {}^*[0,1]$, denoted by $\text{st}(x)$, is the unique real number y such that $x \simeq y$. Every $x \in {}^*[0,1]$ can be uniquely written as $x = y + \varepsilon$, where $y = \text{st}(x)$ and $\varepsilon \in {}^*\mathbb{R}$ is infinitesimal.

Given $x, y \in {}^*\mathbb{R}^+$, say that x is **infinitely greater than** y (or y is **infinitely smaller than** x) if $x > ny$ for all $n \in \mathbb{N}$. Note that if x is infinitely greater than y, then $x > 0$. It is easy to check that x is infinitely greater than y *if and only if* $\text{st}(y/x) = 0$.[3]

Given any *finite* set X, probability measures on X are defined from the power set of X to ${}^*[0,1]$. Such probabilities are called non-standard. We sometimes say that a probability measure is standard if its range is the set $[0,1] \subseteq \mathbb{R}$. We let ${}^*\Delta(X)$ and $\Delta(X)$ denote the sets of, respectively, non-standard and standard probabilities on X. The latter set can be regarded as a subset of the former, and it makes sense to write $\Delta(X) \subseteq {}^*\Delta(X)$.

3 Framework

We introduce the building blocks of the analysis, namely, finite games with observed actions (Section 3.1) and systems of conditional non-standard probabilities (Section 3.2).

3.1 Finite Games with Observed Actions

Throughout, we focus on finite multistage games with perfect monitoring of past actions.[4]

A **finite game with observed actions** is represented by a structure $G := \langle I, \bar{H}, (A_i, u_i)_{i \in I} \rangle$ where:

- I is a finite set of **players**, and, for each $i \in I$, A_i is a finite, nonempty set of feasible **actions**.

[2] In the terminology of non-standard analysis, $\text{Hal}^+(0)$ is the positive part of the **monad**, or **halo**, of zero.

[3] Let $x > ny$ for each $n \in \mathbb{N}$. Then $y/x < 1/n$ for each $n \in \mathbb{N}$, i.e., y/x is infinitesimal, which yields $\text{st}(y/x) = 0$. For the reverse implication, we argue by contraposition. Suppose that $x \leq ny$ for some $n \in \mathbb{N}$. Then $y/x \geq 1/n$ for some $n \in \mathbb{N}$, i.e., y/x is not infinitesimal; hence, $\text{st}(y/x) \neq 0$.

[4] The restriction to finite multistage games with observed actions is only for the sake of notational simplicity. The techniques and results in this paper can be extended to finite sequential games played by agents with perfect recall.

- $\bar{H}$ is a finite tree of feasible **histories**, that is, of sequences of action profiles $a \in A := \times_{i\in I} A_i$. The root of $\bar{H}$ is the **initial history**, or "empty sequence," and is denoted by $\varnothing$. We let Z denote the set of **terminal** histories (or **paths**), and $H := \bar{H}\backslash Z$ is the set of **non-terminal** histories.
- For each $h \in H$, the set of feasible action profiles

$$A(h) := \{a \in A : (h,a) \in \bar{H}\}$$

  is such that $A(h) = \times_{i\in I} A_i(h)$, where $A_i(h)$ is the projection of $A(h)$ on A_i.
- For each $i \in I$, $u_i : Z \to \mathbb{R}$ is the utility function for player i.

The interpretation of G is that, as the game unfolds, each player is informed of the sequence of action profiles that has just occurred. Specifically, it is assumed more: as soon as a history h occurs it becomes common knowledge that h has occurred.

For any $h \in \bar{H}$, we write $h \prec h'$ ($h \preceq h'$) if sequence/history h is a strict (weak) prefix of h', and we say that h **(weakly) precedes** h' if ($h \preceq h'$) $h \prec h'$. Player i is **active** at history $h \in H$ if he has at least two feasible actions, and he is **inactive** otherwise.[5] There are simultaneous moves given h if at least two players are active at h. Game G is **static** if $H = \{\varnothing\}$.

The analysis in this paper focuses on the following derived objects. For each $i \in I$, let $S_i := \times_{h\in H} A_i(h)$ and $S := \times_{i\in I} S_i$. Each $s_i \in S_i$ is a **strategy** of player i, i.e., a function $s_i : H \to A_i$ such that $s_i(h) \in A_i(h)$ for all $h \in H$. The set of strategies of players other than i is $S_{-i} := \times_{j\in I\backslash\{i\}} S_j$.[6]

Each strategy profile $s := (s_i)_{i\in I} \in S$ induces a unique terminal history $\zeta(s)$, where $\zeta : S \to Z$ denotes the **path function**. With this, for each $h \in H$, the set of strategies inducing h is

$$S(h) := \{s \in S : h \prec \zeta(s)\}.$$

The projection

$$S_i(h) := \{s_i \in S_i : \exists s_{-i} \in S_{-i}, (s_i, s_{-i}) \in S(h)\}$$

is the set of strategies of i that **allow** h (i.e., do not prevent the realization of h).[7] Analogously,

$$S_{-i}(h) := \{s_{-i} \in S_{-i} : \exists s_i \in S_i, (s_i, s_{-i}) \in S(h)\}$$

is the set of strategy profiles of i's co-players that allow h. It is worth noting that, in a game with observed actions, $S(h) = \times_{i\in I} S_i(h)$ for each $h \in H$. Finally,

$$U_i := u_i \circ \zeta : S \to \mathbb{R}$$

determines the (strategic-form) payoff $U_i(s) = u_i(\zeta(s))$ of player i as a function of s.

3.2 Conditional Beliefs

For conciseness, the definitions in this section apply to a game $G := \langle I, \bar{H}, (A_i, u_i)_{i\in I}\rangle$. Each player's beliefs are represented by arrays of non-standard probability measures indexed by elements of a (finite) collection of "conditioning events," i.e., sets of the form $S_{-i}(h)$ for $h \in H$. Such events can be thought

[5] Whenever player i is not active at $h \in H$, one can think of the unique element of $A_i(h)$ as the "action" of waiting one's turn to move.

[6] Our notation is standard: for any profile of sets $(X_i)_{i\in I}$, we let $X_{-i} := \times_{j\in I\backslash\{i\}} X_j$ with typical element $x_{-i} := (x_j)_{j\neq i} \in X_{-i}$.

[7] If i carries out strategy s_i, then h can be reached; whether it is reached depends on the play of i's co-players.

of as representing the evidence that a player can obtain in a game. For instance, if history $h \in H$ occurs, player i learns that the co-players are behaving according to a strategy profile in $S_{-i}(h)$. If i's belief before the realization of history h assigned probability 0 to $S_{-i}(h)$, then the occurrence of h falsifies i's earlier belief, which has to be revised, rather than updated according to the chain rule of conditional probabilities (formally introduced below).

Let $\mathscr{S}_{-i} := \{F \subseteq S_{-i} : \exists h \in H, F = S_{-i}(h)\}$ denote the collection of conditioning (or "observable") events about the behavior of i's co-players. We also use the following notation: for any *finite* set X and **event** (subset) $E \subseteq X$, let ${}^*\Delta(E) := \{\mu \in {}^*\Delta(X) : \mu(E) = 1\}$ be the set of non-standard probability measures that assign probability 1 to E. As for standard probabilities, the set $\Delta(E)$ is defined accordingly.

Definition 1 *An array of probability measures $\mu := (\mu(\cdot|C))_{C \in \mathscr{S}_{-i}}$ is a **conditional non-standard probability system (CNPS)** for player $i \in I$ if*

(i) $\mu(\cdot|C) \in {}^*\Delta(C)$ for all $C \in \mathscr{S}_{-i}$, and

(ii) for all $C, D \in \mathscr{S}_{-i}$ and $E \subseteq S_{-i}$,

$$E \subseteq D \subseteq C \Rightarrow \mu(E|C) = \mu(E|D)\mu(D|C). \tag{3.1}$$

*The array μ is **conditional probability system (CPS)** if it is a CNPS such that $\mu(\cdot|C) \in \Delta(C)$ for all $C \in \mathscr{S}_{-i}$.*

Condition (3.1) is the **chain rule** of conditional probabilities, and it can be written as follows: if $E \subseteq D \subseteq C$, then

$$\mu(D|C) > 0 \Rightarrow \mu(E|D) = \frac{\mu(E|C)}{\mu(D|C)}.$$

We let $\Delta^{\mathscr{S}_{-i}}(S_{-i})$ and ${}^*\Delta^{\mathscr{S}_{-i}}(S_{-i})$ denote, respectively, the sets of CPSs and CNPSs for player $i \in I$. As argued above, the former set can be regarded as a subset of the latter, so $\Delta^{\mathscr{S}_{-i}}(S_{-i}) \subseteq {}^*\Delta^{\mathscr{S}_{-i}}(S_{-i})$. Throughout this paper, we will often use the notation $\mu := (\mu(\cdot|S_{-i}(h)))_{h \in H}$ to refer to an arbitrary element of ${}^*\Delta^{\mathscr{S}_{-i}}(S_{-i})$ (or $\Delta^{\mathscr{S}_{-i}}(S_{-i})$).

CPSs are used as representations of players' beliefs to model sophisticated strategic thinking in sequential games, such as forward-induction thinking (Battigalli and Siniscalchi 2002). However, as we will see below (Sections 4-6), there is a sense in which CNPSs are more expressive than CPSs: unlike a CPS, a CNPS also allows to properly formalize forms of strategic thinking based on notions of "caution."

We point out that a CNPS can be viewed as a non-standard analogue of a *system of conditional lexicographic probabilities* (SCLP), a concept introduced by Asheim and Perea (2005) in their analysis on cautious reasoning in sequential games. In contrast to a CNPS, an SCLP represents player i's beliefs—conditional on every $C \in \mathscr{S}_{-i}$—by means of *lexicographic probability systems*, i.e., finite sequences of probability measures, where the different measures are given in a decreasing order of importance. In other words, and using the notation of the current paper, each conditional belief $\mu(\cdot|C)$ for player i is a lexicographic probability system. The axiomatic foundation of non-Archimedean expected utility in Blume et al. (1991) shows that non-standard probabilities and lexicographic probability systems are equivalent in terms of representations of players' preferences.

That said, it makes sense to consider SCLPs as representations of players' conditional beliefs.[8] However, we represent players' beliefs as CNPSs since this representation allows a clear-cut comparison between prudent rationalizability and other solution concepts, such as strong rationalizability.

[8]This alternative option was suggested by Heifetz et al. (2001, Section 6) for the formulation of prudent rationalizability as an iterative reduction procedure for conditional beliefs. Some belief operators defined in Section 4 in terms of non-standard probabilities can be equivalently described in terms of lexicographic probability systems—see Catonini and De Vito (2020, 2024). The alternative characterization in terms of SCLPs makes the analysis more complex, but it would not change the substance of our results.

4 c-Strong Belief

Here, we define the concept of *c-strong belief*, which will be used for the formulation of prudent rationalizability. All definitions in this section apply to a given game $G := \langle I, \bar{H}, (A_i, u_i)_{i \in I} \rangle$. For the reader's convenience, it may be useful to consult Section 2 for the mathematical formalism used here.

Fix a CNPS $\mu := (\mu(\cdot|S_{-i}(h)))_{h \in H}$ for a player $i \in I$ and fix an event $E \subseteq S_{-i}$. Intuitively, we say that E is c-strongly believed under μ if, for every history h consistent with E, every strategy profile s_{-i} in E is deemed infinitely more likely than every strategy profile in not-E (i.e., the complement of E). So, to define c-strong belief, we first need to understand the idea of "infinitely more likely than."

Formally, fix a history $h \in H$ and events $E, F \subseteq S_{-i}(h)$ such that $E \cap F \neq \emptyset$. We say that E is **infinitely more likely than** F under $\mu(\cdot|S_{-i}(h))$ if

$$\text{st}\left(\frac{\mu(F|S_{-i}(h))}{\mu(E|S_{-i}(h))}\right) = 0.$$

Put differently, this definition says that E is infinitely more likely than F if $\mu(E|S_{-i}(h))$ is *infinitely greater than* $\mu(F|S_{-i}(h))$.

Lo (1999) provided a preference-based definition of "infinitely more likely than." As shown in Catonini and De Vito (2020, 2024), Lo's notion admits a characterization in terms of non-standard probabilities which is *equivalent* to the above definition of "infinitely more likely than." Lo's notion is also used—at least implicitly—in other papers, such as Meier and Perea (2024).[9] Similar notions of "infinitely more likely than" can be found in the literature concerning the epistemic foundations for iterated admissibility; see Brandenburger et al. (2008), Lee (2016).

Definition 2 *Fix a player $i \in I$, a CNPS $\mu := (\mu(\cdot|S_{-i}(h)))_{h \in H}$ and an event $E \subseteq S_{-i}$.*

*(i) E is **cautiously believed** under $\mu(\cdot|S_{-i}(h))$ $(h \in H)$ if $\text{st}\left(\frac{\mu(S_{-i} \setminus E|S_{-i}(h))}{\mu(s_{-i}|S_{-i}(h))}\right) = 0$ for all $s_{-i} \in E \cap S_{-i}(h)$.*

*(ii) E is **c-strongly believed** under μ if, for all $h \in H$ such that $E \cap S_{-i}(h) \neq \emptyset$, event E is cautiously believed under $\mu(\cdot|S_{-i}(h))$.*

Before discussing the notions in Definition 2, we point out that c-strong belief admits an equivalent but more cumbersome definition.

Remark 1 *Since $\mu(E|S_{-i}(h)) = \mu(E \cap S_{-i}(h)|S_{-i}(h))$ for all $h \in H$, event E is c-strongly believed under μ if and only if, for all $h \in H$ such that $E \cap S_{-i}(h) \neq \emptyset$, and for all $s_{-i} \in E \cap S_{-i}(h)$,*

$$\text{st}\left(\frac{\mu((S_{-i} \setminus E) \cap S_{-i}(h)|S_{-i}(h))}{\mu(s_{-i}|S_{-i}(h))}\right) = 0.$$

Cautious belief was introduced by Catonini and De Vito (2024) for the analysis of iterated admissibility in finite static games. It requires that *every* strategy in E be deemed infinitely more likely than (every strategy in) not-E. In particular, it embodies a *cautious* attitude of player i towards the "believed" event.

[9] Meier and Perea (2024) formalize the idea of "infinitely more likely than" by means of an "infinitely smaller size than" relation between non-standard numbers. It can be easily shown that such a relation can be given a characterization in terms of the "infinitely greater than" relation—hence, the definition of "infinitely more likely than" in Meier and Perea (2024) is equivalent to ours. Unlike us, Meier and Perea (2024) use a non-Archimedean extension of the real line that is different from the hyperreal line. Conceptually, this difference is immaterial: our definitions and results still hold under the assumption that the non-standard probability measures take value in any non-Archimedean extension of the real field.

Indeed, cautious belief requires that: (1) event E is deemed infinitely more likely than not-E, and (2) before entertaining the possibility that E does not occur, player i takes into account all the possible payoff-relevant consequences (i.e., the strategy profile s_{-i}) of E. Condition (1) corresponds to the notion of **weak belief** (Halpern 2010, Catonini and De Vito 2020), namely $\text{st}(\mu(E|S_{-i}(h))) = 1$. Condition (2) says that the player is cautious towards the (weakly) believed event E: Before considering not-E, he/she takes into account all the possible consequences of E. Thus, cautious belief in E is stronger than weak belief as it captures caution relative to E.

Remark 2 *To further see how cautious belief embodies a form of caution (as suggested by the terminology) relative to E, it could be useful to consider the special case when $E = S_{-i}$. With this, at the initial history $\varnothing$ we have $\mu(s_{-i}|S_{-i}) > 0$ for all $s_{-i} \in S_{-i}$, i.e., every strategy profile of i's co-players is deemed possible. By the chain rule, $\mu(s_{-i}|S_{-i}(h)) > 0$ for all $h \in H$ and $s_{-i} \in S_{-i}(h)$.*

That being said, the intended interpretation of c-strong belief is simple: Player i c-strongly believes event E if he/she cautiously believes it at the beginning of the game; as the play unfolds, he/she continues to do so as long as E is not falsified by the evidence—that is, unless a history h such that $E \cap S_{-i}(h) = \emptyset$ occurs.

For any event $E \subseteq S_{-i}$, let $\text{cSB}_i(E)$ denote the set of CNPSs of player i under which E is c-strongly believed. The reason for the terminology "c-strong belief" is clarified below. Before doing this, we find it convenient to single out the following fact, which will be useful in the next section.

Remark 3 *c-Strong belief is not monotone: for all events $E, F \subseteq S_{-i}$, if $E \subseteq F$, then it is not the case that $\text{cSB}_i(E) \subseteq \text{cSB}_i(F)$.*

In particular, non-monotonicity of c-strong belief follows from non-monotonicity of cautious belief: as shown in Catonini and De Vito (2024), the reason why cautious belief fails monotonicity is that player i may not have towards F the same cautious attitude that he has towards $E \subseteq F$. In particular, there could be some strategies in $F \backslash E$ which are not deemed infinitely more likely than not-F (see Section 4.2 in Catonini and De Vito, 2024, for an example).

The notion of c-strong belief can be viewed as a "cautious analogue" of strong belief, a belief modality put forth by Battigalli and Siniscalchi (2002) for the epistemic analysis of forward-induction reasoning in sequential games. Strong belief is defined in terms of CPSs, but it can also be defined in terms of CNPSs: an event E is **strongly believed** under μ if $\mu(E|S_{-i}(h)) = 1$ for all $h \in H$ such that $E \cap S_{-i}(h) \neq \emptyset$. Unlike c-strong belief, the notion of strong belief does not embody caution. However, like c-strong belief, also strong belief does not satisfy monotonicity (see Battigalli and Siniscalchi 2002, Section 3.2).

5 Solution Procedure

In the remainder of this section, we fix a game $G := \langle I, \bar{H}, (A_i, u_i)_{i \in I} \rangle$. To introduce our proposed definition of prudent rationalizability, we first need a notion of optimality for a strategy, given a CNPS.

Fix a player $i \in I$. For each $s_i \in S_i$ and $h \in H$, let s_i^h denote the minimal modification of s_i allowing h. That is, s_i^h is the "h-replacement" strategy that, at every $h' \prec h$, chooses the unique action $\alpha_i(h', h) \in A_i(h')$ leading from h' toward h, and coincides with s_i at all other histories. For any CNPS μ_i, we let $\rho_i(\mu_i)$ denote the set of all **sequential best replies** to μ_i, that is,

$$\rho_i(\mu_i) := \left\{ s_i \in S_i : \forall h \in H, s_i^h \in \arg\max_{r_i \in S_i(h)} \mathbb{E}_{\mu_i}(U_i(r_i, \cdot)|h) \right\},$$

where

$$\mathbb{E}_{\mu_i}(U_i(r_i,\cdot)|h) := \sum_{s_{-i}\in S_{-i}(h)} U_i(r_i,s_{-i})\,\mu_i(s_{-i}|S_{-i}(h))$$

is the (conditional) expected payoff of r_i given $\mu_i(\cdot|S_{-i}(h))$. With this, we say that a strategy $s_i \in S_i$ is **justified by** μ_i (or μ_i **justifies** s_i) if $s_i \in \rho_i(\mu_i)$.

It should be noted that this notion of sequential best reply (given some μ_i) requires that a strategy s_i be optimal at *all* histories, including those that are precluded by the strategy itself. In Section 6, we will consider an alternative notion—often called *weak sequential optimality*—which requires that s_i be optimal only at histories that s_i allows—see Battigalli et al. (2024, Chapter 10).

Definition 3 *Consider the following procedure.*

(Step 0**)** *For every $i \in I$, let $S_i^0 := S_i$. Also, let $S_{-i}^0 := \times_{j\in I\setminus\{i\}} S_j$ and $S^0 := \times_{i\in I} S_i^0$.*

(Step $n>0$**)** *For every $i \in I$, let*

$$S_i^n := \left\{ s_i \in S_i : \exists \mu_i \in \cap_{m=0}^{n-1} \mathrm{cSB}_i(S_{-i}^m), s_i \in \rho_i(\mu_i) \right\}.$$

Also, let $S_{-i}^n := \times_{j\in I\setminus\{i\}} S_j^n$ and $S^n := \times_{i\in I} S_i^n$.

Finally, let $S_i^\infty := \cap_{n=0}^\infty S_i^n$ and $S^\infty := \times_{i\in I} S_i^\infty$.

Clearly, $(S^n)_{n\geq 0}$ is a decreasing sequence, and, as it can be shown by standard arguments, $S^\infty \neq \emptyset$. In particular, since the game is finite, the sequence becomes constant after some finite number $N \geq 0$ of steps.

Note that, in the recursive step, the procedure requires each player i to c-strongly believe all the events S_{-i}^m $(m = 0,...,n-1)$, not just S_{-i}^{n-1}. *Technically*, if we were to impose only the notion of c-strong belief in S_{-i}^{n-1}, then $\mathrm{cSB}_i(S_{-i}^{n-1})$ would not necessarily be a subset of $\mathrm{cSB}_i(S_{-i}^{n-2})$ because, by Remark 3, c-strong belief is not monotone. Consequently, we would not obtain a well-defined elimination procedure—i.e., $(S^n)_{n\geq 0}$ would not be decreasing.

That being said, we illustrate in some detail the *conceptual* features of the procedure; for simplicity, we focus on steps 1-3. Moreover, we will sometimes refer to the procedure as "prudent rationalizability." The reason for using such a terminology will be justified by the results in Section 6.

At step 1, player i plays a strategy s_i that is justified by some μ_i such that $\mu_i(s_{-i}|S_{-i}(h)) > 0$ for all $h \in H$ and $s_{-i} \in S_{-i}(h)$: to see this, note that $\mathrm{cSB}_i(S_{-i}^0) = \mathrm{cSB}_i(S_{-i})$ and refer to Remark 2. Thus, besides the notion of optimality for strategies, the procedure requires a form of caution (or "prudence") for the justifying CNPSs, in the sense that the players consider all strategy profiles of the co-players possible. In what follows, we will *informally* refer to this combination of optimality and caution as "cautious rationality."

Step 2 of the procedure requires that, on top of being cautiously rational, all the players c-strongly believe that the co-players are cautiously rational as well: formally, for each $i \in I$, the justifying conditional beliefs must belong to the set $\mathrm{cSB}_i(S_{-i}^1)$.

Step 3 allows to point out an important feature of the form of strategic thinking captured by prudent rationalizability. If player i c-strongly believes the events S_{-i}^1 and S_{-i}^2, the he/she c-strongly believes that the co-players' behavior is consistent with a certain degree of strategic sophistication—namely, cautious rationality and c-strong belief in others' cautious rationality. Suppose now that a history $h \in H$ contradicts such a degree of sophistication, that is, $S_{-i}^2 \cap S_{-i}(h) = \emptyset$. In this case c-strong belief in S_{-i}^2 leaves i's beliefs at h unrestricted, so i may well deem the co-players' strategies that are not cautiously

rational infinitely more likely than those consistent with $-i$'s cautious rationality; in particular, this can happen even if $S^1_{-i} \cap S_{-i}(h) \neq \emptyset$.[10] However, c-strong belief in both S^1_{-i} and S^2_{-i} rules out such a situation: whenever $S^1_{-i} \cap S_{-i}(h) \neq \emptyset$, the co-players' "irrational" strategies (i.e., those in $S_{-i} \backslash S^1_{-i}$) are deemed by i infinitely less likely than those consistent with cautious rationality.

The above argument illustrates that prudent rationalizability—as per Definition 3—relies on the **best rationalization principle** (Battigalli 1996): players always ascribe to their co-players the highest degree of strategic sophistication consistent with their past behavior, even when surprised by co-players' behavior.[11] The best rationalization principle is captured by c-strong belief in each event of the sequence $\left(S^n_{-i}\right)_{n\geq 0}$. Specifically, letting $m(h)$ denote the highest degree of strategic sophistication consistent with $h \in H$,[12] we have that $\mu_i \in \cap_{m=0}^{\infty} \text{cSB}_i(S^m_{-i})$ implies $\mu_i \in \cap_{m=0}^{m(h)} \text{cSB}_i(S^m_{-i})$ for all $h \in H$.

Finally, it is noteworthy that if game G is static (i.e., $H = \{\varnothing\}$), then c-strong belief corresponds to (unconditional) cautious belief. In this case, the procedure in Definition 3 is a non-standard version of **lexicographic rationalizability** (Stahl 1995), a solution concept based on lexicographic probabilities as representations of players' beliefs. Indeed, as shown in De Vito (2023), the definition of lexicographic rationalizability relies on the characterization of cautious belief in terms of lexicographic probability systems (cf. Catonini and De Vito 2024).

6 Results

We state the main results of this paper, whose proofs can be found in the appendix. To this end, we first need a formal definition of iteratively admissible strategies for a game. For conciseness, the definitions and results in this section apply to a fixed game $G := \left\langle I, \bar{H}, (A_i, u_i)_{i\in I} \right\rangle$.

Let $\mathscr{Q}$ be the collection of all subsets of S with the form $Q := \times_{i\in I} Q_i$, where $Q_i \subseteq S_i$ for every $i \in I$. In what follows, for any $\sigma_i \in \Delta(S_i)$, let $U_i(\sigma_i, s_{-i}) := \sum_{r_i \in S_i} \sigma_i(r_i)\, U_i(r_i, s_{-i})$.

Definition 4 *Fix a set $Q \in \mathscr{Q}$. A strategy $s_i \in Q_i$ is **weakly dominated with respect to** Q if there exists a mixed strategy $\sigma_i \in \Delta(S_i)$, with $\sigma_i(Q_i) = 1$, such that $U_i(\sigma_i, s_{-i}) \geq U_i(s_i, s_{-i})$ for every $s_{-i} \in Q_{-i}$ and $U_i(\sigma_i, s'_{-i}) > U_i(s_i, s'_{-i})$ for some $s'_{-i} \in Q_{-i}$. Otherwise, say s_i is **admissible with respect to** Q.*

*If $s_i \in S_i$ is weakly dominated (resp. admissible) with respect to S, say s_i is **weakly dominated** (resp. **admissible**).*

The set of iteratively admissible strategies is defined recursively.

Definition 5 *For every $i \in I$, let $\hat{S}^0_i := S_i$, and for every $n \geq 1$, let $\hat{S}^n_i$ be the set of all $s_i \in \hat{S}^{n-1}_i$ that are admissible with respect to $\hat{S}^{n-1} := \times_{j\in I} \hat{S}^{n-1}_j$. A strategy $s_i \in \hat{S}^n_i$ is called n-**admissible**. A strategy $s_i \in \hat{S}^\infty_i := \cap_{n=0}^{\infty} \hat{S}^n_i$ is called **iteratively admissible**.*

By finiteness of the game, it follows from standard arguments that iterated admissibility is a non-empty solution procedure: $\hat{S}^\infty := \times_{i\in I} \hat{S}^\infty_i \neq \emptyset$.

Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, i.e., $\mathbb{N}_0$ is the set of natural number including zero. The first main result of this paper is the following theorem.

[10] Specifically, this can occur when S^2_{-i} is a *strict* subset of S^1_{-i}.

[11] The best rationalization principle was originally formulated with reference to a particular form of sophisticated strategic thinking—namely forward-induction thinking—that underlies the solution concept of strong rationalizability. However, the principle is stated in abstract terms; see Battigalli (1996, p. 179). In our view, in agreement with Pierpaolo Battigalli, it makes sense to refer to the best rationalization principle also for the case of prudent rationalizability.

[12] Formally, $m(h) := \max\left\{m \in \{0, 1, ..., N\} : S^m_{-i} \cap S_{-i}(h) \neq \emptyset\right\}$, where $N \geq 0$ is the number of steps after which the sequence $\left(S^n_{-i}\right)_{n\geq 0}$ becomes constant.

Theorem 1 *For all $n \in \mathbb{N}_0$,*

$$S^n = \hat{S}^n.$$

We now consider the following procedure, which corresponds—for the special case of sequential games without unawareness—to the notion of **prudent rationalizability** in Heifetz et al. (2021) and Meier and Schipper (2024). Unlike Definition 3, the following definition requires that the players' conditional beliefs be represented by CPSs. Let $\mathrm{supp}\nu$ denote the support of a (standard) measure $\nu \in \Delta(X)$.

Definition 6 *Consider the following procedure.*

(Step 0**)** *For every $i \in I$, let $\bar{S}_i^0 := S_i$. Also, let $\bar{S}_{-i}^0 := \times_{j \in I \setminus \{i\}} \bar{S}_j^0$ and $\bar{S}^0 := \times_{i \in I} \bar{S}_i^0$.*

(Step $n > 0$**)** *For every $i \in I$ and every $s_i \in S_i$, let $s_i \in \bar{S}_i^n$ if and only if $s_i \in \bar{S}_i^{n-1}$ and there exists $\mu_i \in \Delta^{\mathscr{S}_{-i}}(S_{-i})$ such that*

1. *$s_i \in \rho_i(\mu_i)$, and*
2. *for every $h \in H$,*

$$\bar{S}_{-i}^{n-1} \cap S_{-i}(h) \neq \emptyset \Rightarrow \mathrm{supp}\mu_i(\cdot|S_{-i}(h)) = \bar{S}_{-i}^{n-1} \cap S_{-i}(h).$$

Also, let $\bar{S}_{-i}^n := \times_{j \in I \setminus \{i\}} \bar{S}_j^n$ and $\bar{S}^n := \times_{i \in I} \bar{S}_i^n$.

Finally, let $\bar{S}_i^\infty := \cap_{n=0}^\infty \bar{S}_i^n$ and $\bar{S}^\infty := \times_{i \in I} \bar{S}_i^\infty$.

Before stating the second result, we find it useful to compare the procedure described in Definition 3 to the notion of prudent rationalizability as per Definition 6. First note that Definition 3 involves an iterative reduction procedure for conditional beliefs: at the recursive step n, the set of CNPSs justifying the strategies in S_i^n is included in the set of CNPSs justifying the strategies in S_i^{n-1}. This in turn implies that $S_i^n \subseteq S_i^{n-1}$.

By way of contrast, Definition 6 involves an iterative deletion of strategies: at the recursive step, it is *required* that $\bar{S}_i^n \subseteq \bar{S}_i^{n-1}$. However, the set of CPSs (not to be confused with CNPSs!) justifying the strategies in $\bar{S}_i^n$ *may not be* included in the set of CPSs justifying the strategies in $\bar{S}_i^{n-1}$. To understand this key point, fix a player $i \in I$ and suppose that, at step $n+1$ of the procedure, $\bar{S}_{-i}^n \subset \bar{S}_{-i}^{n-1}$, where the symbol $\subset$ denotes the *strict* inclusion. Consider a strategy $s_i \in \bar{S}_i^{n+1}$ and a CPS μ_i justifying it. Moreover, consider a history h consistent with $\bar{S}_{-i}^n$, e.g., $h = \varnothing$. In such a case, μ_i must satisfy $\mathrm{supp}\mu_i(\cdot|S_{-i}(\varnothing)) = \bar{S}_{-i}^n \cap S_{-i}(\varnothing) = \bar{S}_{-i}^n$. Yet, μ_i does *not* justify s_i at step n, since $\mathrm{supp}\mu_i(\cdot|S_{-i}(\varnothing)) \subset \bar{S}_{-i}^{n-1}$.

Despite the above differences, the following result shows the equivalence between the procedures in Definition 3 and Definition 6.

Theorem 2 *For all $n \in \mathbb{N}_0$,*

$$S^n = \bar{S}^n.$$

The proof of Theorem 2 is simple: it first shows by induction the equivalence between prudent rationalizability and iterated admissibility, viz. $\bar{S}^n = \hat{S}^n$, which is a known result; with this, the statement follows from Theorem 1.

We conclude this section with two remarks. First, the result in Theorem 2 suggests that, at the recursive step of Definition 6, the "full-support condition" for a justifying CPS μ_i can be thought of as a "standard" approximation of the notion of cautious belief at history h: to capture the idea that $\bar{S}_{-i}^{n-1} \cap S_{-i}(h)$ is cautiously believed under the *standard* measure $\mu_i(\cdot|S_{-i}(h))$, we require that $\mu_i(\cdot|S_{-i}(h))$ must assign positive probability only to the profiles in $\bar{S}_{-i}^{n-1} \cap S_{-i}(h)$. Indeed, under a real-valued measure, an event is deemed infinitely more likely than another only if the latter is assigned zero probability.

Second, we point out that Theorems 1 and 2 admit analogues in terms of "behaviorally equivalent" strategies (cf. Battigalli and De Vito 2021, Section 7). To clarify, fix a player $i \in I$, and let

$$H_i(s_i) := \{h \in H : s_i \in S_i(h)\}$$

denote the set of non-terminal histories allowed by strategy s_i. Say that strategies s_i' and s_i'' are *behaviorally equivalent* if $H_i(s_i') = H_i(s_i'')$ and $s_i'(h) = s_i''(h)$ for each $h \in H_i(s_i')$. It can be shown by standard arguments—see Battigalli et al. (2024, Chapter 9)—that s_i' and s_i'' are behaviorally equivalent if and only if they are *realization equivalent*, that is, $\zeta(s_i', s_{-i}) = \zeta(s_i'', s_{-i})$ for all s_{-i}, which means that they induce the same terminal histories and are observationally indistinguishable. A **reduced strategy** is an element of the partition of S_i induced by the behavioral equivalence relation.[13]

With this, redefine the procedures $(S^n)_{n\in\mathbb{N}}$ and $\left(\bar{S}^n\right)_{n\in\mathbb{N}}$ in, respectively, Definition 3 and Definition 6 with the best-reply correspondence $\rho_i(\cdot)$ replaced by the following weaker version: for every CNPS (or CPS) μ_i,

$$\bar{\rho}_i(\mu_i) := \left\{ s_i \in S_i : \forall h \in H_i(s_i), s_i \in \arg\max_{r_i \in S_i(h)} \mathbb{E}_{\mu_i}(U_i(r_i,\cdot)|h) \right\}.$$

In words, $\bar{\rho}_i(\mu_i)$ is the set of **weak sequential best replies** to μ_i: a strategy $s_i \in \bar{\rho}_i(\mu_i)$ must be optimal only at histories that s_i allows. Consequently, $\bar{\rho}_i(\cdot)$ does not distinguish between strategies in the same equivalence class; put differently, weak sequential optimality determines reduced rather than full strategies.

Redefine also iterated admissibility—i.e., the procedure $\left(\hat{S}^n\right)_{n\in\mathbb{N}}$—in terms of reduced strategies. Going through the proofs in the Appendix, it is easy to see how the same proofs, with minor changes, show that there are analogues of Theorems 1 and 2 in terms of reduced strategies and weak sequential optimality.

Acknowledgements. I thank three anonymous referees, Pierpaolo Battigalli and Emiliano Catonini for their comments and helpful feedback. Needless to say, all mistakes are my own responsibility.

Appendix: Proofs of Theorem 1 and Theorem 2

The proofs of Theorems 1 and 2 make use of the following properties of the standard part—see Theorem 1.6.7 in Hurd and Loeb (1985). For any $x, y \in {}^*\mathbb{R}$,

P1: $\mathrm{st}(x \pm y) = \mathrm{st}(x) \pm \mathrm{st}(y)$,
P2: $\mathrm{st}(xy) = \mathrm{st}(x)\,\mathrm{st}(y)$,
P3: $\mathrm{st}(y/x) = \frac{\mathrm{st}(y)}{\mathrm{st}(x)}$ if $\mathrm{st}(x) \neq 0$,
P4: $\mathrm{st}(x) \leq \mathrm{st}(y)$ if $x \leq y$.

In the remainder of this appendix, we fix a game $G := \left\langle I, \bar{H}, (A_i, u_i)_{i\in I} \right\rangle$. The following lemma will be used in the proof of Theorem 1. As usual, for all $s_i \in S_i$ and $\nu_i \in {}^*\Delta(S_{-i})$, let

$$U_i(s_i, \nu_i) := \sum_{s_{-i} \in S_{-i}} U_i(s_i, s_{-i}) \nu_i(s_{-i}).$$

[13] Often reduced strategies are called "plans of action," suggesting that how a strategy s_i is defined at histories in $H \backslash H_i(s_i)$ is irrelevant for planning.

Lemma 1 *Fix $n \in \mathbb{N}$. Then, for every $s_i \in \hat{S}_i^n$ there exists $\nu_i \in {}^*\Delta(S_{-i})$ such that*
(i) $\nu_i(s_{-i}) > 0$ for every $s_{-i} \in S_{-i}$,
(ii) $s_i \in \operatorname{argmax}_{s_i' \in S_i} U_i(s_i', \nu_i)$, and
(iii) for every $m \in \{0, ..., n-1\}$ and for every $s_{-i} \in \hat{S}_{-i}^m$,

$$\operatorname{st}\left(\frac{\nu_i\left(S_{-i} \backslash \hat{S}_{-i}^m\right)}{\nu_i(s_{-i})}\right) = 0.$$

Proof. By Lemma 2 in Veronesi (1997) (see also Lemma E.1 in Brandenburger et al. 2008), for each $\ell = 0, ..., n-1$, there exists $\nu_\ell \in \Delta(S_{-i})$ such that $\operatorname{supp} \nu_\ell = \hat{S}_{-i}^{n-1-\ell}$ and $s_i \in \operatorname{argmax}_{s_i' \in S_i} U_i(s_i', \nu_\ell)$. Pick any $\varepsilon \in \mathrm{Hal}^+(0)$ such that $\varepsilon > 0$. A (non-standard) measure $\nu_i \in {}^*\Delta(S_{-i})$ is constructed as follows. Define, for all $s_{-i} \in S_{-i}$,

$$\eta(s_{-i}) := \nu_0(s_{-i}) + \sum_{\ell=1}^{n-1} \varepsilon^\ell \nu_\ell(s_{-i}),$$

and

$$\nu_i(s_{-i}) := \frac{\eta(s_{-i})}{\sum_{s_{-i}' \in S_{-i}} \eta\left(s_{-i}'\right)}.$$

Thus, $\nu_i(s_{-i}) > 0$ for each $s_{-i} \in S_{-i}$, and it can be easily checked that $s_i \in \operatorname{argmax}_{s_i' \in S_i} U_i(s_i', \nu_i)$. It remains to be shown that Condition (iii) holds. By definition of ν_i, it is enough to show that, for all $m \leq n-1$ and $s_{-i} \in \hat{S}_{-i}^m$,

$$\operatorname{st}\left(\frac{\sum_{s_{-i}' \in S_{-i} \backslash \hat{S}_{-i}^m} \eta\left(s_{-i}'\right)}{\eta(s_{-i})}\right) = 0.$$

To this end, fix any $m \leq n-1$. Let $s_{-i} \in \hat{S}_{-i}^m$. Notice that, if $s_{-i} \in \hat{S}_{-i}^{n-1}$, then $\operatorname{st}(\eta(s_{-i})) = \nu_0(s_{-i}) > 0$. Thus, using P1-P3,

$$\operatorname{st}\left(\frac{\sum_{s_{-i}' \in S_{-i} \backslash \hat{S}_{-i}^m} \eta\left(s_{-i}'\right)}{\eta(s_{-i})}\right) = \frac{\operatorname{st}\left(\sum_{\ell=1}^{n-1} \varepsilon^\ell \nu_\ell\left(S_{-i} \backslash \hat{S}_{-i}^m\right)\right)}{\operatorname{st}(\eta(s_{-i}))} = 0.$$

With this, consider the case when $s_{-i} \in \hat{S}_{-i}^m \backslash \hat{S}_{-i}^{n-1}$. If $s_{-i} \in \hat{S}_{-i}^{n-2}$ (which implies $m < n-1$), then $\nu_0(s_{-i}) = 0$, so that

$$\begin{aligned} \eta(s_{-i}) &= \varepsilon\left[\nu_1(s_{-i}) + \frac{\varepsilon^2}{\varepsilon}\nu_2(s_{-i}) + ... + \frac{\varepsilon^{n-1}}{\varepsilon}\nu_{n-1}(s_{-i})\right] \\ &= \varepsilon\left(\sum_{\ell=1}^{n-1} \frac{\varepsilon^\ell}{\varepsilon}\nu_\ell(s_{-i})\right). \end{aligned}$$

Hence,

$$\frac{\sum_{s_{-i}' \in S_{-i} \backslash \hat{S}_{-i}^m} \eta\left(s_{-i}'\right)}{\eta(s_{-i})} = \frac{\sum_{\ell=1}^{n-1} \varepsilon^\ell \nu_\ell\left(S_{-i} \backslash \hat{S}_{-i}^m\right)}{\varepsilon} \cdot \frac{1}{\sum_{\ell=1}^{n-1} \frac{\varepsilon^\ell}{\varepsilon}\nu_\ell(s_{-i})}.$$

It turns out that

$$\operatorname{st}\left(\frac{1}{\sum_{\ell=1}^{n-1} \frac{\varepsilon^\ell}{\varepsilon}\nu_\ell(s_{-i})}\right) = \frac{1}{\nu_1(s_{-i})},$$

and

$$\begin{aligned}\text{st}\left(\frac{\sum_{\ell=1}^{n-1}\varepsilon^{\ell}\nu_{\ell}\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)}{\varepsilon}\right) &= \text{st}\left(\frac{\sum_{\ell=2}^{n-1}\varepsilon^{\ell}\nu_{\ell}\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)}{\varepsilon}\right)\\ &= \text{st}\left(\sum_{\ell=2}^{n-1}\frac{\varepsilon^{\ell}}{\varepsilon}\nu_{\ell}\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)\right)\\ &= 0,\end{aligned}$$

where the second equality follows from the fact that $m<n-1$, which yields $\nu_1\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)=0$ (recall that $\operatorname{supp}\nu_1=\hat{S}_{-i}^{n-2}$). Hence,

$$\text{st}\left(\frac{\sum_{s'_{-i}\in S_{-i}\backslash\hat{S}_{-i}^{m}}\eta\left(s'_{-i}\right)}{\eta\left(s_{-i}\right)}\right)=0\cdot\frac{1}{\nu_1\left(s_{-i}\right)}=0.$$

The case when $s_{-i}\in\hat{S}_{-i}^{m}\backslash\hat{S}_{-i}^{n-2}$ and $s_{-i}\in\hat{S}_{-i}^{n-3}$ (which implies $m<n-2$) is similar: it turns out that $\nu_1\left(s_{-i}\right)=0$ and

$$\begin{aligned}\eta\left(s_{-i}\right) &= \varepsilon^2\nu_2\left(s_{-i}\right)+\ldots+\varepsilon^{n-1}\nu_{n-1}\left(s_{-i}\right)\\ &= \varepsilon^2\sum_{\ell=2}^{n-1}\frac{\varepsilon^{\ell}}{\varepsilon^2}\nu_{\ell}\left(s_{-i}\right).\end{aligned}$$

By noting that $m<n-2$, it is easily deduced that $\nu_1\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)=\nu_2\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)=0$. Thus,

$$\text{st}\left(\frac{\varepsilon^3\nu_2\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)+\ldots+\varepsilon^{n}\nu_{n-1}\left(S_{-i}\backslash\hat{S}_{-i}^{m}\right)}{\varepsilon^2}\right)=0$$

and

$$\text{st}\left(\frac{1}{\sum_{\ell=2}^{n-1}\frac{\varepsilon^{\ell}}{\varepsilon^2}\nu_{\ell}\left(s_{-i}\right)}\right)=\frac{1}{\nu_2\left(s_{-i}\right)},$$

which yields the result. Proceeding this way, the claim follows. ■

Proof of Theorem 1. By induction on $n\in\mathbb{N}_0$.

(*Basis step*) Immediate, since $S^0=\hat{S}^0=S$.

(*Inductive step*) Suppose that the result is true for each $m\leq n$. We show that it is true for each $m\leq n+1$.

Fix a player i and $s_i\in S_i$. Arguing by contraposition, suppose that $s_i\notin\hat{S}_i^{n+1}$. Since s_i is weakly dominated with respect to $\hat{S}^{n-1}$, there exists $\sigma_i\in\Delta(\hat{S}_i^{n+1})$ such that $U_i(\sigma_i,s_{-i})\geq U_i(s_i,s_{-i})$ for every $s_{-i}\in\hat{S}_{-i}^{n}$ and $U_i(\sigma_i,s'_{-i})>U_i(s_i,s'_{-i})$ for some $s'_{-i}\in\hat{S}_{-i}^{n}$. By the inductive hypothesis, $\hat{S}_{-i}^{n}=S_{-i}^{n}$. Pick any $\mu_i\in\cap_{l=0}^{n}\text{cSB}_i(S_{-i}^{l})$. Since every $s_{-i}\in S_{-i}=S_{-i}(\varnothing)$ is assigned strictly positive probability by $\mu_i(\cdot|S_{-i})$, it follows that, for all $s_{-i}\in S_{-i}^{n}$,

$$U_i(\sigma_i,s_{-i})\mu_i\left(s_{-i}|S_{-i}\right)-U_i(s_i,s_{-i})\mu_i\left(s_{-i}|S_{-i}\right)\geq 0$$

and, for $s'_{-i}\in S_{-i}^{n}$,

$$U_i(\sigma_i,s_{-i})\mu_i\left(s'_{-i}|S_{-i}\right)-U_i(s_i,s'_{-i})\mu_i\left(s'_{-i}|S_{-i}\right)>0.$$

Then

$$\sum_{s_{-i}\in S^n_{-i}} (U_i(\sigma_i,s_{-i})\mu_i(s_{-i}|S_{-i}) - U_i(s_i,s_{-i})\mu_i(s_{-i}|S_{-i})) > 0,$$

which implies $U_i(s_i,\mu_i(s_{-i}|S_{-i})) < U_i(\sigma_i,\mu_i(s_{-i}|S_{-i}))$ for all $s_{-i}\in S^n_{-i}$. Hence, $s_i\notin\rho_i(\mu_i)$. (For, if $s_i\in\rho_i(\mu_i)$, then $U_i(s_i,\mu_i(s_{-i}|S_{-i}))\geq U_i(\sigma_i,\mu_i(s_{-i}|S_{-i}))$ for all $\sigma_i\in\Delta(S_i)$ and $s_{-i}\in S_{-i}$.) Thus, $s_i\notin S^{n+1}_i$. This argument shows that $S^{n+1}\subseteq\hat{S}^{n+1}$.

Conversely, let $s_i\in\hat{S}^{n+1}_i$. By Lemma 1, there exists $\nu_i\in{}^*\Delta(S_{-i})$ such that

(i) $\nu_i(s_{-i})>0$ for every $s_{-i}\in S_{-i}$,

(ii) $s_i\in\operatorname{argmax}_{s'_i\in S_i}U_i(s'_i,\nu_i)$, and

(iii) for every $m\in\{0,\ldots,n\}$ and for every $s_{-i}\in\hat{S}^m_{-i}$,

$$\operatorname{st}\left(\frac{\nu_i\left(S_{-i}\backslash\hat{S}^m_{-i}\right)}{\nu_i(s_{-i})}\right)=0. \tag{6.1}$$

With this, define a CNPS $\mu_i:=(\mu_i(\cdot|S_{-i}(h)))_{h\in H}\in{}^*\Delta^{\mathscr{S}_{-i}}(S_{-i})$ by letting $\mu_i(\cdot|S_{-i}):=\nu_i$, and, for each $h\neq\varnothing$, the measure $\mu_i(\cdot|S_{-i}(h))$ is defined by conditioning. Since $s_i\in\operatorname{argmax}_{s'_i\in S_i}U_i(s'_i,\mu_i(\cdot|S_{-i}))$ and $\mu_i(S_{-i}(h)|S_{-i})>0$ for all $h\neq\varnothing$, it follows from standard results (see, e.g., Battigalli et al. 2024, Theorem 34) that $s_i\in\rho_i(\mu_i)$. It remains to be shown that, for each $m=1,\ldots,n$, the event S^m_{-i} is c-strongly believed under μ_i. By the inductive hypothesis, $S^m_{-i}=\hat{S}^m_{-i}$ for all $m=1,\ldots,n$. So, consider any S^m_{-i} $(m\leq n)$ and $h\in H$ such that $S^m_{-i}\cap S_{-i}(h)\neq\emptyset$. Then

$$\mu_i\left(\hat{S}^m_{-i}|S_{-i}(h)\right)=\frac{\nu_i\left(\hat{S}^m_{-i}\cap S_{-i}(h)\right)}{\nu_i(S_{-i}(h))}.$$

Hence, for all $s_{-i}\in S^m_{-i}\cap S_{-i}(h)$,

$$\operatorname{st}\left(\frac{\mu\left(S_{-i}\backslash\hat{S}^m_{-i}|S_{-i}(h)\right)}{\mu\left(s_{-i}|S_{-i}(h)\right)}\right)=\operatorname{st}\left(\frac{\nu_i\left(\left(S_{-i}\backslash\hat{S}^m_{-i}\right)\cap S_{-i}(h)\right)}{\nu_i(s_{-i})}\right)=0,$$

where the last equality follows from (6.1) and P4. Thus, any S^m_{-i} $(m\leq n)$ is c-strongly believed under μ_i, as required. ■

Proof of Theorem 2. We show, by induction on $n\in\mathbb{N}_0$, that $\hat{S}^n=\bar{S}^n$. With this, Theorem 1 yields the result.

(*Basis step*) Immediate, since $\hat{S}^0=\bar{S}^0=S$.

(*Inductive step*) Suppose that the result is true for each $m\leq n$. We show that it is true for each $m\leq n+1$.

Let $s_i\in\hat{S}^{n+1}_i$. Then $s_i\in\hat{S}^n_i$, and so, by the inductive hypothesis, $s_i\in\bar{S}^n_i$. So, $s_i\in\rho_i(\mu_i)$ for some CPS $\mu_i\in\Delta^{\mathscr{S}_{-i}}(S_{-i})$ such that, for every $h\in H$,

$$\bar{S}^{n-1}_{-i}\cap S_{-i}(h)\neq\emptyset\Rightarrow\operatorname{supp}\mu_i(\cdot|S_{-i}(h))=\bar{S}^{n-1}_{-i}\cap S_{-i}(h).$$

By Lemma E.1 in Brandenburger et al. (2008) (or Lemma 2 in Veronesi 1997), there exists $\nu\in\Delta(S_{-i})$ such that $\operatorname{supp}\nu=\hat{S}^n_{-i}=\bar{S}^n_{-i}$ and $s_i\in\operatorname{argmax}_{s'_i\in S_i}U_i(s'_i,\nu)$. Define a CPS $\hat{\mu}_i\in\Delta^{\mathscr{S}_{-i}}(S_{-i})$ as follows: for each $h\in H$,

$$\hat{\mu}_i(\cdot|S_{-i}(h)):=\begin{cases}\frac{\nu(\cdot\cap S_{-i}(h))}{\nu(S_{-i}(h))}, & \text{if } \bar{S}^n_{-i}\cap S_{-i}(h)\neq\emptyset,\\ \mu_i(\cdot|S_{-i}(h)), & \text{otherwise.}\end{cases}$$

It is immediate to check that $\hat{\mu}_i$ satisfies the required properties and $s_i \in \rho_i(\hat{\mu}_i)$.

For the converse, pick any $s_i \in \bar{S}_i^{n+1}$. Then, by the inductive hypothesis, $s_i \in \hat{S}_i^m$ for all $m \leq n$. It must be shown that $s_i \in \hat{S}_i^{n+1}$. Since $s_i \in \bar{S}_i^{n+1}$, there exists $\mu_i \in \Delta^{\mathscr{S}_{-i}}(S_{-i})$ such that $\text{supp}\mu_i(\cdot|S_{-i}) = \bar{S}_{-i}^n = \hat{S}_i^n$ and $s_i \in \rho_i(\mu_i)$. In particular, $s_i \in \text{argmax}_{s_i' \in S_i} U_i(s_i', \mu_i(\cdot|S_{-i}))$. By Lemma 4 in Pearce (1984), s_i is admissible with respect to $\hat{S}^n$. Hence, $s_i \in \hat{S}_i^{n+1}$. ■

References

[1] G.B. Asheim & A. Perea (2005): *Sequential and Quasi-perfect Rationalizability in Extensive Games*. Games and Economic Behavior 53, pp. 15–42, doi:10.1016/j.geb.2004.06.015.

[2] P. Battigalli (1996): *Strategic Rationality Orderings and the Best Rationalization Principle*. Games and Economic Behavior 13, pp. 178–200, doi:10.1006/game.1996.0033.

[3] P. Battigalli (1997): *On Rationalizability in Extensive Games*. Journal of Economic Theory 74, pp. 40–61, doi:10.1006/jeth.1996.2252.

[4] P. Battigalli, E. Catonini & N. De Vito (2024): *Game Theory: Analysis of Strategic Thinking*. Lecture Notes, Bocconi University.

[5] 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.

[6] P. Battigalli & M. Siniscalchi (2002): *Strong Belief and Forward Induction Reasoning*. Journal of Economic Theory 106, pp. 356–391, doi:10.1006/jeth.2001.2942.

[7] 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.

[8] L. Blume, A. Brandenburger & E. Dekel (1991): *Lexicographic Probabilities and Choice Under Uncertainty*. Econometrica 59, pp. 61–79, doi:10.2307/2938240.

[9] 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.

[10] E. Catonini & N. De Vito (2020): *Weak Belief and Permissibility*. Games and Economic Behavior 120, pp. 154–179, doi:10.1016/j.geb.2019.11.007.

[11] E. Catonini & N. De Vito (2024): *Cautious Belief and Iterated Admissibility*. Journal of Mathematical Economics 110:102918, doi:10.1016/j.jmateco.2023.102918.

[12] N. De Vito (2023): *Directed Lexicographic Rationalizability*. Economics Letters 227:111134, doi:10.1016/j.econlet.2023.111134.

[13] L. Goldblatt (1998): *Lectures on the Hyperreals. An Introduction to Nonstandard Analysis*. Springer-Verlag, doi:10.1007/978-1-4612-0615-6.

[14] J.Y. Halpern (2010): *Lexicographic Probability, Conditional Probability, and Nonstandard Probability*. Games and Economic Behavior 68, pp. 155–179, doi:10.1016/j.geb.2009.03.013.

[15] A. Heifetz, M. Meier & B. Schipper (2013): *Dynamic Unawareness and Rationalizable Behavior*. Games and Economic Behavior 81, pp. 50–68, doi:10.1016/j.geb.2013.04.003.

[16] A. Heifetz, M. Meier & B. Schipper (2021): *Prudent Rationalizability in Generalized Extensive-form Games with Unawareness*. The B.E. Journal of Theoretical Economics 21, pp. 525–556, doi:10.1515/bejte-2020-2013.

[17] A.E. Hurd & P.A. Loeb (1985): *An Introduction to Nonstandard Real Analysis*. Academic Press, doi:10.1604/9780123624406.

[18] B.S. Lee (2016): *Admissibility and Assumption*. Journal of Economic Theory 163, pp. 42–72, doi:10.1016/j.jet.2016.01.006.

[19] K.C. Lo (1999): *Nash Equilibrium without Mutual Knowledge of Rationality*. Economic Theory 14, pp. 621–633, doi:10.1007/s001990050344.

[20] W.A.J. Luxemburg (1962): *Non-standard Analysis. Lectures on A. Robinson's Theory of Infinitesimals and Infinitely Large Numbers*. California Institute of Technology, Pasadena.

[21] M. Meier & A. Perea (2024): *Reasoning about Your Own Future Mistakes*. Available at `https://epicenter.name/Perea/Papers/Future-mistakes.pdf`.

[22] M. Meier & B. Schipper (2024): *Conditional Dominance in Games with Unawareness*. Available at `https://faculty.econ.ucdavis.edu/faculty/schipper/condom.pdf`.

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

[24] D.O. Stahl (1995): *Lexicographic Rationalizability and Iterated Admissibility*. Economics Letters 47, pp. 155–159, doi:10.1016/0165-1765(94)00530-F.

[25] P. Veronesi (1997): *Lexicographic Rationality Orderings and Iterative Weak Dominance*. In P. Battigalli, A. Montesano & F. Panunzi, editors: Decisions, Games and Markets. Studies in Risk and Uncertainty, 8, Springer, pp. 125–147, doi:10.1007/978-1-4615-6337-2_5.

Complexity of Łukasiewicz Modal Probabilistic Logics

Daniil Kozhemiachenko
Aix Marseille Univ, CNRS, LIS, Marseille, France
daniil.kozhemiachenko@lis-lab.fr

Igor Sedlár*
The Czech Academy of Sciences, Institute of Computer Science, Prague, Czech Republic
sedlar@cs.cas.cz

Modal probabilistic logics provide a framework for reasoning about probability in modal contexts, involving notions such as knowledge, belief, time, and action. In this paper, we study a particular family of these logics, extending the modal Łukasiewicz many-valued logic. These logics are shown to be capable of expressing nuanced probabilistic concepts, including upper and lower probabilities. Our main contribution is a PSPACE-completeness result for two variants of the local consequence problem, providing a precise computational characterisation.

1 Introduction

Probabilistic logics have been studied and applied in various fields for decades [19, 28]. In particular, they provide formal languages that express information about probabilistic models in a compact yet unambiguous way, and they can be used to facilitate and verify reasoning about these models. The *many-valued approach* to probabilistic logic goes back to the work of Hájek et al. [14, 15, 16, 24], where 'α is probable' is taken as an *imprecise statement* whose truth value (or truth *degree*) is identified with the probability of α. This approach contrasts with the better-known logic of Fagin et al. [10], which uses formulas involving linear inequalities over probabilistic terms to represent *precise statements* about probability – statements which are either true or false. Many-valued probabilistic logics are tailored for expressing *imprecise* information about probabilistic models in terms of properties that models do not simply "have" or "not have", but which models have *to some degree*. A particular example are imprecise comparisons, indicating the degree to which two probabilities differ. Another advantage of the many-valued approach is its simpler syntax and axiomatisation. Nevertheless, as shown in [3], the many-valued approach and the linear approach are linked by entailment-preserving translations.

In computer science, artificial intelligence and economics, *modal* probabilistic logics are of particular importance. These logics formalise reasoning about probability in the presence of modal notions such as knowledge, belief, time and action [1, 2, 7, 9]. In [26], a many-valued framework for modal probabilistic logic was introduced by combining modal Łukasiewicz logic [21] with the propositional probabilistic logics of Hájek et al. Decidability of the basic many-valued modal probabilistic logic and some of its extensions was established by reduction to basic non-probabilistic modal Łukasiewicz logic, which is known to be decidable [29]. However, the reduction is not polynomial and so it does not by itself yield complexity results for many-valued modal probabilistic logics. This paper makes two contributions to this line of research. First, using the tableaux method, we show that the problem of deciding local consequence over *finitely branching* frames is PSPACE-complete even for a rather expressive extension of

*Work of the second author was supported by the grant 22-16111S of the Czech Science Foundation. We would like to thank the reviewers for their comments, which helped us improve the paper.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 350–364, doi:10.4204/EPTCS.437.28

© D. Kozhemiachenko & I. Sedlár
This work is licensed under the
Creative Commons Attribution License.

the language used in [26]. We also show that the problem of deciding local consequence over arbitrary frames is PSPACE-complete if we restrict ourselves to a specific fragment of the full language. Secondly, we argue that our framework is sufficiently expressive to formalise a variety of practically relevant concepts and scenarios, including upper and lower probabilities and their imprecise comparisons. These two contributions together establish many-valued modal logics as a viable framework for formalising the interplay between probability and modality.

The paper is structured as follows. Section 2 introduces our basic semantic structures, probabilistic frames, and provides several examples of these structures. It also discusses their relationship with similar structures proposed in the literature. Section 3 discusses a many-valued modal logic for reasoning about probabilistic frames. This logic is based on a modal extension of $\mathbf{\L\Pi\frac{1}{2}}$ [8], which is an expressive combination of Łukasiewicz and product fuzzy logics [14]. Section 4 shows how our logic can be used to represent and reason about a variety of practically relevant concepts and scenarios. Section 5 establishes our main technical results on the PSPACE-completeness of the local consequence problem (i) over finitely branching frames using the full language and (ii) over arbitrary frames using a specific fragment of the full language. Section 6 concludes the paper and lists some tasks left for the future.

2 Probabilistic frames

Recall that a *finitely additive probability measure* on a Boolean algebra X is a function $P : X \to [0,1]$ such that (i) $P(1_X) = 1$ and (ii) $P(x \vee y) = P(x) + P(y)$ for all x, y such that $x \wedge y = 0_X$. A probability measure on a finite Boolean algebra is uniquely determined by its values on the atoms of the algebra.

Definition 1 (Probabilistic frames)**.** Let A be an at most countable set. A *probabilistic* A*-frame* is a tuple $\mathfrak{F} = \langle W, R, S, \mu \rangle$ such that $\langle W, R \rangle$ is a Kripke A-frame comprising a non-empty set W and a function R from A to binary relations on W, S is a Boolean subalgebra of 2^W containing W, and $\mu = \langle \mu_w \rangle_{w \in W}$ with each $\mu_w : S \to [0,1]$ being a finitely additive probability measure on S.

The set of modal indices A can be used to represent various kinds of objects, for instance, agents or actions. As usual in modal logic, Kripke A-frames represent a collection of possible states of affairs, or *possible worlds*, connected with *accessibility relations* R_a for $a \in \mathsf{A}$. These relations can represent, for example, the qualitative uncertainty of agents (wR_au means that the information of agent a in world w is not sufficient to exclude u as a possibility) or the effects of actions (wR_au means that world u is a possible outcome of performing action a in world w). The collection S of subsets of W represents *events* (or *propositions*). Intuitively, events correspond to features of possible worlds to which probabilities will be assigned. This assignment is carried out using the collection of functions μ: for every world $w \in W$, $\mu_w(E)$ expresses the probability of event $E \in S$ *in* world w — the crucial feature of *modal* probabilistic frames is that the probabilities of events can vary from world to world, much as the truth values of propositions can vary from world to world in standard modal semantics.

Example 1. Suppose a robot is sent to retrieve an item from a warehouse. If the warehouse light is on (i.e. L is true), then the probability of retrieving the correct item (i.e. I being true) is 0.8. On the other hand, if L is false, then the probability of I being true is 0.2. There are two possible start states, which represent the situation before the robot operates and are distinguished by whether the light is on or off. There are also four possible end states, as shown in Figure 1. The start states, s_L and $s_{\neg L}$, are shown on the bottom line, and the end states $e_{L\wedge I}$, $e_{L\wedge\neg I}$, $e_{\neg L\wedge I}$ and $e_{\neg L\wedge\neg I}$ are shown on the top line. The arrows labelled with b represent the possible outcomes of the robot's actions (it is assumed that the robot cannot turn the light switch). In the state s_L, the probability of the event $E_I = \{e_{L\wedge I}, e_{\neg L\wedge I}\}$, indicated using the thick circle, is 0.8 and the probability of the event $E_{\neg I} = \{e_{L\wedge\neg I}, e_{\neg L\wedge\neg I}\}$ is 0.2. In the state $s_{\neg L}$,

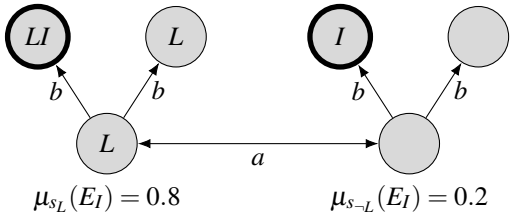


Figure 1: The robot in a warehouse example.

the probabilities are reversed: E_I has probability 0.2 and $E_{\neg I}$ has probability 0.8.[1] The arrows labelled by a represent the action of switching the warehouse light switch before the robot operates. This action clearly changes the probability of E_I – the probability of E_I increases when the light is switched on and decreases when it is switched off.

Example 2. Take the scenario from Example 1, but now suppose we have an agent a who lacks information about the warehouse lights. In standard probabilistic fashion, a's uncertainty could be modelled *quantitatively* by a subjective probability distribution on $\{s_L, s_{\neg L}\}$. However, such a distribution expresses more information than we seem to have. We know that s_L and $s_{\neg L}$ are possibilities for the agent, but we don't know how likely the agent thinks these possibilities are. Without this additional information, we can use the resources of epistemic modal logic and model the agent's uncertainty *qualitatively* by a binary equivalence relation. This is again depicted in Figure 1 if the arrow labelled by a is seen as representing agent a's qualitative uncertainty. The idea of modelling qualitative uncertainty by a *set* of probability distributions is an old one [19, Ch. 2.3].

Example 3. Well-known kinds of probabilistic relational structures also provide examples of probabilistic frames. Example 1 suggests the following construction. A *discrete-time Markov chain* is a triple $\mathbf{M} = \langle \{1,\ldots,n\}, \mathfrak{p}, \mathfrak{q} \rangle$ where $N = \{1,\ldots,n\}$ represents a set of states, $\mathfrak{p} : N \times N \to [0,1]$ is a transition probability function such that $\sum_{j=1}^{n} \mathfrak{p}(i,j) = 1$ for all $i \in N$, and $\mathfrak{q} : N \to [0,1]$ is a start probability function such that $\sum_{i=1}^{n} \mathfrak{q}(i) = 1$. For each $m \in \mathbb{N}$, a Markov chain $\mathbf{M}$ gives rise to a probability measure on the set of all paths over N of length m: $P(i_0,\ldots,i_m) = q(i_0) \cdot \prod_{j=1}^{m} \mathfrak{p}(i_{j-1}, i_j)$. Every $\mathbf{M} = \langle N = \{1,\ldots,n\}, \mathfrak{p}, \mathfrak{q} \rangle$ can be seen as a probabilistic N-frame $\mathfrak{F}_{\mathbf{M}} = \langle N_0, R, 2^S, \langle \mu_i \rangle_{i \in N_0} \rangle$ where $N_0 = N \cup \{0\}$, $R_k(i,j)$ iff $j = k$ for $j,k \in N$ and $i \in N_0$, and $\mu_i(\{j\}) = \mathfrak{p}(i,j)$ if $i,j \in N$ and $\mu_0(\{i\}) = \mathfrak{q}(i)$.

Example 4. The following is a generalisation of probabilistic frames. A *multi-agent probabilistic* A-*frame* is a structure $\langle W, R, S, \mu \rangle$ such that $\langle W, R, S \rangle$ is as before and $\mu = \langle \mu_{w,a} \rangle_{w \in W, a \in \mathsf{G}}$ for some $\mathsf{G} \subseteq \mathsf{A}$ such that $\mu_{w,a} : S \to [0,1]$ is a finitely additive probability measure. Modal indices in G represent agents and $\mu_{w,a}$ represents the subjective probability distribution of agent a in w. That is, in multi-agent frames, probability measures μ have a world index *and* an agent index while accessibility relations R are indexed by all $a \in \mathsf{A}$ as before. Probabilistic frames correspond to the special case with only one fixed agent index $a_0 \in \mathsf{A}$ which can be omitted. Multi-agent probabilistic frames where W is finite, $S = 2^W$, $\mathsf{A} = \mathsf{G}$ and R_a is an equivalence relation correspond to finite semantic knowledge-belief systems of [1], structures that extend finite Harsanyi type spaces [22] with S5-style knowledge operators. Intuitively, the relation R_a represents the knowledge state of agent a in a manner similar to Example 2. We confine the more general multi-agent framework to this example mainly for reasons of notational simplicity.

Remark 1. Probabilistic frames are similar to structures that have appeared in the literature. Aumann's *infinite* semantic knowledge-based systems [1] assume that S is closed under countable unions (σ-algebra)

[1] We assume for the sake of simplicity that, in each end state e, the only event that has non-zero probability is $\{e\}$ and that the probability of this event is 1.

and that $\mu_a(\cdot, E) : w \mapsto \mu_{w,a}(E)$ is an S-measurable mapping from W to $[0,1]$. This condition also appears in the definition of *Markov kernels*, probabilistic transition systems on infinite state spaces. The condition allows to define probabilities of events of the form "The probability of event E is at least r" for $r \in [0,1] \cap \mathbb{Q}$; see [22, 25, 31]. In this paper we focus only on probability assignments to Boolean events, and so we can use simpler structures. Fagin and Halpern [9] use similar A-frames $\langle W, R, \mathscr{P} \rangle$ with $\mathscr{P}(a,w) = \langle W_{a,w}, S_{a,w}, \mu_{a,w} \rangle$ being a probability space for $W_{a,w} \subseteq W$, $S_{a,w}$ a Boolean algebra of subsets of $W_{a,w}$ containing $W_{a,w}$, and $\mu_{a,w}$ a probability measure on $S_{a,w}$. That is, not only the probability measure but also the space of measurable events can vary from world to world.[2] Our structures are simpler.

3 Modal many-valued probabilistic logic

We have seen that probabilistic frames are a general and versatile type of semantic structure. In this section, we introduce a language based on many-valued modal logic for expressing and reasoning about properties of probabilistic frames. We build on the work of Hájek et al. [14, 15, 16, 24], who focus on variants of the well-known Łukasiewicz fuzzy logic and its combinations with product fuzzy logic. The propositional fragment of our modal language can be seen as the language of $\mathbf{\L\Pi\frac{1}{2}}$ logic, a combination of Łukasiewicz and product logic and one of the most expressive propositional fuzzy languages.

Let $\mathtt{Prop}$ be a countably infinite set of propositional variables. Let $\mathscr{L}_{\{\sim,\wedge\}}$ be the set of Boolean terms generated via the following grammar:

$$\mathscr{L}_{\{\sim,\wedge\}} \ni \alpha ::= p \in \mathtt{Prop} \mid {\sim}\alpha \mid (\alpha \wedge \alpha)$$

We use Greek letters α, β to range over $\mathscr{L}_{\{\sim,\wedge\}}$.

Definition 2. The modal probabilistic language $\mathscr{L}^{\mathsf{A}}_{\mathsf{Pr}}$ is given by

$$\mathscr{L}^{\mathsf{A}}_{\mathsf{Pr}} \ni \phi ::= \mathsf{Pr}(\alpha) \mid \neg\phi \mid (\phi \to \phi) \mid (\phi \bullet \phi) \mid (\phi \to_\Pi \phi) \mid \Box_a \phi \mid \tfrac{\mathbf{1}}{\mathbf{2}}$$

for $\alpha \in \mathscr{L}_{\{\sim,\wedge\}}$ and $a \in \mathsf{A}$. We will assume a fixed A from now on and omit explicit reference to it in most notation. In what follows, we will call formulas of the form $\mathsf{Pr}(\alpha)$ *probabilistic atoms*.

Convention 1 (Notation). Given $\phi \in \mathscr{L}_{\mathsf{Pr}}$, we use $\mathtt{Prop}(\phi)$ to denote the set of propositional variables occurring in ϕ. Furthermore, $\ell(\phi)$ denotes the length of ϕ, i.e., the number of occurrences of symbols in it. Similarly, for $\Gamma \subseteq \mathscr{L}_{\mathsf{Pr}}$, we use $\mathsf{Pr}[\Gamma]$ and $\ell[\Gamma]$ for the set of propositional variables in Γ and its length. Finally, given a binary relation R on W, we set $R(w) := \{w' \mid wRw'\}$.

Definition 3 (Modal depth)**.** Let $\phi \in \mathscr{L}_{\mathsf{Pr}}$. Its *modal depth* ($md(\phi)$) is defined inductively as follows: $md(p) = 0$, $md(\neg\phi) = md(\phi)$, $md(\phi \circ \chi) = \max(md(\phi), md(\chi))$ (with $\circ \in \{\to, \bullet, \to_\Pi\}$), $md(\Box_a\phi) = md(\phi) + 1$. Given $\Gamma \subseteq \mathscr{L}_{\mathsf{Pr}}$, the modal depth of Γ is ($md[\Gamma]$) is defined as follows: $md[\Gamma] = \max\{md(\phi) \mid \phi \in \Gamma\}$.

Definition 4 (Semantics of $\mathbf{K\L\Pi\frac{1}{2}}^{\mathsf{Pr}}$)**.** A $\mathbf{K\L\Pi\frac{1}{2}}^{\mathsf{Pr}}$*-model* is a pair $\mathfrak{M} = \langle \mathfrak{F}, E \rangle$ with $\mathfrak{F}$ a *probabilistic* frame and an *event function* or *valuation* $E : \mathtt{Prop} \to S$ extended to a homomorphism $\mathscr{L}_{\{\sim,\wedge\}} \to S$ in an obvious way. The *probabilistic interpretation induced by* $\mathfrak{M}$ is a function $\mathscr{I}_{\mathfrak{M}} : \mathscr{L}_{\mathsf{Pr}} \times W \to [0,1]$ s.t.

[2] We note that Fagin and Halpern's frames do not require that $\mu_a(\cdot, E) : w \mapsto \mu_{w,a}(E)$ be measurable; indeed, such a requirement requires a "common" space of measurable events. Instead, they use inner measures to define probabilities of events of the form "the probability of event E is at least r".

$\mathscr{I}_{\mathfrak{M}}(\frac{\mathbf{1}}{\mathbf{2}},w)=\frac{1}{2}$ and

$$\mathscr{I}_{\mathfrak{M}}(\mathsf{Pr}(\alpha),w)=\mu_w(E(\alpha)) \qquad \mathscr{I}_{\mathfrak{M}}(\phi\to\chi,w)=\min(1,1-\mathscr{I}_{\mathfrak{M}}(\phi,w)+\mathscr{I}_{\mathfrak{M}}(\chi,w))$$

$$\mathscr{I}_{\mathfrak{M}}(\phi\bullet\chi,w)=\mathscr{I}_{\mathfrak{M}}(\phi,w)\cdot\mathscr{I}_{\mathfrak{M}}(\chi,w) \qquad \mathscr{I}_{\mathfrak{M}}(\phi\to_{\Pi}\chi,w)=\begin{cases}1 & \text{if } \mathscr{I}_{\mathfrak{M}}(\phi,w)\le\mathscr{I}_{\mathfrak{M}}(\chi,w)\\ \frac{\mathscr{I}_{\mathfrak{M}}(\chi,w)}{\mathscr{I}_{\mathfrak{M}}(\phi,w)} & \text{otherwise}\end{cases}$$

$$\mathscr{I}_{\mathfrak{M}}(\neg\phi,w)=1-\mathscr{I}_{\mathfrak{M}}(\phi,w) \qquad \mathscr{I}_{\mathfrak{M}}(\Box_a\phi,w)=\inf\{\mathscr{I}_{\mathfrak{M}}(\phi,w')\mid w'\in R_a(w)\}$$

We say that $\phi\in\mathscr{L}_{\mathsf{Pr}}$ is $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid on a pointed frame* $\langle\mathfrak{F},w\rangle$ (denoted $\mathfrak{F},w\models_{\mathsf{Pr}}\phi$) if $\mathscr{I}_{\mathfrak{M}}(\phi,w)=1$ for every $E:\mathtt{Prop}\to S$; ϕ is $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid on* $\mathfrak{F}$ ($\mathfrak{F}\models_{\mathsf{Pr}}\phi$) if $\mathfrak{F},w\models_{\mathsf{Pr}}\phi$ for every $w\in\mathfrak{F}$. Given a class $\mathbb{F}$ of probabilistic frames, we say that ϕ is $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid in* $\mathbb{F}$ ($\mathbb{F}\models_{\mathsf{Pr}}\phi$) if $\mathfrak{F}\models_{\mathsf{Pr}}\phi$ for every $\mathfrak{F}\in\mathbb{F}$. Finally, ϕ *is* $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid* ($\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}\models\phi$) if it is valid on all frames, and $\Gamma\subseteq\mathscr{L}_{\mathsf{Pr}}$ *entails* χ ($\Gamma\models_{\mathsf{Pr}}\chi$) if $\mathscr{I}_{\mathfrak{M}}(\chi,w)=1$ in every $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$-model $\mathfrak{M}$ and $w\in\mathfrak{M}$ s.t. $\mathscr{I}_{\mathfrak{M}}(\phi,w)=1$ for all $\phi\in\Gamma$.

Other Łukasiewicz connectives can be defined in a standard manner:

$$\top\coloneqq\mathsf{Pr}(p)\to\mathsf{Pr}(p) \quad \triangle\phi\coloneqq\neg\phi\to_{\Pi}\bot \quad \phi\oplus\chi\coloneqq\neg\phi\to\chi \quad \phi\odot\chi\coloneqq\neg(\neg\phi\oplus\neg\chi)$$
$$\bot\coloneqq\neg\top \quad \Diamond_a\phi\coloneqq\neg\Box_a\neg\phi \quad \phi\vee\chi\coloneqq(\phi\to\chi)\to\chi \quad \phi\wedge\chi\coloneqq\neg(\neg\phi\vee\neg\chi)$$
$$\phi\leftrightarrow\chi\coloneqq(\phi\to\chi)\odot(\chi\to\phi)$$

Using Definition 4, one can obtain their semantics:

$$\mathscr{I}_{\mathfrak{M}}(\top,w)=1 \qquad \mathscr{I}_{\mathfrak{M}}(\phi\oplus\chi,w)=\min(1,\mathscr{I}_{\mathfrak{M}}(\phi,w)+\mathscr{I}_{\mathfrak{M}}(\chi,w))$$
$$\mathscr{I}_{\mathfrak{M}}(\bot,w)=0 \qquad \mathscr{I}_{\mathfrak{M}}(\phi\odot\chi,w)=\max(0,\mathscr{I}_{\mathfrak{M}}(\phi,w)+\mathscr{I}_{\mathfrak{M}}(\chi,w)-1)$$
$$\mathscr{I}_{\mathfrak{M}}(\triangle\phi,w)=\begin{cases}1 & \text{if } \mathscr{I}_{\mathfrak{M}}(\phi,w)=1\\ 0 & \text{otherwise}\end{cases} \qquad \begin{aligned}\mathscr{I}_{\mathfrak{M}}(\phi\vee\chi,w) &= \max(\mathscr{I}_{\mathfrak{M}}(\phi,w),\mathscr{I}_{\mathfrak{M}}(\chi,w))\\ \mathscr{I}_{\mathfrak{M}}(\phi\wedge\chi,w) &= \min(\mathscr{I}_{\mathfrak{M}}(\phi,w),\mathscr{I}_{\mathfrak{M}}(\chi,w))\end{aligned}$$
$$\mathscr{I}_{\mathfrak{M}}(\phi\leftrightarrow\chi,w)=1-|\mathscr{I}_{\mathfrak{M}}(\phi,w)-\mathscr{I}_{\mathfrak{M}}(\chi,w)| \qquad \mathscr{I}_{\mathfrak{M}}(\Diamond_a\phi,w)=\sup\{\mathscr{I}_{\mathfrak{M}}(\phi,w')\mid w'\in R_a(\phi)\}$$

We note that the semantics of $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$ can be equivalently reformulated if we redefine frames by allowing S to be an arbitrary Boolean algebra independent from W. We call these frames and their respective models *sample-independent* (SI-frames and SI-models).

Definition 5.

- An *SI probabilistic* A*-frame over* X (SI-frame) is a tuple $\mathfrak{F}=\langle W,R,X,\mu\rangle$ with $\langle W,R\rangle$ being a Kripke A-frame, $X\neq\varnothing$ a Boolean algebra, and $\mu=\langle\mu_w\rangle_{w\in W}$ being a tuple of finitely additive probability measures on X.
- An *SI-*$\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-model* is a tuple $\mathfrak{M}=\langle\mathfrak{F},E\rangle$ with $\mathfrak{F}$ being a sample-independent probabilistic frame and $E:\mathtt{Prop}\to X$. The notions of probabilistic interpretations induced by $\mathfrak{M}$, entailment, and validity are the same as in Definition 4.

Theorem 1. *Let* $\phi\in\mathscr{L}_{\mathsf{Pr}}$*. Then* ϕ *is* $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid iff* ϕ *is* $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$*-valid on SI-models.*

Proof. Clearly, if ϕ is $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$-valid on SI-models, then it is valid on $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$-models. For the converse, let ϕ be *not* $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$-valid on SI-models. Thus, $\mathscr{I}_{\mathfrak{M}}(\phi,w)=x<1$ for some SI-model $\mathfrak{M}=$

$\langle W,R,X,\mu,E\rangle$. If $|X| \leq |2^W|$, it is clear that $\mathfrak{M}$ is also a $\mathbf{K\Pi\frac{1}{2}}^{\mathsf{Pr}}$-model. Otherwise let $|X| = 2^{|W|+n}$. We define a $\mathbf{K\Pi\frac{1}{2}}^{\mathsf{Pr}}$ model $\mathfrak{M}' = \langle W',R',X,\mu,E\rangle$ as follows. Let $W' = W \cup \{w_1,\ldots,w_n\}$ for new states w_1, ..., w_n; let further, wR'_aw' iff $\{w,w'\} \subseteq W$ and wR_aw'. Now, using that the measures and valuations of propositional variables in $\mathfrak{M}$ and $\mathfrak{M}'$ coincide, one can show by induction on ϕ that $\mathcal{I}_{\mathfrak{M}}(\phi,w) = \mathcal{I}_{\mathfrak{M}'}(\phi,w)$ for every $w \in W$. The result follows. □

An important property of SI-models is that given $\mathfrak{M}$ and $\phi \in \mathscr{L}_{\mathsf{Pr}}$, one can assume that $X = 2^{2^{\mathtt{Prop}(\phi)}}$. Intuitively, this means that to verify whether an $\mathscr{L}_{\mathsf{Pr}}$-formula is valid, it suffices to check probability distributions over Boolean terms that can be composed of its variables.

Proposition 1. *Let $\mathfrak{M} = \langle W,R,X,\mu,E\rangle$ be an SI-model. Define $\mathfrak{M}^{\mathsf{C}}_{\phi} = \langle W,R,2^{2^{\mathtt{Prop}(\phi)}},\mu^{\mathsf{C}},E^{\mathsf{C}}\rangle$ s.t.*

$$\forall p \in \mathtt{Prop} : E^{\mathsf{C}}(p) = \{Y \mid p \in Y,\ Y \subseteq \mathtt{Prop}(\phi)\}$$
$$\forall w \in W, Y \subseteq \mathtt{Prop}(\phi) : \mu^{\mathsf{C}}_w(\{Y\}) = \mu_w\left(E\left(\bigwedge_{p\in Y} p \wedge \bigwedge_{q\notin Y} \neg q\right)\right)$$

Then $\mathcal{I}_{\mathfrak{M}}(\phi,w) = \mathcal{I}_{\mathfrak{M}^{\mathsf{C}}_{\phi}}(\phi,w)$ for every $w \in W$ and $\phi \in \mathscr{L}_{\mathsf{Pr}}$.

Proof. Induction on ϕ. For the basis case, we let w.l.o.g. $\phi = \mathsf{Pr}(\alpha)$ for some $\alpha = \bigvee\limits_{i=1}^{n} \tau_i$ in the DNF each part of which contains all variables of ϕ. Now, $\mathcal{I}_{\mathfrak{M}}(\mathsf{Pr}(\alpha),w) = x$ means that $\mu_w\left(E\left(\bigvee\limits_{i=1}^{n} \tau_i\right)\right) = x$, i.e., $\sum\limits_{i=1}^{n} \mu_w(E(\tau_i)) = x$. From here, using the fact that $Y \vDash \tau$ iff $Y = \{p \mid p \in \tau\}$, we obtain that $\mu^{\mathsf{C}}_w(E^{\mathsf{C}}(\tau_i)) = \mu_w((\tau_i))$ for every $i \in \{1,\ldots,n\}$. Thus, $\sum\limits_{i=1}^{n} \mu^{\mathsf{C}}_w(E^{\mathsf{C}}(\tau_i)) = \sum\limits_{i=1}^{n} \mu_w(E(\tau_i)) = x$, as required.

The cases of propositional connectives can be shown by a straightforward application of the induction hypothesis. The case of $\phi = \Box\chi$ follows easily since $\langle W,R\rangle$-reducts in $\mathfrak{M}$ and $\mathfrak{M}^{\mathsf{C}}_{\phi}$ coincide. □

In what follows, we will use $\mathscr{L}_{\Box}$ to denote the fragment of $\mathscr{L}_{\mathsf{Pr}}$ s.t. only the probabilistic atoms of the form $\mathsf{Pr}(p)$ with $p \in \mathtt{Prop}$ are allowed. We will also use $\mathbf{K\Pi\frac{1}{2}}$ to denote the $\mathscr{L}_{\Box}$-fragment of $\mathbf{K\Pi\frac{1}{2}}^{\mathsf{Pr}}$ and $\models$ for its entailment (satisfaction) relation. As every two distinct probabilistic atoms concern two *independent events*, it is clear that $\mathbf{K\Pi\frac{1}{2}}$-models can be built on Kripke frames, *ignoring measures*, i.e., treating each $\mathsf{Pr}(p)$ as a *propositional variable*. Namely, given a *Kripke frame* $\mathfrak{F} = \langle W,R\rangle$, and a map $v : \mathtt{Prop} \times W \to [0,1]$, a $\mathbf{K\Pi\frac{1}{2}}$*-model* is a pair $\mathfrak{M} = \langle \mathfrak{F},v\rangle$. A $\mathbf{K\Pi\frac{1}{2}}$*-interpretation* $\mathcal{I}_{\mathfrak{M}}$ *induced by* $\mathfrak{M}$ is defined as for $\mathbf{K\Pi\frac{1}{2}}^{\mathsf{Pr}}$ with the exception that $\mathcal{I}_{\mathfrak{M}}(\mathsf{Pr}(p),w) = v(p,w)$.

Proposition 2. *For finite $\Gamma \cup \{\chi\} \subseteq \mathscr{L}_{\Box}$, $\Gamma \models \chi$ iff $\Gamma \models_{\mathsf{Pr}} \chi$.*

Proof. If $\Gamma \not\models_{\mathsf{Pr}} \chi$, then take any countermodel $\mathfrak{M}$ and construct a $\mathbf{K\Pi\frac{1}{2}}$-model $\mathfrak{M}'$ by $v(p,w) := \mathcal{I}_{\mathfrak{M}}(\mathsf{Pr}(p),w)$. It is clear that $\mathcal{I}_{\mathfrak{M}}(\phi,w) = \mathcal{I}_{\mathfrak{M}'}(\phi,w)$ for all $\phi \in \mathscr{L}_{\Box}$ and so $\Gamma \not\models \chi$. Conversely, if $\Gamma \not\models \chi$, then take any countermodel $\mathfrak{M}$ and define an *SI*-model $\mathfrak{M}'$ where $X = 2^{\mathtt{Prop}[\Gamma\cup\{\chi\}]}$ and where μ_w is determined by $\mu_w(\{p\}) = v(p,w)$. Moreover, let $E(p) = \{p\}$ if p occurs in $\Gamma \cup \{\chi\}$ and let $E(p) = \varnothing$ otherwise. It is clear that $\mathcal{I}_{\mathfrak{M}}(\phi,w) = \mathcal{I}_{\mathfrak{M}'}(\phi,w)$ for all $\phi \in \mathscr{L}_{\Box}$ such that ϕ is a subformula of a formula occurring in $\Gamma \cup \{\chi\}$. Hence, by the obvious extension of Theorem 1 to local consequence, $\Gamma \not\models_{\mathsf{Pr}} \chi$. □

4 Expressivity

Note that the truth degree of an implication $\phi \to \psi$ is inversely proportional to the truncated difference between the truth degrees of ϕ and ψ.[3] Thus, $\phi \to \psi$ can be read as "The truth degree of ϕ is *not much higher* than the truth degree of ψ". If the truth degree of ϕ is less or equal to the truth degree of ψ, then the truth degree of $\phi \to \psi$ is 1. If, on the other hand, the truth degree of ϕ is higher than the truth degree of ψ, then the truth degree of $\phi \to \psi$ is 1 minus the difference. Similarly, the truth degree of $\phi \leftrightarrow \psi$ is inversely proportional to the *absolute difference* between the values of ϕ and ψ. Consequently, we can read $\phi \leftrightarrow \psi$ as "The truth degree of ϕ does not *differ much* from the truth degree of ψ".

Recall that, given a set $\mathscr{P}$ of probability measures on a Boolean algebra X, the *lower* and *upper probability* of $x \in X$, respectively, are defined as $\mathscr{P}_*(x) = \inf\{\mu(x) \mid \mu \in \mathscr{P}\}$ and $\mathscr{P}^*(x) = \sup\{\mu(x) \mid \mu \in \mathscr{P}\}$ (see [18] and [19, Ch. 2.3]). Lower and upper probabilities are useful for expressing qualitative uncertainty. For example, the interval $[\mathscr{P}_*(x), \mathscr{P}^*(x)]$ can be seen as putting lower and upper bounds on an agent's ignorance concerning the probability of x. A crucial feature of the many-valued modal framework is that lower and upper probabilities are directly expressed by formulas of the language. Given a model $\mathfrak{M}$ and a world w, the modal formula $\Box_a \mathsf{Pr}(\alpha)$ can be seen as a probabilistic term expressing the lower probability of α with respect to $\mathscr{P} = \{\mu_{w'} \mid w' \in R_a(w)\}$ since

$$\mathscr{I}_{\mathfrak{M}}(\Box_a \mathsf{Pr}(\alpha), w) = \inf\{\mathscr{I}_{\mathfrak{M}}(\mathsf{Pr}(\alpha), w') \mid w' \in R_a(w)\} = \inf\{\mu_{w'}(E(\alpha)) \mid w' \in R_a(w)\} = \mathscr{P}_*(E(\alpha)).$$

Similarly, $\Diamond_a \mathsf{Pr}(\alpha)$ expresses the upper probability of α. It follows that, on the epistemic interpretation of $a \in \mathsf{A}$, the formula $\Box_a \mathsf{Pr}(\alpha) \leftrightarrow \Diamond_a \mathsf{Pr}(\alpha)$ expresses the amount of qualitative certainty of agent a concerning α. In particular, the truth degree of $\Box_a \mathsf{Pr}(\alpha) \leftrightarrow \Diamond_a \mathsf{Pr}(\alpha)$ is inversely proportional to the size of the interval $[\mathscr{P}_*(E(\alpha)), \mathscr{P}^*(E(\alpha))]$. Let us write $\mathsf{Cert}_a(\alpha)$ instead of $\Box_a \mathsf{Pr}(\alpha) \leftrightarrow \Diamond_a \mathsf{Pr}(\alpha)$. We can use Łukasiewicz implication to express comparisons between the amounts of qualitative uncertainty of two agents. For instance, $\mathsf{Cert}_a(\alpha) \to \mathsf{Cert}_b(\alpha)$ says that a is not much more certain about α than b.[4]

Remark 2. Halpern and Pucella [18] introduce a propositional probabilistic logic suitable for reasoning about upper probabilities. A Halpern-Pucella model is, using our notation, a tuple $\langle W, S, \mathscr{P}, E\rangle$ where W is a set, S is a Boolean subalgebra of 2^W and E is a function from a set of propositional variables to S – similarly as in our models – and $\mathscr{P}$ is a set of probability measures on S. The language Halpern and Pucella use is a variant of the non-modal linear inequality language of [10] with a primitive upper probability operator. Halpern and Pucella provide a sound and complete axiomatisation of their logic and show that the corresponding satisfiability problem is NP-complete. Marchioni [27] introduces a propositional many-valued logic for reasoning about upper probabilities in semantic structures derived from Halpern-Pucella models.[5] The logic is based on Rational Łukasiewicz logic [12] and it does not contain modal operators. Marchioni provides a sound and (finitely) strongly complete axiomatisation of his logic, and he shows that the corresponding satisfiability problem is NP-complete.

Our framework can be seen as a generalisation of both [18] and [27]. Halpern-Pucella models are a KD45-like version of our K-like probabilistic models. Instead of having a fixed set of "globally" accessible states giving rise to $\mathscr{P}$, we introduce an accessibility relation which expresses the idea that each state has its own set of "locally" accessible states. This difference is well reflected in the comparison

[3] In particular, $\mathscr{I}_{\mathfrak{M}}(\phi \to \psi, w) = 1 - \max\{\mathscr{I}_{\mathfrak{M}}(\phi, w) - \mathscr{I}_{\mathfrak{M}}(\psi, w), 0\}$.

[4] A similar interpretation applies, of course, if we read $a \in \mathsf{A}$ as actions. Then $\mathsf{Cert}_a(\alpha)$ expresses the extent to which action a "fixes" the probability of α.

[5] Instead of a set of probability measures $\mathscr{P}$, Marchioni uses an upper probability measure $\pi : S \to [0,1]$, that is, a function that satisfies certain conditions that allow it to represent an upper probability given by a set of probability measures. The reader is referred to [18, 27] for details.

of the computational complexity: as shown below, deduction problems in our framework are PSPACE-complete as opposed to coNP-complete in the case of [18] and [27]. Another difference is in the language we are using. Instead of a primitive upper/lower probability operator, we use a combination $\Diamond\mathsf{Pr}/\Box\mathsf{Pr}$ of a modal operator and the probability operator Pr.[6] The same syntactic approach to representing lower probabilities is taken by Corsi et al. [6]. Their semantics is a special case of our probabilistic models with only one equivalence accessibility relation (mono-modal S5-style probabilistic models). On the other hand, Corsi et al. [6] work with a richer language that allows for the inclusion of $\Box$ within the scope of Pr. They demonstrate that their modal probabilistic framework can formalise reasoning about probability, belief functions, and lower probabilities within a unified framework. The main results of [6] are related to complete axiomatisations, and complexity is not studied.

We can also use Łukasiewicz implication to express how actions modify the probability of certain events. For example, $\mathsf{Pr}(\alpha) \to \Box_a\mathsf{Pr}(\alpha)$ says that action *a cannot decrease* the probability of α *much* — the truth degree in w is 1 if $\mu_w(E(\alpha)) \leq \inf\{\mu_{w'}(E(\alpha)) \mid w' \in R_a(w)\}$, that is, the probability of α cannot decrease at all, and it is 1 minus the difference otherwise. Similarly, $\Diamond_a\mathsf{Pr}(\alpha) \to \mathsf{Pr}(\alpha)$ says that a cannot *increase* the probability of α much. We can write $\mathsf{NoDec}_a(\alpha)$ instead of $\mathsf{Pr}(\alpha) \to \Box_a\mathsf{Pr}(\alpha)$ and, as before, use formulas of the form $\mathsf{NoDec}_a(\alpha) \to \mathsf{NoDec}_b(\alpha)$ to express comparisons between actions a and b. Similarly for $\Diamond_a\mathsf{Pr}(\alpha) \to \mathsf{Pr}(\alpha)$ written as $\mathsf{NoInc}_a(\alpha)$. We reiterate that these comparisons are not precise statements (true or false) but rather imprecise – that is, true to a degree.

It is well known that in the presence of the constant $\frac{\mathbf{1}}{\mathbf{2}}$ and the "arithmetic" operators $\bullet$ and $\to_\Pi$, one can define constants $\mathbf{q}$ for all rational numbers $q \in [0,1] \cap \mathbb{Q}$ in Łukasiewicz logic. In combination with Łukasiewicz $\to$ and $\leftrightarrow$, this allows us to express comparisons with rational thresholds. For example, the formula $\mathbf{0.7} \to \Diamond_a\mathsf{Pr}(\alpha)$ read in terms of actions says that the supremum ("best case") of the possible probabilities of α resulting from action a is not much lower than 0.7. The definable operator $\triangle$ turns imprecise statements into precise ones. For instance, $\triangle(\mathbf{0.7} \to \mathsf{Pr}(\alpha))$ has truth degree 1 if probability of α is at least 0.7 and has truth degree 0 otherwise. This means that we may define a restricted version of the "Markovian" modalities of [1, 22, 25] etc.: $L_{\mathbf{q}}(\alpha) \coloneqq \triangle(\mathbf{q} \to \mathsf{Pr}(\alpha))$ is true (has truth degree 1) in states where the probability of α is at least $q \in [0,1] \cap \mathbb{Q}$ and false (has truth degree 0) otherwise.[7] In addition, if a is interpreted epistemically, then $\Box_a L_{\mathbf{q}}(\alpha)$ expresses a crisp statement about epistemic attitudes towards probabilities — namely, "Agent a believes that the probability of α is at least q".

We observe briefly that just as in the propositional case, constants do not change the expressivity of the language. The following statement generalises [14, Lemmas 3.3.11 and 3.3.13].

Proposition 3. *Let $\Gamma \subseteq \mathscr{L}_{\mathsf{Pr}}$ be finite and n be the least common denominator of the constants in Γ. Then there is Γ' without constants s.t. Γ is satisfiable iff Γ' is and Γ' is of polynomial size w.r.t. $\max(\ell[\Gamma], n)$.*

Proof. As Γ is finite, n exists. We can express $1/n$ as follows. For fresh $\{q_1, \ldots, q_n, q\} \subseteq \mathtt{Prop}$, we set

$$\mathbf{1/n}(q) \coloneqq (\underbrace{\mathsf{Pr}(q) \oplus \ldots \oplus \mathsf{Pr}(q)}_{n-1 \text{ times}}) \leftrightarrow \neg\mathsf{Pr}(q) \qquad \mathbf{m/n}(q_m) \coloneqq \mathsf{Pr}(q_m) \leftrightarrow (\underbrace{\mathsf{Pr}(q) \oplus \ldots \oplus \mathsf{Pr}(q)}_{m \text{ times}})$$

It is clear that $\mathscr{I}(\mathbf{1/n}(q), w) = 1$ iff $\mathscr{I}_{\mathfrak{M}}(\mathsf{Pr}(q), w) = 1/n$ and $\mathscr{I}_{\mathfrak{M}}(\mathbf{m/n}(q_m), w) = 1$ iff $\mathscr{I}(\mathsf{Pr}(q_m), w) = m/n$ (and $v(q) = 1/n$). Now, given $\phi \in \mathscr{L}_{\mathsf{Pr}}$, we define $\phi^\flat$ as follows:

$$\mathsf{Pr}(\alpha)^\flat = \mathsf{Pr}(\alpha) \qquad \mathbf{m/n}^\flat = \mathsf{Pr}(q_m) \qquad (\heartsuit\phi)^\flat = \heartsuit\phi^\flat \qquad (\phi \circ \chi)^\flat = \phi^\flat \circ \chi^\flat$$
$$(\heartsuit \in \{\neg, \triangle, \Box\}, \circ \in \{\to, \to_\Pi, \bullet\})$$

[6] In addition, our syntax and semantics are both multi-modal, allowing for modelling multiple agents or actions.

[7] This is only a restricted version since $L_{\mathbf{q}}$ cannot be nested in our language.

Now let $\mathscr{d}[\Gamma] = \max\{\mathscr{d}(\phi) \mid \phi \in \Gamma\}$ and define

$$\Gamma^\flat = \{\phi^\flat \mid \phi \in \Gamma\} \cup \{\Box^i \mathbf{1}/\mathbf{n}(q) \mid i \in \{0,\ldots,\mathscr{d}[\Gamma]\}\} \cup \bigcup_{i=0}^{\mathscr{d}[\Gamma]} \{\Box^i \mathbf{m}/\mathbf{n}(q_m) \mid \mathbf{m}/\mathbf{n} \text{ occurs in } \Gamma\}$$

It is clear that $\Gamma^\flat$ is only polynomially larger than Γ when $\ell(\Gamma) \geq n$.

Now let $\mathfrak{M}$ be a $\mathbf{K\L\Pi}^{\mathsf{Pr}}_{\frac{1}{2}}$-model and $w \in \mathfrak{M}$ s.t. $\mathscr{I}_{\mathfrak{M}}(\phi, w) = 1$ for each $\phi \in \Gamma$. By Proposition 1, we may w.l.o.g. assume that $\mathfrak{M} = \langle W, R, 2^{2^{\mathtt{Prop}[\Gamma]}}, \mu, E^{\mathsf{C}} \rangle$ is an SI-model over $2^{2^{\mathtt{Prop}[\Gamma]}}$. We define an SI-model $\mathfrak{M}^\flat = \langle W, R, 2^{2^{\mathtt{Prop}[\Gamma^\flat]}}, \mu^\flat, E^{\flat\mathsf{C}} \rangle$ s.t. $\mu^\flat_w(E^{\flat\mathsf{C}}(\alpha)) = \mu_w(E^{\mathsf{C}}(\alpha))$ for each α s.t. $\mathtt{Prop}(\alpha) \cap \{q_1,\ldots,q_n,q\} = \varnothing$, $\mu^\flat_w(E(q_m)) = \frac{m}{n}$ for every q_m, and $\mu^\flat_w(E(q)) = \frac{1}{n}$. It is now clear that $\mathscr{I}_{\mathfrak{M}^\flat}(\mathtt{Prop}(q_m), w) = \frac{m}{n}$ in every $w \in \mathfrak{M}$. Thus, as values of formulas depend only on states not further than their modal depth we have $\mathscr{I}_{\mathfrak{M}}(\phi, w) = \mathscr{I}_{\mathfrak{M}^\flat}(\phi^\flat, w)$ for every $\phi \in \Gamma$. The result follows. □

In [16], $\to_\Pi$ is used to express conditional probabilities $\mathsf{Pr}(\alpha \mid \beta) \coloneqq \mathsf{Pr}(\beta) \to_\Pi \mathsf{Pr}(\alpha \wedge \beta)$. It follows from this definition that $\mathsf{Pr}(\alpha \mid \beta)$ has truth degree 1 if the probability of β is 0. A many-valued probabilistic logic with primitive formulas expressing conditional probabilities was proposed in [13]. Flaminio [11] shows that reasoning about conditional probabilities can be carried out in a probabilistic logic based on an extension of Rational Łukasiewicz logic (RŁ) with the $\triangle$ operator, where the satisfiability problem is NP-complete. Our reason to base our logic on $\mathbf{\L\Pi}\frac{1}{2}$ instead of RŁ with $\triangle$ is that the former can express additional statements about probability, such as independence statements.

We now turn to linking the expressive resources of our language to the examples discussed in Section 2. Example 1 can be formalised using $L, I, S \in \mathtt{Prop}$, expressing "light on", "item retrieved" and "start state", respectively. It is clear how the event function E should be defined on the frame depicted in Figure 1 — for instance, $E(\neg S \wedge I) = \{e_{L \wedge I}, e_{\neg L \wedge I}\}$. The formula $\mathsf{Pr}(\neg S \wedge I)$ expresses the probability of the event $E(\neg S \wedge I)$, "the robot will retrieve the correct item", and the formula $\Box_a \mathsf{Pr}(\neg S \wedge I)$ expresses the probability of this event after the execution of action a, that is, after switching the light switch. In each start state, the action a has exactly one output state — s_L is switched to $s_{\neg L}$ and vice versa. The formula $\mathsf{Pr}(\neg S \wedge I) \to \Box_a \mathsf{Pr}(\neg S \wedge I)$ says that switching the light switch cannot decrease the probability of the robot retrieving the correct item *much*. In state s_L, the truth degree of this formula is 0.4 since the truth degree of $\mathsf{Pr}(\neg S \wedge I)$ in s_L is 0.8 and in $s_{\neg L}$ it is 0.2. On the other hand, the truth degree of $\mathsf{Pr}(\neg S \wedge I) \to \Box_a \mathsf{Pr}(\neg S \wedge I)$ in $s_{\neg L}$ is 1. The formula $\mathsf{Pr}(\neg S \wedge I) \leftrightarrow \Box_a \mathsf{Pr}(\neg S \wedge I)$, saying that the probability of the robot retrieving the correct item is not *changed much* by a, has truth degree 0.4 in both start states s_L and $s_{\neg L}$.

Example 3 shows that finite discrete-time Markov chains can be seen as probabilistic frames. Take a Markov chain $\mathbf{M} = \langle N, \mathfrak{p}, \mathfrak{q} \rangle$ where $N = \{1,\ldots,n\}$, and consider an extension of $\mathfrak{F}_{\mathbf{M}}$ with an event function E such that $E(\alpha_i) = \{i\}$ for $i \in N_0$ and a tuple of Boolean formulas α_i. The probability of a path $P(i_0,\ldots,i_m) = \mathfrak{q}(i_0) \cdot \prod_{j=1}^m \mathfrak{p}(i_{j-1}, i_j)$ corresponds to the truth degree of the formula

$$\mathsf{Pr}(i_0,\ldots,i_m) \coloneqq \mathsf{Pr}(\alpha_{i_0}) \bullet \Diamond_{i_1}(\mathsf{Pr}(\alpha_{i_1}) \bullet \Diamond_{i_2}(\ldots(\Diamond_{i_{m-1}} \mathsf{Pr}(\alpha_{i_m}))\ldots))$$

in the state 0.[8] For instance, the probability of $P(1,2,3)$ is the truth degree of $\mathsf{Pr}(\alpha_1) \bullet \Diamond_1(\mathsf{Pr}(\alpha_2) \bullet \Diamond_2 \mathsf{Pr}(\alpha_3))$. As before, we can use the resources of our language to express comparisons between probabilities of various paths, e.g., $\mathsf{Pr}(1,2,3) \to \mathsf{Pr}(1,3,2)$ or $\triangle(\mathsf{Pr}(1,2,3) \to \mathsf{Pr}(1,3,2))$.

[8] Note that, for all $m > 0$, the power set of the set of all paths of length m forms a finite Boolean algebra. This justifies the notation $P(i_0,\ldots,i_m)$.

In practical applications, one very often encounters *finitely branching* frames, that is, probabilistic frames where $R_a(w)$ is finite for all $a \in \mathsf{A}$ and all $w \in W$. For example, take frames where R_a is a partial function for all $a \in \mathsf{A}$, representing deterministic actions.[9] Another obvious example are finite frames where worlds are individuated by values of a finite set of discrete random variables. We write $\models^{\mathsf{fb}}_{\mathsf{Pr}}$ and $\models^{\mathsf{fb}}$ to denote the entailment (satisfaction) relations over the class of finitely-branching frames in $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}{}^{\mathsf{Pr}}$ and $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}$, respectively.

5 Decidability and complexity

Let us now tackle the complexity of reasoning in $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}{}^{\mathsf{Pr}}$. We will consider two fundamental tasks. First, *model-checking* — given a $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}{}^{\mathsf{Pr}}$-model $\mathfrak{M}$, a state $w \in \mathfrak{M}$, and a formula ϕ, determine the truth degree of $\mathscr{I}_{\mathfrak{M}}(\phi, w)$. The second task is *finitary entailment* over finitely branching models — given a finite $\Gamma \cup \{\chi\}$, decide whether $\Gamma \models^{\mathsf{fb}}_{\mathsf{Pr}} \chi$.

As expected, model-checking can be done in polynomial time.

Theorem 2. *Given a finite model $\mathfrak{M} = \langle W, R, S, \mu, E\rangle$, $w \in W$, and $\phi \in \mathscr{L}_{\mathsf{Pr}}$, it takes polynomial time w.r.t. $|W| + |S| + \ell(\phi)$ to determine the truth degree of $\mathscr{I}_{\mathfrak{M}}(\phi, w)$.*

Proof. We proceed by induction on ϕ. For a basis case of $\phi = \mathsf{Pr}(\alpha)$, we first determine $\|\alpha\|$ which takes polynomial time w.r.t. $\ell(\phi) + |S|$. Then, using the definition of μ_w, we calculate $\mu_w(\|\alpha\|)$ which also takes polynomial time. The cases of propositional connectives can be tackled by straightforward application of the induction hypothesis. Finally, to determine $v(\Box\chi, w)$, we calculate $v(\chi, w')$ in every $w' \in R(w)$, which takes polynomial time by the induction hypothesis. Then, we take the minimal among those, which gives us the result. □

Let us now tackle the complexity of $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}{}^{\mathsf{Pr}}$-entailment over finitely-branching frames. We begin with the presentation of tableaux for $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}\mathsf{fb}}$ — the logic $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}}$ over finitely-branching frames.

Definition 6 (Constraints)**.** We fix a countable set $\mathsf{W} = \{w, w', w_0, w_1, \ldots\}$ of *state labels* and define:

- a *labelled formula* to be $w : \phi$ with $w \in \mathsf{W}$ and $\phi \in \mathscr{L}_{\mathsf{Pr}}$;
- a *formulaic constraint* as $w : \phi \nabla i$ s.t. $w : \phi$ is a labelled formula, $\nabla \in \{\leq, \geq\}$ and i a polynomial;
- a *numerical constraint* as $w : i \triangleleft j$ with $w \in \mathsf{W}$, $\triangleleft \in \{\leq, <\}$, and i and j polynomials.

Definition 7 ($\mathscr{T}(\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}\mathsf{fb}})$ — constraint tableaux for $\mathbf{K\Pi}^{\mathbf{1}}_{\mathbf{2}\mathsf{fb}}$)**.** A *constraint tableau* is a downward branching tree of *formulaic constraints*, *numerical constraints*, and *relational terms* of the form wR_aw' ($w, w' \in \mathsf{W}$). A branch can be extended by an application of a rule from Fig. 2. Given a tableau branch $\mathscr{B} = \mathscr{F} \cup \mathscr{N}$, we let $\mathsf{W}_{\mathscr{B}} = \{w \mid w \in \mathsf{W}, \text{ and } w \text{ occurs in } \mathscr{B}\}$ and set

$$\mathscr{F} = \{w_1 : \phi_1 \nabla i_1, \ldots, w_m : \phi_m \nabla i_m\} \qquad \mathscr{N} = \{w'_1 : i'_1 \nabla j'_1, \ldots, w'_n : i'_n \nabla j'_n, w''_1 : \tfrac{\mathbf{1}}{\mathbf{2}} \nabla j_1, \ldots, w''_r : \tfrac{\mathbf{1}}{\mathbf{2}} \nabla j_r\}$$

We define $\mathscr{F}^{\mathbb{I}}_w = \{x_{w:\phi} \nabla i \mid w : \phi \nabla i \in \mathscr{F}\}$ and $\mathscr{N}^{\mathbb{I}}_w = \{x \nabla y \mid w : (x \nabla y) \in \mathscr{N}\}$ for every $w \in \mathsf{W}_{\mathscr{B}}$ and set $\mathscr{B}^{\mathbb{I}}_w = \mathscr{F}^{\mathbb{I}}_w \cup \mathscr{N}^{\mathbb{I}}_w$. Furthermore, we set $\mathscr{F}^{\mathbb{I}} = \bigcup_{w \in \mathscr{B}_{\mathsf{W}}} \mathscr{F}^{\mathbb{I}}_w$, $\mathscr{N}^{\mathbb{I}} = \bigcup_{w \in \mathscr{B}_{\mathsf{W}}} \mathscr{N}^{\mathbb{I}}_w$, and $\mathscr{B}^{\mathbb{I}} = \bigcup_{w \in \mathscr{B}_{\mathsf{W}}} \mathscr{B}^{\mathbb{I}}_w$.

A tableau branch $\mathscr{B}$ is called *open* if its corresponding system of inequalities $\mathscr{B}^{\mathbb{I}}$ has a solution over $[0, 1]$ and is *closed* otherwise. An open branch is *complete* if its conclusion is also present on the branch for every premise of a rule on the branch. A tableau is closed when all its branches are closed.

[9] Finitely branching Kripke frames are used in the semantics of Strict Deterministic Propositional Dynamic Logic [20], for example. Finitely branching probabilistic frames provide semantics for a fragment of a probabilistic version of this logic.

$$\neg_{\leq}: \frac{w:\neg\phi \leq i}{w:\phi \geq 1-i} \qquad \rightarrow_{\leq}: \frac{w:\phi\rightarrow\chi \leq i}{i\geq 1 \;\Big|\; \begin{array}{c} w:\phi \geq 1-i+j \\ w:\chi \leq j \\ w:j \leq i \end{array}} \qquad \Box_{\leq}: \frac{w:\Box_a\phi \leq i}{\begin{array}{c} wR_aw' \\ w':\phi \leq i \end{array}}$$

$$\neg_{\geq}: \frac{w:\neg\phi \geq i}{w:\phi \leq 1-i} \qquad \rightarrow_{\geq}: \frac{w:\phi\rightarrow\chi \geq i}{\begin{array}{c} w:\phi \leq 1-i+j \\ w:\chi \geq j \end{array}} \qquad \Box_{\geq}: \frac{\begin{array}{c} w:\Box_a\phi \geq i \\ wR_au \end{array}}{u:\phi \geq i}$$

$$\bullet: \frac{w:\phi\bullet\chi\triangledown i}{\begin{array}{c} w:\phi = j_1 \\ w:\chi = j_2 \\ w:j_1\cdot j_2\triangledown i \end{array}} \qquad \rightarrow_{\Pi\leq}: \frac{w:\phi\rightarrow_{\Pi}\chi \leq i}{i\geq 1 \;\Big|\; \begin{array}{c} w:\phi = j_1 \\ w:\chi = j_2 \\ w:j_2/j_1 \leq i \end{array}} \qquad \rightarrow_{\Pi\geq}: \frac{w:\phi\rightarrow_{\Pi}\chi \geq i}{\begin{array}{c} w:\phi \geq j \\ w:\chi \leq j \end{array} \;\Big|\; \begin{array}{c} w:\phi = j_1 \\ w:\chi = j_2 \\ w:j_2/j_1 \geq i \end{array}}$$

Figure 2: Tableaux rules: j is a new variable; w' is fresh on the branch and u is present on the branch; $w:\psi = j$ is a shorthand for $\{w:\psi \geq j, w:\psi \leq j\}$.

Definition 8 (Realising model)**.** Let $\mathscr{B}$ be a complete open branch, $\mathscr{B}^{\mathbb{I}}$ its corresponding system of inequalities, $\mathcal{S}_{\mathscr{B}^{\mathbb{I}}}$ its solution, and given a variable $x_{w:\phi}$, let $\mathcal{S}_{\mathscr{B}^{\mathbb{I}}}(x_{w:\phi})$ denote its value in $\mathcal{S}_{\mathscr{B}^{\mathbb{I}}}$. We say that a $\mathbf{K\Pi\frac{1}{2}}$*-model* $\mathfrak{M} = \langle W, R, v\rangle$ *realises* $\mathcal{S}_{\mathscr{B}^{\mathbb{I}}}$ if $W = \{w \mid w \text{ occurs on } \mathscr{B}\}$, $R_a = \{\langle w, w'\rangle \mid wR_aw' \in \mathscr{B}\}$, and $\mathcal{I}_{\mathfrak{M}}(\phi, w) = \mathcal{S}_{\mathscr{B}^{\mathbb{I}}}(x_{w:\phi})$ for each ϕ. In particular, if $w:\phi\triangledown i \in \mathscr{B}$ and $\mathcal{I}_{\mathfrak{M}}(\phi, w) = \mathcal{S}_{\mathscr{B}^{\mathbb{I}}}(x_{w:\phi})$, we say that $\mathfrak{M}$ realises $w:\phi\triangledown i$. We say that $\mathfrak{M}$ *realises* $\mathscr{B}$ if there is some solution of $\mathscr{B}^{\mathbb{I}}$ that is realised by $\mathfrak{M}$.

Observe from the definition above that to construct a full countermodel, it suffices to have *one* complete open branch. Let us show the completeness of $\mathscr{T}(\mathbf{K\Pi\frac{1}{2}}_{\mathsf{fb}})$. The next theorem can be proved in a standard manner.

Theorem 3. *Let* $\Gamma \cup \{\chi\} \subseteq \mathscr{L}_{\mathsf{Pr}}$ *be finite. Then* $\Gamma \models^{\mathsf{fb}} \chi$ *iff the tableau beginning with* $\{w:\phi \geq 1 \mid \phi \in \Gamma\} \cup \{w:\chi \leq d, d < 1\}$ *is closed.*

Proof. For the sake of simplicity, we assume that our language is mono-modal. The proof can be carried out in the same manner when $|\mathsf{A}| \geq 2$.

For the soundness part, we observe first that closed branches do not have realising models. Now, it suffices to show that if $\mathfrak{M}$ realises the premises of a rule, it also realises the conclusion. For propositional rules, this follows straightforwardly from Definition 4. Let us now consider modal rules. Assume that $\mathfrak{M}$ realises $w:\Box\phi \leq i$. Then $\mathcal{I}(\Box\phi, w) \leq i$ and there must be some $w' \in R(w)$ s.t. $\mathcal{I}(\phi, w') \leq i$. Hence, the conclusion is also realised. The rule $\Box_{\geq}$ can be dealt with similarly.

For the completeness part, we show that every complete open branch $\mathscr{B}$ has a realising model. Let $\mathcal{S}_{\mathscr{B}^{\mathbb{I}}}$ be a solution of the system of inequalities $\mathscr{B}^{\mathbb{I}}$ that corresponds to $\mathscr{B}$. We construct the model $\mathfrak{M} = \langle W, R, v\rangle$ as follows: $W = \{w \mid w \text{ occurs on } \mathscr{B}\}$, $R = \{\langle w, w'\rangle \mid wRw' \in \mathscr{B}\}$, $v(p, w) = \mathcal{S}_{\mathscr{B}^{\mathbb{I}}}(x_{w:p})$. It remains to show that $\mathcal{I}_{\mathfrak{M}}(\chi, w) = \mathcal{S}_{\mathscr{B}^{\mathbb{I}}}(x_{w:\chi})$ for every χ occurring on $\mathscr{B}$.

We proceed by induction on χ. The base case of $\chi \in \mathtt{Prop}$ holds by construction of $\mathfrak{M}$. The propositional cases can be verified by a straightforward application of the induction hypothesis. Let us now consider the case of $\chi = \Box\psi$. Assume that $w:\Box\psi \leq i \in \mathscr{B}$. As $\mathscr{B}$ is complete, $\{wRw', w':\psi \leq i\} \subseteq \mathscr{B}$. By the induction hypothesis, $\mathfrak{M}$ realises $w':\psi \leq i$. Thus, there is $w' \in R(w)$ s.t. $\mathcal{I}_{\mathfrak{M}}(\psi, w') \leq i$. It follows that $\mathcal{I}_{\mathfrak{M}}(\Box\psi, w) \leq i$, as required. Now assume that $w:\Box\psi \geq i \in \mathscr{B}$. If there is no u s.t. $wRu \in \mathscr{B}$,

it means that $R(w) = \varnothing$, whence $\mathcal{I}_{\mathfrak{M}}(\Box\psi, w) = 1$. Thus, $w : \Box\psi \geq i$ is realised. If there is some u s.t. $wRu \in \mathscr{B}$, then for each such u, $u : \psi \geq i \in \mathscr{B}$. Again, by the induction hypothesis, all these constraints are realised, whence $\mathcal{I}_{\mathfrak{M}}(\Box\psi, w) \leq i$, as required. □

We are finally ready to establish the computational complexity of $\models^{\mathsf{fb}}_{\mathsf{Pr}}$. We adapt the technique from [17] and combine it with a standard decision algorithm for **K** from [4].

Theorem 4. *Let $\Gamma \cup \{\chi\} \subseteq \mathscr{L}_{\mathsf{Pr}}$ be finite. Then it is* PSPACE*-complete to decide whether $\Gamma \models^{\mathsf{fb}}_{\mathsf{Pr}} \chi$.*

Proof. Let us begin with the PSPACE-hardness. We provide a reduction from the validity in **K**, which is known to be PSPACE-complete. Namely, let ϕ be a formula over $\{\neg, \to, \Box\}$ and let further, $\phi^{\triangle}$ be the result of replacing each variable p occurring in ϕ with $\triangle\mathsf{Pr}(p)$. Now observe that $\triangle$-formulas have values in $\{0,1\}$ and $\mathbf{K\Pi}\frac{1}{2}^{\mathsf{Pr}}$-connectives behave classically on $\{0,1\}$. As **K** is complete over finitely branching frames, ϕ is **K**-valid iff $\phi^{\triangle}$ is $\mathbf{K\Pi}\frac{1}{2}^{\mathsf{Pr}}$-valid on finitely branching frames.

Let us now provide a PSPACE decision procedure. We assume that $\Xi = \Gamma \cup \{\chi\}$, that there are r different probabilistic atoms in Ξ, and $\mathtt{Prop}[\Xi] = \{p_1, \ldots, p_m\}$. Observe that tableaux for finite sets of formulas terminate since rules have the branching factor of at most 2 and decompose formulas. Consider the following set: $\mathscr{T}(\Xi) = \{w^0_1 : \chi \leq c, c < 1\} \cup \{w^0_1 : \phi \geq 1 \mid \phi \in \Gamma\}$. By [5, Theorem 3.3], it is in PSPACE to determine whether the given branch is closed as $\mathscr{B}^{\mathbb{I}}_w$'s are systems of *polynomial* inequalities over $\mathbb{R}$.

The algorithm works in two stages. First, we build a $\mathbf{K\Pi}\frac{1}{2}$-tableau in a depth-first manner. Second, we recall that by Proposition 1, we can assume that all α's are evaluated on $2^{2^{\mathtt{Prop}(\phi)}}$. At this stage, we check that the values of modal atoms are coherent as a measure on $S = 2^{2^{\mathtt{Prop}[\Xi]}}$, i.e., that there is a probability measure μ s.t. $\mu_w(E(\alpha)) = x_{w:\mathsf{Pr}(\alpha)}$ for every modal atom $\mathsf{Pr}(\alpha)$ and state label w in Ξ.

We begin by applying propositional rules to $\mathscr{T}(\Xi)$, guessing the branch if needed and checking whether the resulting system of inequalities has solutions. Once there are no formulas in w^0_1 with a propositional principal connective, we pick a constraint $w^0_1 : \Box\psi \leq i$ and apply the $\Box_{\leq}$ rule to it. This produces a new state $w^1_{1,1}$[10] and a relational term $w^0_1 R w^1_{1,1}$. Now we can apply rules $\Box_{\geq}$ using this new state. In this new state, we apply propositional rules until no formulas with principal propositional connectives remain. We proceed this way until we reach a state, say $w^n_{1,1}$ generated from $w^{n-1}_{1,1}$, which does not contain modal formulas. Observe that $n \in \mathscr{O}(\ell[\Xi])$ because the length of the branch in the model is bounded by the maximal number of nested modalities in a formula of Ξ.

Once formulas in $w^n_{1,1}$ are decomposed into modal atoms, we have the following system of polynomial inequalities that corresponds to the current branch $\mathscr{B}$ and state $w^n_{1,1}$ (ψ's are formulas in $w^n_{1,1}$):

$$x_{w^n_{1,1}:\psi_1} \nabla l'_1 \quad \ldots \quad x_{w^n_{1,1}:\psi_k} \nabla l'_k \quad x_{w^n_{1,1}:\mathsf{Pr}(\alpha_1)} \nabla l_1 \quad \ldots \quad x_{w^n_{1,1}:\mathsf{Pr}(\alpha_r)} \nabla l_n \quad j_1 \lhd j'_1 \quad \ldots \quad j_{r'} \lhd j'_{r'} \quad (\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$$

We now need to check that the values of $x_{w^n_{1,1}:\mathsf{Pr}(\alpha)}$'s are coherent as a probability measure. To do this, we consider the following system of *linear* inequalities for every $i \in \{1, \ldots, r\}$:

$$\sum_{e \in \{0,1\}^m} u_e = 1 \qquad \sum_{e \in \{0,1\}^m} a_{i,e} \cdot u_e = x_{w^n_{1,1}:\mathsf{Pr}(\alpha_i)} \qquad (\exp \mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$$

Here, the e indices represent the subsets of $\mathtt{Prop}[\Xi]$, i.e., the atoms of $2^{2^{\mathtt{Prop}[\Xi]}}$: $a_{i,e} = 1$ when $e \models \alpha_i$ (this takes polynomial time to establish), and the value of u_e is the measure assignment to e. For example, if

[10] Here, the upper index denotes the 'depth' of the state on the branch; the lower left index stands for the position of the state on the level; the lower right index identifies the position of the predecessor on its level.

there are only two variables: p and q, then u_{01} corresponds to the measure of $\{q\}$ and u_{11} to the measure of $\{p,q\}$. Thus, $(\exp\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ has solutions iff the values of $x_{w^n_{1,1}:\mathsf{Pr}(\alpha_i)}$'s are coherent as a measure.

Now observe that u_e's do not occur in $(\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ and that $(\exp\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ has only $r+1$ lines. Thus, even though there are $2^{|\mathtt{Prop}[\Xi]|}$ variables u_e, we can guess a list L of $r+1$ words e over $\{0,1\}$ using [10, Lemma 2.5] and let $u_e = 0$ for each $e \notin L$. This produces the following system of inequalities.

$$\sum_{e\in L} u_e = 1 \qquad\qquad \sum_{e\in L} a_{i,e}\cdot u_e = x_{w^n_{1,1}:\mathsf{Pr}(\alpha_i)} \qquad\qquad (\mathrm{poly}\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$$

Clearly, $(\mathrm{poly}\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ is of polynomial size w.r.t. $\ell[\Xi]$ and has solutions iff $(\exp\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ does. We can now treat $(\mathrm{poly}\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ and $(\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ as quantifier-free formulas of the real field theory and consider their conjunction. By [5, Theorem 3.3], it is in PSPACE to check whether it is satisfiable over $[0,1]$. If it is *unsatisfiable*, there is no solution to $(\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$ that makes the values of $x_{w^n_{1,1}:\mathsf{Pr}(\alpha)}$'s coherent with a measure and the values of $x_{w^n_{1,1}:\psi}$'s with the semantics of $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$. In this case, we close $\mathscr{B}$. Otherwise, we delete $(\mathscr{B}^{\mathbb{I}}_{w^n_{1,1}})$, mark the constraint $w^{n-1}_{1,1} : \Box\tau \le j$ that produced $w^n_{1,1}$ as 'safe', and delete the term $w^{n-1}_{1,1} R w^n_{1,1}$. We then pick the next constraint in $w^{n-1}_{1,1}$ that can produce a new state and repeat the process. Once all constraints of the form $w^{n-1}_{1,1} : \Box\tau' \le j'$ are marked safe, we guess $\mathrm{poly}\mathscr{B}^{\mathbb{I}}_{w^{n-1}_{1,1}}$ and consider its conjunction with $\mathscr{B}^{\mathbb{I}}_{w^{n-1}_{1,1}}$. Note, furthermore, that if all probabilistic atoms occurring in $w^{n-1}_{1,1}$ are in the scope of $\Box$, it suffices to check that $\mathscr{B}^{\mathbb{I}}_{w^{n-1}_{1,1}}$ has solutions. We repeat the process until either w^0_1 is marked safe (in which case, $\Gamma \not\models^{\mathsf{fb}}_{\mathsf{Pr}} \chi$) or all branches of our tableau are closed (whence, $\Gamma \models^{\mathsf{fb}}_{\mathsf{Pr}} \chi$).

Finally, recall that the depth of the model is bounded from above by $\mathscr{d}[\Xi]+1$; moreover, each state w contains only $\mathscr{B}^{\mathbb{I}}_w$ and $\mathrm{poly}\mathscr{B}^{\mathbb{I}}_w$, whose sizes are polynomial w.r.t. $\ell[\Xi]$. Thus, we need polynomial space to execute the procedure. The result follows. □

We finish the section with two brief observations. First, in Example 4, we considered multi-agent probabilistic frames. One can readily see that our language can be straightforwardly expanded to incorporate such frames by allowing probabilistic atoms of the form $\mathsf{Pr}_a(\alpha)$ with $a \in \mathsf{A}$. The proof of PSPACE-completeness of the resulting logic can also be adapted expectedly. Second, we were dealing with the complexity of reasoning in $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$ over *finitely-branching* frames. Note, however, that we do not need the restriction to finitely branching frames if we consider the fragment $\mathscr{L}_{\mathsf{add}}$ of $\mathscr{L}_{\mathsf{Pr}}$ that contains all rational constants but *does not* contain $\bullet$ and $\to_\Pi$. We use $\mathbf{KL}^{\mathsf{Pr}}_{\mathbb{Q}}$ to denote the $\mathscr{L}_{\mathsf{add}}$-fragment of $\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}^{\mathsf{Pr}}$.

Theorem 5. *Let $\Gamma\cup\{\chi\} \subseteq \mathscr{L}_{\mathsf{add}}$ be finite. Then, it is* PSPACE*-complete to decide whether $\Gamma \models_{\mathsf{Pr}} \chi$.*

Proof sketch. By Proposition 3, we can eliminate rational constants from $\Gamma\cup\{\chi\}$ without the use of $\bullet$ and $\to_\Pi$. Thus, by [30, Lemma 4.6], the resulting logic is *complete w.r.t. witnessed models*. Hence, the $\mathscr{T}(\mathbf{K\Pi}\frac{\mathbf{1}}{\mathbf{2}}_{\mathsf{fb}})$-rules are sound. Now we use the same procedure as in the proof of Theorem 4. □

6 Conclusion

Building on the work of Hájek et al. [14, 15, 16, 24], this paper investigated many-valued logics for reasoning about probability in the presence of modal notions such as knowledge, belief and action. Our main technical result, extending the contributions of [26], is that the problem of deciding local

consequence over finitely branching frames is PSPACE-complete for a rather expressive logic. Over arbitrary frames, it is PSPACE-complete for a specific fragment of the full logic.

Several interesting research problems remain to be tackled in the future. First, we would like to extend our complexity result for $\mathbf{KL\Pi\frac{1}{2}}^{\mathsf{Pr}}$ beyond finitely branching frames and the result for $\mathbf{KL}_{\mathbb{Q}}^{\mathsf{Pr}}$ to stronger fragments. Interestingly enough, even for some fragments weaker than $\mathbf{KL\Pi\frac{1}{2}}^{\mathsf{Pr}}$, this would require solving some open problems in many-valued logic. For example, $\mathbf{KPL}^{\mathsf{Pr}}$ (the fragment without $\frac{1}{2}$ and $\to_{\Pi}$) is PSPACE-complete if first-order **PŁ** [23] is complete with respect to witnessed models. We would also like to extend our results to modal logics stronger than **K**. Second, we would like to compare the expressivity of our framework with the corresponding fragment of [9], especially when it comes to expressing statements about lower and upper probabilities.

References

[1] R.J. Aumann (1999): *Interactive epistemology II: Probability*. International Journal of Game Theory 28(3), pp. 301–314, doi:10.1007/s001820050112.

[2] F. Bacchus, J.Y. Halpern & H.J. Levesque (1999): *Reasoning about noisy sensors and effectors in the situation calculus*. Artificial Intelligence 111(1-2), pp. 171–208, doi:10.1016/S0004-3702(99)00031-4.

[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] P. Blackburn, M. de Rijke & Y. Venema (2010): *Modal logic*, 4. print. with corr. edition. Cambridge tracts in theoretical computer science 53, Cambridge University Press.

[5] J. Canny (1988): *Some algebraic and geometric computations in PSPACE*. In: Proceedings of the twentieth annual ACM symposium on Theory of computing — STOC '88, ACM Press, New York, New York, USA, pp. 460–467, doi:10.1145/62212.62257.

[6] E.A. Corsi, T. Flaminio, L. Godo & H. Hosni (2023): *A modal logic for uncertainty: a completeness theorem*. In E. Miranda, I. Montes, E. Quaeghebeur & B. Vantaggi, editors: Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, Proceedings of Machine Learning Research 215, PMLR, pp. 119–129.

[7] D. Doder & A. Perović (2020): *Probabilistic Temporal Logics*. In Z. Ognjanović, editor: Probabilistic Extensions of Various Logical Systems, Springer International Publishing, pp. 71–108, doi:10.1007/978-3-030-52954-3_3.

[8] F. Esteva, L. Godo & F. Montagna (2001): *The LΠ and L$\Pi\frac{1}{2}$ logics: two complete fuzzy systems joining Łukasiewicz and Product Logics*. Archive for Mathematical Logic 40(1), pp. 39–67, doi:10.1007/s001530050173.

[9] R. Fagin & J.Y. Halpern (1994): *Reasoning about knowledge and probability*. Journal of the ACM 41(2), pp. 340–367, doi:10.1145/174652.174658.

[10] R. Fagin, J.Y. Halpern & N. Megiddo (1990): *A logic for reasoning about probabilities*. Information and computation 87(1–2), pp. 78–128, doi:10.1016/0890-5401(90)90060-U.

[11] Tommaso Flaminio (2007): *NP-containment for the coherence test of assessments of conditional probability: a fuzzy logical approach*. Archive for Mathematical Logic 46(3–4), pp. 301–319, doi:10.1007/s00153-007-0045-3.

[12] Brunella Gerla (2001): *Rational Łukasiewicz logic and DMV-algebras*. Neural Networks World 11, pp. 579–584.

[13] L. Godo & E. Marchioni (2006): *Coherent Conditional Probability in a Fuzzy Logic Setting*. Logic Journal of IGPL 14(3), pp. 457–481, doi:10.1093/jigpal/jzl019.

[14] P. Hájek (1998): *Metamathematics of Fuzzy Logic*. Trends in Logic 4, Springer, Dordrecht, doi:10.1007/978-94-011-5300-3.

[15] P. Hájek, L. Godo & F. Esteva (1995): *Fuzzy logic and probability*. In P. Besnard & S. Hanks, editors: *UAI'95: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, pp. 237–244, doi:10.5555/2074158.2074185.

[16] P. Hájek, L. Godo & F. Esteva (2000): *Reasoning about probability using fuzzy logic*. Neural Network World 10(5), pp. 811–824.

[17] P. Hájek & S. Tulipani (2001): *Complexity of fuzzy probability logics*. Fundamenta Informaticae 45(3), pp. 207–213, doi:10.3233/FUN-2001-45304.

[18] J. Y. Halpern & R. Pucella (2002): *A Logic for Reasoning about Upper Probabilities*. Journal of Artificial Intelligence Research 17, pp. 57–81, doi:10.1613/jair.985.

[19] J.Y. Halpern (2017): *Reasoning about uncertainty*, 2nd edition. The MIT Press, doi:10.7551/mitpress/10951.001.0001.

[20] J.Y. Halpern & J.H. Reif (1983): *The propositional dynamic logic of deterministic, well-structured programs*. Theoretical Computer Science 27(1–2), pp. 127–165, doi:10.1016/0304-3975(83)90097-x.

[21] G. Hansoul & B. Teheux (2013): *Extending Łukasiewicz Logics with a Modality: Algebraic Approach to Relational Semantics*. Studia Logica 101(3), pp. 505–545, doi:10.1007/s11225-012-9396-9.

[22] A. Heifetz & P. Mongin (2001): *Probability Logic for Type Spaces*. Games and Economic Behavior 35(1–2), pp. 31–53, doi:10.1006/game.1999.0788.

[23] R. Horčík & P. Cintula (2004): *Product Łukasiewicz Logic*. Archive for Mathematical Logic 43(4), pp. 477–503, doi:10.1007/s00153-004-0214-6.

[24] P. Hájek & D. Harmancová (1995): *Medical fuzzy expert systems and reasoning about beliefs*. In: *Artificial Intelligence in Medicine. AIME 1995*, Springer Berlin Heidelberg, pp. 403–404, doi:10.1007/3-540-60025-6_159.

[25] D. Kozen, R. Mardare & P. Panangaden (2013): *Strong Completeness for Markovian Logics*. In K. Chatterjee & J. Sgall, editors: *Mathematical Foundations of Computer Science 2013*, Lecture Notes in Computer Science 8087, Springer, Berlin, Heidelberg, pp. 655–666, doi:10.1007/978-3-642-40313-2_58.

[26] O. Majer & I. Sedlár (2025): *On Many-Valued Modal Probabilistic Logics*. In: *IEEE International Symposium on Multiple-Valued Logic (ISMVL 2025)*, pp. 26–31, doi:10.1109/ISMVL64713.2025.00014.

[27] E. Marchioni (2008): *Representing Upper Probability Measures over Rational Lukasiewicz Logic*. Mathware and Soft Computing 15(2), pp. 159–173. Available at http://eudml.org/doc/42219.

[28] Z. Ognjanović, M. Rašković & Z. Marković (2016): *Probability Logics: Probability-Based Formalization of Uncertain Reasoning*. Springer International Publishing, Cham, doi:10.1007/978-3-319-47012-2.

[29] A. Vidal (2021): *On transitive modal many-valued logics*. Fuzzy Sets and Systems 407, pp. 97–114, doi:10.1016/j.fss.2020.01.011.

[30] A. Vidal (2022): *Undecidability and non-axiomatizability of modal many-valued logics*. The Journal of Symbolic Logic 87(4), pp. 1576–1605, doi:10.1017/jsl.2022.32.

[31] C. Zhou (2007): *Complete Deductive Systems for Probability Logic with Application to Harsanyi Type Spaces*. PhD Thesis, Indiana University.

Bifurcation Logic: Separation Through Ordering

Didier Galmiche
Université de Lorraine, CNRS, LORIA
F-54000 Nancy, France
didier.galmiche@loria.fr

Timo Lang
University College London
London, UK
timo.lang@ucl.ac.uk

Daniel Méry
Université de Lorraine, CNRS, LORIA
F-54000 Nancy, France
daniel.mery@loria.fr

David Pym
UCL and Institute of Philosophy
University of London, UK
d.pym@ucl.ac.uk, david.pym@sas.ac.uk

We introduce Bifurcation Logic, *BL*, which combines a basic classical modality with separating conjunction, $*$, together with its naturally associated multiplicative implication, $-\!\!*$, that is defined using the modal ordering. Specifically, a formula $\phi_1 * \phi_2$ is true at a world w if and only if each ϕ_i holds at worlds w_i that are each above w, on separate branches of the order, and have no common upper bound. We provide a labelled tableaux calculus for *BL* and establish soundness and completeness relative to its relational semantics. The standard finite model property fails for *BL*. However, we show that, in the absence of $-\!\!*$, but in the presence of $*$, every model has an equivalent finite representation and that this is sufficient to obtain decidability. We illustrate the use of *BL* through an example of modelling multi-agent access control that is quite generic in its form, suggesting many applications.

1 Introduction

We introduce Bifurcation Logic, *BL*, which combines a basic classical modality with separating conjunction, $*$, together with its naturally associated multiplicative implication, $-\!\!*$, that is defined using the modal ordering. We provide a relational semantics and a labelled tableaux calculus for *BL* and establish soundness and completeness relative to *BL*'s relational semantics. The standard finite model property fails for *BL*. However, we show that, in the absence of $-\!\!*$, but in the presence of $*$, every model has a finite representation and that this is sufficient to obtain decidability. We illustrate the use of *BL* in logical modelling through an example of multi-agent access control.

The key property of interest in *BL* is the semantics of its multiplicatives, the 'separating' $*$ and $-\!\!*$, which is given in terms of the ordering that is used to define the classical modality. This stands in contrast to the set-up in, say, bunched implications (*BI* — e.g., [16, 18, 12]) in which specific relational structure is used for their definition — see [13] for a through discussion of *BI*'s semantics. Kamide's account of Kripke semantics for modal substructural logics [15] also employs a binary operation on worlds to give a treatment of the multiplicative conjunction that is similar to that of *BI*'s elementary semantics (e.g., [16, 18, 12, 13]). Galmiche, Kimmel, and Pym [10] consider an epistemic modal extension of boolean *BI* in which the semantics of the multiplicative conjunction employs a monoidal product on worlds. Došen [7] considers a range of issues in the relationship between modal and substructural logics from the perspective of translations between proof systems, and Ono [17] has also considered the proof theory of modal and substructural logics.

The basic idea in *BL* is that a formula $\phi_1 * \phi_2$ is true at a world w if and only if each ϕ_i holds at worlds w_i that are each above w, on separate branches of the order, and have no common upper bound — that is, they are bifurcated. Consequently, the semantics of the multiplicative implication has the property that

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 365–378, doi:10.4204/EPTCS.437.29

the implicational formula $\phi \multimap \psi$ and its subformula ϕ are required to hold at bifurcated worlds above the world at which ψ holds. This use of this feature is illustrated in a substantive modelling example given in Section 3. The semantics of the classical connectives and modality is standard.

In Section 2, we introduce the language of Bifurcation Logic and its models, based on frames with a ternary relation structure for the bifurcation semantics. In Section 3, we give an extended, quite generic — i.e., evidently mappable to other settings — example of the use of *BL* in modelling access control, suggesting wider application in knowledge representation and reasoning. This example, albeit somewhat idealized, illustrates the interaction between the classical modality and the multiplicative connectives, especially the somewhat unusual semantic form of the implication, in a simple and direct way. We also discuss some related work in this section.

In Section 4, we give a system of labelled tableaux for *BL*. The form of the calculus follows the pattern established in, for example, [12, 10] and allows, in Section 5, soundness and completeness results to be established (cf. [12, 10]). The proofs are provided in the appendix.

While the usual finite model property fails for *BL*, a modified form of it does hold. Section 6 explains, through a counterexample, why the standard form fails and introduces a modified form through the concept of 'model with back links'. Intuitively, *BL* is a logic with the subformula property — in the sense evident from the tableaux system — and is about paths in finitely branching trees. Back links describe how the paths go back to already-seen configurations (see Section 6 for a formal explanation). Using this modified finite model property, the decidability of *BL* is obtained.

Finally, before proceeding to our formal development, we consider a few interesting outstanding questions, among many others, for further work:

- we would aim to give a (Hilbert-type) axiomatization of *BL*;
- we would explore natural deduction and sequent calculus presentations of *BL*;
- we would seek to establish the complexity of deciding *BL*;
- we would seek to explore the addition to BL of multi-agent and epistemic modalities, as well as quantifiers, so extending its potential as a modelling tool (cf. [10], for example).

The question of whether there is an interesting intuitionistic version of *BL* seems quite challenging, as it would combine the difficulties of intuitionistic modal logics [21] with the need to handle the multiplicatives in a coherent way.

2 Bifurcation Logic

In this section, we introduce Bifurcation Logic, *BL*, by giving a definition in terms of ternary relational semantics in the style of Routley-Meyer [20], with some similarity to the work of Fuhrmann and Mares [9].

Definition 1 (Language)**.** *Let P be a countable set of propositional letters. The formulae of BL, the set of which is denoted by* Φ*, are given by the following grammar:*

$$\phi ::= p\,(\in P) \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi \mid \phi * \phi \mid \phi \multimap \phi$$

The connectives $\vee$*,* $\rightarrow$*,* $\leftrightarrow$*,* $\Diamond$*, and the units* $\top$*,* $\bot$ *are defined as follows:* $\phi \vee \psi = \neg(\neg\phi \wedge \neg\psi)$*,* $\phi \rightarrow \psi = \neg\phi \vee \psi$*,* $\phi \leftrightarrow \psi = (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$*,* $\top = \phi \vee \neg\phi$*,* $\bot = \phi \wedge \neg\phi$*,* $\Diamond\phi = \neg\Box\neg\phi$*.* □

To minimize the use of parentheses; we use the following strict order of precedence (with right associativity): $\Box, \Diamond, \neg > * > \wedge, \vee > \multimap > \rightarrow > \leftrightarrow$.

Definition 2 (Tree)**.** *Let $(W,\leq)$ be a partial order. Two elements $w_1,w_2 \in W$ are* separated (or disjoint)*, denoted $w_1 \perp w_2$, if neither $w_1 \leq w_2$ nor $w_2 \leq w_1$. We call $(W,\leq)$* rooted *if there exists $w \in W$ such that $w \leq w'$ for all $w' \in W$. $(W,\leq)$ has the* persistent separation property *if it satisfies the following condition:*

$$(P) \quad \textit{for all } w_1,w_2,w_1' \in W, \textit{ if } w_1 \perp w_2 \textit{ and } w_1 \leq w_1', \textit{ then } w_1' \perp w_2.$$

$(W,\leq)$ is called a tree *if it is rooted and has the persistent separation property.* □

Definition 3 (Frame)**.** *A BL* frame *is a structure $\mathscr{F} = (W,\leq,R)$, where $(W,\leq)$ is a tree of elements called* worlds *and R is the ternary relation on worlds defined as follows:*

$$(B) \quad \textit{for all } w,w_1,w_2 \in W, \; R(w,w_1,w_2) \textit{ iff } w \leq w_1, w \leq w_2 \textit{ and } w_1 \perp w_2.$$

That is, w_1 and w_2 belong to distinct futures of w. □

Definition 4 (Model)**.** *A BL* model *is a triple $\mathscr{M} = (\mathscr{F},V,\Vdash)$, where $\mathscr{F}$ is a BL frame and V is a* valuation function *from W to $\wp(P)$. The satisfaction relation $\Vdash$ is inductively defined as the smallest relation on $W \times \Phi$ such that*

$$\begin{array}{rcl}
\mathscr{M},w \Vdash p & \textit{iff} & p \in V(w) \\
\mathscr{M},w \Vdash \neg\phi & \textit{iff} & \mathscr{M},w \nVdash \phi \\
\mathscr{M},w \Vdash \phi \wedge \psi & \textit{iff} & \mathscr{M},w \Vdash \phi \textit{ and } \mathscr{M},w \Vdash \psi \\
\mathscr{M},w \Vdash \Box\phi & \textit{iff} & \textit{for all } w' \in W, \textit{ if } w \leq w' \textit{ then } \mathscr{M},w' \Vdash \phi \\
\mathscr{M},w \Vdash \phi * \psi & \textit{iff} & \textit{for some } w_1,w_2 \in W \textit{ such that } R(w,w_1,w_2), \\
 & & \mathscr{M},w_1 \Vdash \phi \textit{ and } \mathscr{M},w_2 \Vdash \psi \\
\mathscr{M},w \Vdash \phi \mathrel{-\!\!*} \psi & \textit{iff} & \textit{for all } w_1,w_2 \in W \textit{ such that } R(w_2,w,w_1), \\
 & & \textit{if } \mathscr{M},w_1 \Vdash \phi \textit{ then } \mathscr{M},w_2 \Vdash \psi
\end{array}$$

A formula ϕ is satisfied in a model $\mathscr{M}$*, denoted $\mathscr{M} \Vdash \phi$, if $\mathscr{M},w \Vdash \phi$ for all worlds w in $\mathscr{M}$. We write $w \Vdash \phi$ instead of $\mathscr{M},w \Vdash \phi$ whenever the model is clear from the context. ϕ is* satisfiable *if it is satisfied in some model $\mathscr{M}$, and* valid*, denoted $\Vdash\phi$, if it is satisfied in all models.* □

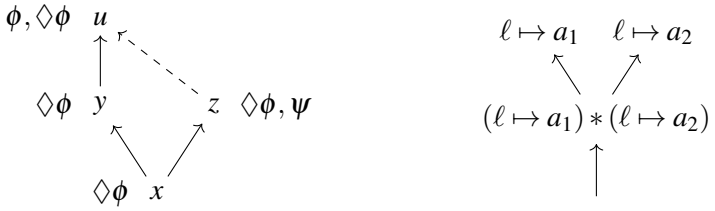


Figure 1: Examples of *BL* structures

BL uses frames that obey the persistent separation property as we believe that they better correspond to an intuitive understanding of bifurcation. We then define non-persistent (or lax) *BL* as the extension of *BL* that deals with frames that are not required to obey the persistent separation property. For example, the formula $\varphi = (\Diamond\phi * \psi) \leftrightarrow (\phi * \psi)$ is valid in *BL* (a tableau proof is given in Section 4), but it is not valid in lax *BL* as it can be falsified in the direct acyclic graph (DAG) given on the left-hand side of Figure 1. Indeed, we have $x \Vdash \Diamond\phi * \psi$ because we have $R(x,y,z)$, $y \Vdash \Diamond\phi$ and $z \Vdash \psi$, but we do not have $x \Vdash \phi * \psi$ because the only world satisfying ϕ is u and u does not belong to a distinct future of z, which is the only world satisfying ψ.

Proposition 5. *BL is a conservative extension of the modal logic S*4.

Proof. This follows immediately from observing that the only conditions imposed on the order relation are for the purpose of defining $*$ and $-\!\!*$. □

Although *BL* is a conservative extension of *S*4, it differs from both bunched and other separating logics. First, we remark that since *BL* addresses separation as an ordering problem rather than a resource composition problem (as in, for example, bunched logics, where $*$ usually corresponds to a product in a monoid), we do not include a unit $\top^*$ for the multiplicative conjunction $*$. Having the multiplicative unit $\top^*$ would unnecessarily complicate our definition of the bifurcation relation or rule out many partial order structures. Indeed, we would need to satisfy $w \Vdash \phi * \top^*$ iff $w \Vdash \phi$ for all worlds w and all formulae ϕ. Hence, by definition of $*$, we would need worlds w_1, w_2 such that $R(w, w_1, w_2)$, $w_1 \Vdash \phi$ and $w_2 \Vdash \top^*$. In particular, consider a *BL* frame with only one world w and set $\phi = \top$. We have $w \Vdash \top$, but not $w \Vdash \top * \top^*$ because $R(w,w,w)$ is impossible to achieve as it would imply both $w \leq w$ and $w \not\leq w$ by definition of R.

Second, *BL* also differs from Separation Logic, as illustrated in the right-hand side of Figure 1. In Separation Logic [14, 19], the built-in points-to predicate $(\ell \mapsto a)$ intuitively denotes a memory heap with only one cell whose location (address) is ℓ and whose value is a. Heaps are defined as partial functions from locations to values and composition of heaps is given by the union of functions with disjoint domains. Therefore, the formula $(\ell \mapsto a_1) * (\ell \mapsto a_2)$ is not satisfiable in Separation Logic as it denotes the disjoint composition of two one-cell heaps that share the same location ℓ. In *BL*, as $*$ represents bifurcation, a node satisfying $(\ell \mapsto a_1) * (\ell \mapsto a_2)$ simply implies that the location ℓ might have two distinct futures, one in which it points to the value a_1 and the other one in which it points to the value a_2. More interestingly, the formula $\Box(\ell \mapsto a)$, when satisfied by some world w, would imply that in all possible futures of w, the location ℓ should point to the same value a. This would be useful, for example, to state that an interrupt vector always points to the address where its legitimate handler resides.

3 Modelling With Bifurcation Logic

Logics can be used not merely to describe reasoning itself, but also to describe reasoning about systems. This use of logics as modelling tools has delivered significant advances in many areas — too numerous to describe here — including program analysis and verification, with one leading example, making essential use of multiplicative conjunction, being Separation Logic [14, 19]. Another highly effective example in the same spirit is Context Logic [3, 4]. More abstractly, substructural modal logics provide reasoning tools for models of systems in the 'distributed systems metaphor' (e.g., [1, 5, 11]). Demri and Deter [6] have surveyed connections between modality and separation, but they do not consider separating connectives defined through ordering. More detailed connections with Separation Logic, as mentioned above, are beyond the scope of this paper; so, instead of describing how connections with that might work, we give a quite generic example — many examples involving obtaining, representing, and verifying knowledge and information will be very similar — of how *BL*'s modality and multiplicatives interact through an example in the context of multi-agent access control.

Example 6 (Crimson Tide [22] Part 1)**.** *The plot of the film* Crimson Tide *[22] takes place mainly on board a United States Navy Ballistic Missile Submarine (the USS Alabama). During a period of international tension, the submarine receives an order to launch nuclear-armed missiles. For the release of the weapons to be authorized, a composite key must be authenticated:*

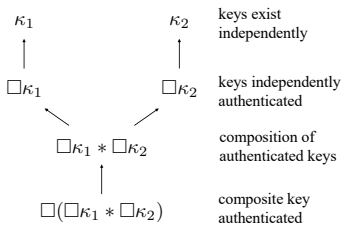


Figure 2: Joint access

1. *An order to release weapons is received in a message that contains a code.*
2. *Two senior officers (neither the Captain nor the Executive Officer) must independently authenticate the message by verifying the code against a local authentication device. Authentication (think 'necessarily correct') is denoted using* $\Box$.
3. *This yields a composite —* $\Box\kappa_1 * \Box\kappa_2$ *in Figure 2 — of authenticated keys that is passed (verbally) to the Executive Officer and the Captain.*
4. *The Executive Officer confirms that the correct protocol has been followed and so authenticates the composite key, thereby yielding* $\Box(\Box\kappa_1 * \Box\kappa_2)$ *in Figure 2.*
5. *The Captain then confirms the authentication and may order the release of the weapons.*

This concludes the first part of our example. $\Box$

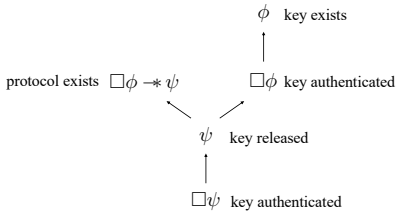


Figure 3: Protocol for obtaining a key

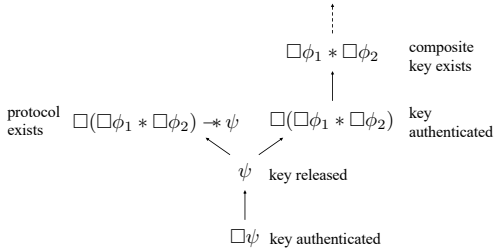


Figure 4: Protocol for obtaining a composite key

Example 7 (Crimson Tide [22] Part 2)**.** *The use of a protocol can also be represented in Bifurcation Logic. This is illustrated in Figure 3. Here the idea is that a protocol is modelled by an implicational formula that, given an authenticated key, yields a key, which may then need to be authenticated. (Informally, the implicational formula may be thought of as the type of a function that returns a key.) We can*

see how in an inessential, slightly more detailed, variant of the set-up, Figure 4 represents the use of protocols in Crimson Tide *as follows:*

1. *In the discussion based on Figure 2, the role of protocols is suppressed.*
2. *Figure 3 illustrates the use of protocols in general.*
3. *In the case of our example, we represent the Captain's key permitting the release of weapons by ψ. For the key to be authenticated, it must have been released through the protocol.*
4. *The protocol is given, in Figure 4, by the implicational formula $\Box(\Box\phi_1 * \Box\phi_2) \mathrel{-\!\!*} \psi$. That is, the authenticated composite key allows access to the Captain's key, ψ, which, when authenticated, allows the weapons to be released.*

Here we use the modality $\Box$ to denote authentication, but what is the role of separation? The role of $*$ should be clear: it enforces the independence of multiple authentications. But what of $\mathrel{-\!\!*}$? It ensures that there is no interference between the authentication of the composite key and its use to make available the Captain's key, ψ (which must itself be authenticated for use). Note the essential use of the semantics of $\mathrel{-\!\!*}$: first, the separation between worlds afforded by $\mathrel{-\!\!*}$, as opposed to $\rightarrow$, ensures no interference between the existence of the protocol and existence of the (authenticated) key to which it applies. Second, the particular semantic form of this implication — in that its application 'looks back' down the ordering — captures exactly the release of the Captains's key, ψ, through access to the authenticated composite key. □

4 A Tableaux Calculus for *BL*: T_{BL}

The tableaux calculus for classical propositional logic [8] can be adapted systematically to calculi for many non-classical logics by the addition of labelling. The basic idea is that the structure of a Kripke model for a given non-classical logic is used to define a tableaux calculus for that logic by reflecting its structure in an algebra of labels that is used to impose side-conditions on the tableaux rules. Through this mechanism, the basic classical and/or tableaux figures are modified to capture non-classical connectives.

T_{BL} is presented in Figure 5. T_{BL} has logical rules that capture the meaning of the connectives, structural rules that capture the properties of *BL* models, and closure rules (whose conclusion is a cross mark) that capture (logical or structural) inconsistencies. As usual, closure rules put an end to the expansion of a branch. T_{BL} can address either *BL* or its lax variant depending on the inclusion or not of the optional persistency rule $\| P$. All the results for *BL* presented in this section also hold for lax *BL*.

Definition 8. *Let L be a countable set of symbols called* labels. *A* labelled formula *is a pair (ϕ, x), written $\phi : x$, where ϕ is a formula and x is a label. A* label constraint *is an expression of the form $x \prec y$, where x, y are labels.* □

Definition 9. *Let Sg be the set $\{T, F\}$ of* signs. *A* signed labelled formula *is a triple (S, ϕ, x), written $S\,\phi : x$, where S is a sign and $\phi : x$ is a labelled formula. Similarly, a* signed label constraint *is a label constraint prefixed with a sign.* □

We define $T\,x \sim y$ as a shorthand for the expression $T\,x \prec y, T\,y \prec x$. Similarly, $T\,x \,\|\, y$ is a shorthand for $F\,x \prec y, F\,y \prec x$ and $T\,x \prec y \,\|\, z$ is a shorthand for $T\,x \prec y, T\,x \prec z, T\,y \,\|\, z$.

Definition 10. *A* tableau for $\phi : x$ *is a finitely branching rooted tree built inductively according to the rules given in Figure 5 and the root node of which is the signed labelled formula $F\,\phi : x$.* □

Definition 11. *A tableau branch b is* closed *if it ends with a closure rule. A tableau t is* closed *if all of its branches are closed.* □

Logical rules

$$\frac{F\,\neg\phi : x}{T\,\phi : x}\,F\neg \qquad \frac{T\,\neg\phi : x}{F\,\phi : x}\,T\neg \qquad \frac{T\,\phi \wedge \psi : x}{\begin{array}{l} T\,\phi : x \\ T\,\psi : x \end{array}}\,T\wedge \qquad \frac{F\,\phi \wedge \psi : x}{F\,\phi : x \mid F\,\psi : x}\,F\wedge$$

$$\frac{F\,\Box\phi : x}{\begin{array}{l} T\,x \prec u \\ F\,\phi : u \end{array}}\,F\Box \qquad \frac{T\,\Box\phi : x \quad T\,x \prec y}{T\,\phi : y}\,T\Box$$

$$\frac{T\,\phi * \psi : x}{\begin{array}{l} T\,x \prec u \,\|\, v \\ T\,\phi : u \\ T\,\psi : v \end{array}}\,T* \qquad \frac{F\,\phi * \psi : x \quad T\,x \prec y \,\|\, z}{F\,\phi : y \mid F\,\psi : z}\,F* \qquad \frac{T\,\phi \mathrel{-\!\!*} \psi : x \quad T\,z \prec x \,\|\, y}{F\,\phi : y \mid T\,\psi : z}\,T-\!\!*$$

$$\frac{F\,\phi \mathrel{-\!\!*} \psi : x}{\begin{array}{l} T\,v \prec x \,\|\, u \\ T\,\phi : u \\ F\,\psi : v \end{array}}\,F-\!\!* \qquad \frac{S\,\phi : x \quad T\,x \sim y}{S\,\phi : y}\,S\sim$$

Structural rules

$$\frac{S\,\phi : x,\ S\,\psi : y}{T\,x \prec y \mid T\,y \prec x \mid T\,x \,\|\, y}\,CD \qquad \frac{S\,\phi : x}{T\,x \prec x}\,{\prec}R \qquad \frac{T\,x \prec y \quad T\,y \prec z}{T\,x \prec z}\,{\prec}T$$

$$\left[\frac{\begin{array}{l} T\,x_1 \,\|\, x_2 \\ T\,x_i \prec y \end{array}}{T\,x_j \,\|\, y}\,\|P\right]$$

$$\frac{T\,x \,\|\, z}{F\,x \prec z,\ F\,z \prec x}\,\| \qquad \frac{T\,x \prec y \,\|\, z}{T\,x \prec y,\ T\,x \prec z,\ T\,y \,\|\, z}\,B \qquad \frac{T\,x \prec y \quad T\,y \prec x}{T\,x \sim y}\,\sim$$

Closure rules

$$\frac{\begin{array}{l} T\,\phi : x \\ F\,\phi : x \end{array}}{\times}\,\times_\Phi \qquad \frac{T\,x \prec y \quad F\,x \prec y}{\times}\,\times_\prec$$

Derivable rules

$$\frac{F\,\phi \vee \psi : x}{\begin{array}{l} F\,\phi : x \\ F\,\psi : x \end{array}}\,F\vee \qquad \frac{T\,\phi \vee \psi : x}{T\,\phi : x \mid T\,\psi : x}\,T\vee \qquad \frac{T\,\phi \to \psi : x}{F\,\phi : x \mid T\,\psi : x}\,T\to \qquad \frac{F\,\phi \to \psi : x}{\begin{array}{l} T\,\phi : x \\ F\,\psi : x \end{array}}\,F\to$$

$$\frac{T\,\phi \leftrightarrow \psi : x}{\begin{array}{l|l} T\,\phi : x & F\,\phi : x \\ T\,\psi : x & F\,\psi : x \end{array}}\,T\leftrightarrow \qquad \frac{F\,\phi \leftrightarrow \psi : x}{\begin{array}{l|l} F\,\phi : x & T\,\phi : x \\ T\,\psi : x & F\,\psi : x \end{array}}\,F\leftrightarrow \qquad \frac{T\,\Diamond\phi : x}{\begin{array}{l} T\,x \prec u \\ T\,\phi : u \end{array}}\,T\Diamond \qquad \frac{F\,\Diamond\phi : x \quad T\,x \prec y}{F\,\phi : y}\,F\Diamond \qquad \frac{T\,x \,\|\, y}{T\,y \,\|\, x}\,\|C$$

$$\frac{F\,\top : x}{\times}\,\times_\top \qquad \frac{T\,\bot : x}{\times}\,\times_\bot \qquad \frac{T\,x_1 \,\|\, x_2 \quad T\,x_i \prec x_j}{\times}\,\times_\| \qquad \frac{T\,x \prec y_1 \,\|\, y_2 \quad F\,x \prec y_i}{\times}\,\times_{BA} \qquad \frac{T\,x \prec y_1 \,\|\, y_2 \quad T\,y_i \prec y_j}{\times}\,\times_{BS}$$

Side-conditions and comments

- In rules introducing u and/or v, u and v are distinct labels that are fresh in the branch.
- In rules involving i and/or j, $i \in \{1,2\}$ and $j = 3 - i$.
- Double lines indicate rules that can be used bottom-up or top-down.
- The brackets indicate that the rule is optional (included for *BL*, excluded for lax *BL*).

Figure 5: Rules of the T_{BL} calculus

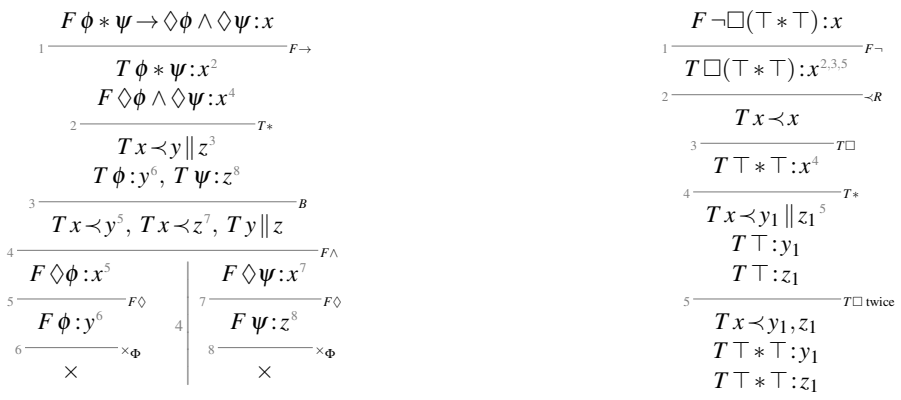
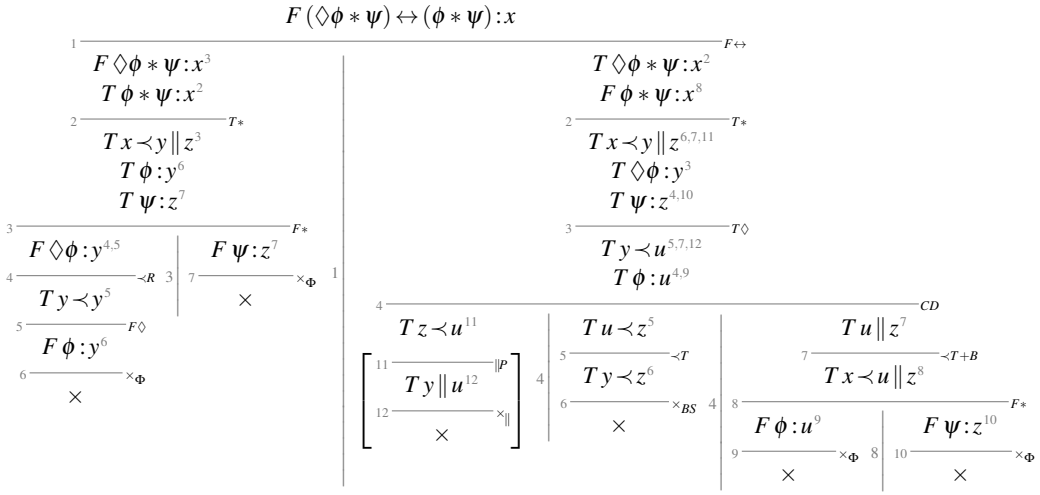


Figure 6: Tableaux examples

Definition 12. *Let $\phi : x$ be a labelled formula. A T_{BL}-*proof of $\phi : x$ *is a* closed tableau for $\phi : x$. *We write $\vdash \phi : x$ if $\phi : x$ is* provable in T_{BL}, *that is if there exists a T_{BL}-proof of $\phi : x$. Similarly, a formula ϕ is* provable *in T_{BL}, written $\vdash \phi$, if $\vdash \phi : x$ for some label x.* □

Example 13. *The tableau depicted on the left-hand side of Figure 6 is a T_{BL}-proof of $\phi * \psi \to \Diamond\phi \wedge \Diamond\psi$. Step 3 is only given to improve readability and is not really necessary as it is just the explicit expansion of the shorthand for $T\,x \prec y \,\|\, z$.* □

Example 14. *A T_{BL}-proof of $\varphi \equiv (\Diamond\phi * \psi) \leftrightarrow (\phi * \psi)$ is given in the tableau below.*



After Step 1, the tableau splits into two parts, the interesting one being the second one (on the right-hand side of the first vertical rule) that eventually leads to four closed branches. Step 4 is an example of the case distinction rule. We remark that the first case only leads to a closed branch when the optional rule $\|P$ is used (resulting in the optional steps Step 11 and Step 12). Indeed, without persistence of separation, one can build a countermodel of φ as illustrated in Figure 1. Hence, φ is a formula that distinguishes persistent from non-persistent BL. □

Example 15. *The tableau depicted on the right-hand side of Figure 6 is an example of an infinite open tableau for $\neg\Box(\top * \top)$. The signed formula $T\,\Box(\top * \top) : x$ introduced in Step 1 must be expanded for all*

labels u such that $T\, x \prec u$ occurs in the branch. The first such label is x, so that $T\, \top * \top : x$ is introduced in Step 3. The expansion of $T\, \top * \top : x$ in Step 4 generates two new successors of x, namely y_1, z_1, and then induces Step 5, where two new expansions of $T\, \Box(\top * \top) : x$ are performed (for simplicity, in one step for both y_1 and z_1). Step 5 results in the introduction of $T\, \top * \top : y_1$ and $T\, \top * \top : z_1$, the expansion of which creates four new successors of x, two of y_1 and two of z_1. The whole process described previously then repeats itself infinitely often as the BL models such that $x \Vdash \Box(\top * \top)$ are the ones in which all sucessors of x have at least two distinct successors. This point is further discussed at the beginning of Section 6. □

5 Soundness and Completeness

Definition 16. *The* domain of a tableau branch *b, denoted $D(b)$, is the set $\{ x \mid (S\, \phi : x) \in b \}$ of all labels occurring in b.* □

Definition 17 (Realization)**.** *Let b be a tableau branch. A* realization *of b in a BL model $\mathscr{M} = (W, \leq, R, V, \Vdash)$ is a function ρ from $D(b)$ to W such that*

- *if $(T\, \phi : x) \in b$, then $\rho(x) \Vdash \phi$ and if $(F\, \phi : x) \in b$, then $\rho(x) \nVdash \phi$*
- *if $(T\, x \prec y) \in b$, then $\rho(x) \leq \rho(y)$, and if $(F\, x \prec y) \in b$, then $\rho(x) \not\leq \rho(y)$.*

A tableau branch is realizable *if has at least one realization in some BL model. A tableau is* realizable *if has at least one realizable branch.* □

Lemma 18. *If a tableau branch is closed, then it is not realizable.*

Proof. Let b be a closed tableau branch. Assume that b is realizable. Then, there exists a realization ρ of b in some *BL* model.

- If b is closed because both $(T\, \phi : x) \in b$ and $(F\, \phi : x) \in b$, then by definition of a realization, we have both $\rho(x) \Vdash \phi$ and $\rho(x) \nVdash \phi$, which is a contradiction.
- If b is closed because both $(T\, x \prec y) \in b$ and $(F\, x \prec y) \in b$, then by definition of a realization, we have both $\rho(x) \leq \rho(y)$ and $\rho(x) \not\leq \rho(y)$, which is a contradiction.

Therefore, b cannot be realizable. □

Lemma 19. *If a tableau t is realizable, then expanding t using one of the tableau expansion rules given in Figure 5 results in a realizable tableau t'.*

Proof. Suppose that t is realizable. Then, it has at least one realizable branch b. If t' is obtained from t by expanding a branch that is distinct from b then t' remains realizable since it still contains the unchanged realizable branch b. Otherwise, t' is obtained by expanding b into b'. We then proceed by case analysis on the tableau rule expanding b. Let ρ be a realization of b is some *BL* model.

We consider just a few illustrative cases, the others being similar.

- Case $T{-\!\!*}$:

  If $(T\, z \prec x \,\|\, y) \in b$ and $(T\, \phi \mathrel{-\!\!*} \psi : x) \in b$, then by Definition 17, we get $R(\rho(z), \rho(x), \rho(y))$ and $\rho(x) \Vdash \phi \mathrel{-\!\!*} \psi$. It then follows from Definition 4 that $\rho(y) \Vdash \phi$ and $\rho(z) \nVdash \psi$. Therefore, ρ realizes b'.

- Case $F{-\!\!*}$:

  If $(F\, \phi \mathrel{-\!\!*} \psi : x) \in b$, then by Definition 17, we get $\rho(x) \nVdash \phi \mathrel{-\!\!*} \psi$. By Definition 4, there exist $w_1, w_2 \in W$ such that $R(w_2, \rho(x), w_1)$, $w_1 \Vdash \phi$ and $w_2 \nVdash \psi$. Since b' extends b with $(T\, v \prec x \,\|\, u)$, $(T\, \phi : u)$ and $(F\, \psi : v)$, where u and b are distinct fresh labels, we can extend ρ into a realization of b' by setting $\rho(u) = w_1$ and $\rho(v) = w_2$.

The other cases are similar. □

Theorem 20 (Soundness)**.** *If* $\vdash\phi$*, then* $\vDash\phi$*.*

Proof. If $\vdash\phi$ then we have closed tableau t for $\phi:x$ for some label x. Assume that $\nvDash\phi$. Any tableau construction procedure that results in t begins with the tableau t_0 that consists in the single node $F\,\phi:x$. Since t_0 is realizable, Lemma 19 implies that t should also be realizable and should therefore contain at least one realizable branch b. By Lemma 18, b cannot be closed. It then follows that t is open, which is a contradiction. Thus, we have $\vDash\phi$. □

Definition 21. *A tableau branch* b *is* saturated *if it satisfies all of the following conditions:*

1. *if* $(S\,\phi:x)\in b$*, then* $(T\,x\prec x)\in b$
2. *if* $(S\,\phi:x, S\,\psi:y)\in b$*, then* $(T\,x\prec y)\in b$ *or* $(T\,y\prec x)\in b$ *or* $(T\,x\parallel y)\in b$
3. *if* $(T\,x\prec y)\in b$ *and* $(T\,y\prec z)\in b$*, then* $(T\,x\prec z)\in b$
4. *if* $(T\,x\sim y)\in b$ *and* $(S\,\phi:x)\in b$*, then* $(S\,\phi:y)\in b$
5. *if* $(T\,x\parallel y)\in b$ *and* $(T\,x\prec z)\in b$*, then* $(T\,z\parallel y)\in b$
6. *if* $(T\,\neg\phi:x)\in b$*, then* $(F\,\phi:x)\in b$
7. *if* $(F\,\neg\phi:x)\in b$*, then* $(T\,\phi:x)\in b$
8. *if* $(T\,\phi\wedge\psi:x)\in b$*, then* $(T\,\phi:x)\in b$ *and* $(T\,\psi:x)\in b$
9. *if* $(F\,\phi\wedge\psi:x)\in b$*, then* $(F\,\phi:x)\in b$ *or* $(F\,\psi:x)\in b$
10. *if* $(T\,\Box\phi:x)\in b$*, then for all* $y\in D(b)$*, if* $(T\,x\prec y)\in b$*, then* $(T\,\phi:y)\in b$
11. *if* $(F\,\Box\phi:x)\in b$*, then for some* $y\in D(b)$*,* $(T\,x\prec y)\in b$ *and* $(F\,\phi:y)\in b$
12. *if* $(T\,\phi * \psi:x)\in b$*, then, for some* $y,z\in D(b)$*,* $(T\,x\prec y\parallel z)\in b$*,* $(T\,\phi:y)\in b$*, and* $(T\,\psi:z)\in b$
13. *if* $(F\,\phi * \psi:x)\in b$*, then, for all* $y,z\in D(b)$*, if* $(T\,x\prec y\parallel z)\in b$*, then* $(F\,\phi:y)\in b$ *or* $(F\,\psi:z)\in b$
14. *if* $(T\,\phi \mathrel{-\!\!*} \psi:x)\in b$*, then, for all* $y,z\in D(b)$*, if* $(T\,z\prec x\parallel y)\in b$*, then* $(F\,\phi:y)\in b$ *or* $(T\,\psi:z)\in b$
15. *if* $(F\,\phi \mathrel{-\!\!*} \psi:x)\in b$*, then, for some* $y,z\in D(b)$*,* $(T\,z\prec x\parallel y)\in b$*,* $(T\,\phi:y)\in b$*, and* $(F\,\psi:z)\in b$*.* □

Lemma 22. *Let* b *be a saturated open tableau branch. The binary relation* $\prec_b$ *over* $D(b)$ *defined as* $x\prec_b y$ *iff* $(T\,x\prec y)\in b$ *is a quasi-order over* $D(b)$ *such that if* $(F\,x\prec y)\in b$ *then* $x\not\prec_b y$*.*

Proof. Transitivity and reflexivity of $\prec_b$ clearly follow from Conditions 1 and 3 of Definition 21. Now, if $(F\,x\prec y)\in b$, then $(T\,x\prec y)\notin b$ because b is open. Hence, $x\not\prec_b y$ by definition of $\prec_b$. □

Lemma 23. *Let* b *be a saturated open branch. Let* $\sim_b$ *be the equivalence relation over* $D(b)$ *defined as* $x\sim_b y$ *iff* $x\prec_b y$ *and* $y\prec_b x$*. For all* $x\in D(b)$*, let* $[x]=\{\,y\in D(b)\mid x\sim_b y\,\}$ *denote the equivalence class of* x *under* $\sim_b$ *and let* W_b *denote the quotient* $D(b)/\sim_b$*. Then, the binary relation* $\leq_b$ *over* W_b *defined as* $[x]\leq_b[y]$ *iff* $x\prec_b y$ *is a partial order over* W_b *that satisfies the persistent separation property.*

Proof. First, we show that $\leq_b$ is well defined by showing that the following conditions are equivalent for all $[u],[v]\in W_b$:

(*a*) $u\prec_b v$

(*b*) $x\prec_b y$ for all $x\in[u]$ and all $y\in[v]$

(*c*) $x\prec_b y$ for some $x\in[u]$ and some $y\in[v]$

It is clear that (a) implies (c) and that (b) implies (c) (and (a)). Therefore, we only have to show that (c) implies (b). Assume $x \prec_b y$ for some $x \in [u]$ and some $y \in [v]$. Then, $x \sim_b u$ implies $u \prec_b x$ and $y \sim_b v$ implies $y \prec_b v$. Pick an arbitrary $x' \in [u]$, then $x' \sim_b u$ implies $x' \prec_b u$. Since $u \prec_b x$, we get $x' \prec_b x$. Pick an arbitrary $y' \in [v]$, then $y' \sim_b v$ implies $v \prec_b y'$ Since $y \prec_b v$, we get $y \prec_b y'$. Finally, since we assumed $x \prec_b y$, $x' \prec_b x$ and $y \prec_b y'$ imply $x' \prec_b y'$.

Second, we show that $\leq_b$ is a partial order over W_b. Since $\prec_b$ is a quasi-order over $D(b)$ by Lemma 22, it immediately follows that $\leq_b$ is both reflexive and transitive. It remains to show that $\leq_b$ is anti-symmetric. Assume $[x] \leq_b [y]$ and $[y] \leq_b [x]$, then, by definition of $\leq_b$, we have $x \prec_b y$ and $y \prec_b x$, from which we get $x \sim_b y$ by definition of $\sim_b$. Hence, $[x] = [y]$.

Last, we show that $\leq_b$ satisfies the separation persistence property stated in Definition 2. We pick arbitrary $[x], [y], [z] \in W_b$ such that $[x] \perp [y]$ and $[x] \leq_b [z]$ and show that $[z] \perp [y]$. We have the following facts:

(i) By definition of $\perp$, $[x] \perp [y]$ implies $[x] \not\leq_b [y]$ and $[y] \not\leq_b [x]$.

(ii) By definition of $\leq_b$, $[x] \leq_b [z]$ implies $x \prec_b z$.

Assume that $[z] \leq_b [y]$. Then, $z \prec_b y$ by definition of $\leq_b$. Since $x \prec_b z$ and $z \prec_b y$ imply $x \prec_b y$, we get $[x] \leq_b [y]$ by definition of $\leq_b$, which contradicts $[x] \perp [y]$ (i). Hence, $[z] \not\leq_b [y]$.
Assume that $[y] \leq_b [z]$. Then, $y \prec_b z$ by definition of $\leq_b$, from which we get $(T\, y \prec z) \in b$ by definition of $\prec_b$. Besides, $[x] \not\leq_b [y]$ and $[y] \not\leq_b [x]$ imply $x \not\prec_b y$ and $y \not\prec_b x$ by definition of $\leq_b$. By definition of $\prec_b$, we get $(T\, x \prec y) \notin b$ and $(T\, y \prec x) \notin b$. Since b is saturated, $(T\, x \,\|\, y) \in b$ then follows from Condition 2 of Definition 21. In turn, $(T\, x \,\|\, y) \in b$ and $(T\, y \prec z) \in b$ imply $(T\, x \,\|\, z) \in b$ by Condition 5 of Definition 21. $(T\, x \,\|\, z) \in b$ implies $(F\, x \prec z) \in b$ by definition of $\|$, but from $(F\, x \prec z) \in b$, Lemma 22 implies $x \not\prec_b z$, which contradicts $x \prec_b z$ (ii). Hence, $[y] \not\leq_b [z]$. From, $[z] \not\leq_b [y]$ and $[y] \not\leq_b [z]$, we conclude $[y] \perp [z]$ by definition of $\perp$. □

Lemma 24 (Model existence)**.** *Let b be a saturated open tableau branch. Then, b induces a BL model $\mathscr{M}_b = (W, \leq, R, V, \Vdash)$, where $W = W_b$ and $\leq = \leq_b$ as per Lemma 23, R is the ternary relation induced by $\leq_b$ as per Definition 3, and, for all worlds $[x] \in W_b$, $V([x]) = \{\, p \mid (T\, p : x) \in b \,\}$. Moreover, $\mathscr{M}_b$ is such that if $(T\, \phi : x) \in b$ then $[x] \Vdash \phi$ and if $(F\, \phi : x) \in b$ then $[x] \nVdash \phi$.*

Proof. By mutual induction on the structure of ϕ for all labels x.

Base case $\phi = p$:

If $(T\, p : x) \in b$, then $p \in V([x])$ by definition of V, which implies $[x] \Vdash p$ by Definition 4.

If $(F\, p : x) \in b$, then since b is open, we have $(T\, p : x) \notin b$. Hence, we get $p \notin V([x])$ by definition of V, which implies $[x] \nVdash p$ by Definition 4.

Case $\phi = \phi_1 \mathrel{-\!\!*} \phi_2$:

If $(F\, \phi_1 \mathrel{-\!\!*} \phi_2 : x) \in b$, then for some $y, z \in D(b)$, we have $(T\, z \prec x \,\|\, y) \in b$, $(T\, \phi_1 : y) \in b$ and $(F\, \phi_2 : z) \in b$. $(T\, z \prec x \,\|\, y) \in b$ implies $R([z], [x], [y])$ by Lemma 23. Moreover, by induction hypothesis, $(T\, \phi_1 : y) \in b$ and $(F\, \phi_2 : z) \in b$ imply $[y] \Vdash \phi_1$ and $[z] \nVdash \phi_2$. Hence, by Definition 4, we have $[x] \nVdash \phi_1 \mathrel{-\!\!*} \phi_2$.

If $(T\, \phi_1 \mathrel{-\!\!*} \phi_2 : x) \in b$, then for all $y, z \in D(b)$, if $(T\, z \prec x \,\|\, y) \in b$ then $(F\, \phi_1 : y) \in b$ or $(T\, \phi_2 : z) \in b$. Pick arbitrary $[u], [v] \in W_b$ such that $R([v], [x], [u])$ and $[u] \Vdash \phi_1$. Since $[v] \leq_b [x]$ and $[v] \leq_b [u]$, we have $(T\, v \prec x) \in b$ and $(T\, v \prec u) \in b$ by definition of $\leq_b$. Similarly, since $[x] \not\leq_b [u]$ and $[u] \not\leq_b [x]$, we have $(T\, x \prec u) \notin b$ and $(T\, u \prec x) \notin b$, which by Condition 2 implies $(T\, x \,\|\, u) \in b$. It then follows that we have $(T\, v \prec x \,\|\, u) \in b$, which implies $(F\, \phi_1 : u) \in b$ or $(T\, \phi_2 : v) \in b$. By the induction

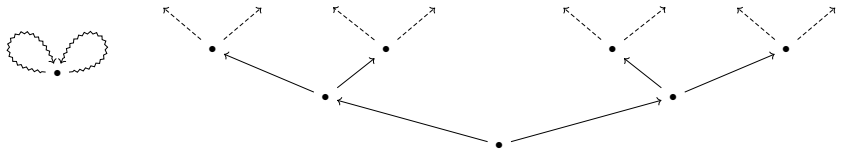


Figure 7: A finite model with backlinks (left) and its infinite unfolding (right)

hypothesis, we then get $[u] \nVdash \phi_1$ or $[v] \Vdash \phi_2$. Since we assume $[u] \Vdash \phi_1$, we necessarily have $[v] \Vdash \phi_2$. Therefore, $[x] \Vdash \phi_1 \mathrel{-\!\!*} \phi_2$.

The other cases are similar. □

Corollary 25. *If b is a saturated open tableau branch in a tableau for $\phi : x$, then the induced model $\mathscr{M}_b = (W_b, \leq_b, R, V, \Vdash)$ is such that $[x] \nVdash \phi$.*

Proof. Since $(F\,\phi : x)$ is the root of any tableau for $\phi : x$, The concluding property of Lemma 24 implies that $x \nVdash \phi$. □

Theorem 26 (Completeness)**.** *If $\Vdash \phi$, then $\vdash \phi$.*

Proof. It is standard to define a tableau construction procedure that applies the rules given in Figure 5 with a fair strategy. Such a procedure will either result in a finite closed tableau for ϕ, in which case we get a T_{BL}-proof of ϕ, or build at least one (possibly infinite) complete open branch b, in which case b gives rise to a *BL* model $\mathscr{M}_b$ such that $\mathscr{M}_b \not\models \phi$ by Corollary 25. □

6 A Finite Model Property and Decidability

The formula $\Box(\top * \top)$ enforces that every world has at least two distinct successors and can, therefore, only be satisfied in infinite models like the one shown in Figure 7 (right). Consequently, the finite model property in its traditional formulation fails for *BL*.

However, models such as the one in Figure 7 (right) are sufficiently regular to allow for a finite schematic representation. Consider, for example, the scheme in Figure 7 (left). Here, each squiggly edge represents a *backlink*, indicating that a copy of the graph originating from its endpoint should be attached to its starting point via a single edge. By unfolding such a schematic model, we obtain precisely the full binary tree shown in Figure 7 (right).

For simplicity, we only present an informal description of *models with backlinks*. These models are trees augmented with additional backlinks — edges that point backward in the tree order. Formulas in BL can be interpreted on such models by first unfolding them into (infinite) trees. One can then establish the following result:

Lemma 27. *If φ is $\mathrel{-\!\!*}$-free and satisfiable in BL, then it has a finite model with backlinks.*

Proof. Let Σ be the set of subformulae of φ. For a subset $\Sigma_0 \subseteq \Sigma$, consider all immmediate successors of the root node that satisfy exactly those formulae in Σ that are in Σ_0. If we remove all but two of such nodes and the branches stemming from them we obtain a different model that still satisfies φ at the root. Note that we need to retain two nodes–unlike in standard modal logic where one representative suffices– as the root might satisfy a statement $\psi * \psi$ that enforces the existence of two *different* successors, even

with equal theories. Applying this idea repeatedly we can create a model of φ that is finitely branching up to any given height.

On the other hand, if a branch is longer than $2^{|\Sigma|}$, we will encounter worlds $w < v$ satisfying exactly the same formulae from Σ. In this case, we can remove the branch stemming from v and instead create a back link from the immediate predecessor of v to w. This too preserves truth of φ at the root. In the end, we obtain a finitely branching model with backlinks that is of bounded height, and therefore finite. □

The proof is a modification of a standard filtration and pruning argument used in other modal logics [2]. These arguments rely on the fact that the truth of a formula depends only on the truth of its subformulae at successor nodes — a property that fails once one considers backwards-looking modalities. This explains the restriction on $-\!\!*$.

Corollary 28. *It is decidable whether a $-\!\!*$-free formula is valid in BL.*

Proof. By Theorem 26, the $-\!\!*$-free formulae of *BL* are computably enumerable. Furthermore, Lemma 27 implies that and it follows that their complement is computably enumerable too, since we can systematically generate all finite models with backlinks. Thus, decidability follows. □

Acknowledgements

1. Galmiche and Méry gratefully acknowledge the partial support of the ANR Projet NARCO (ANR-21-CE48-0011).
2. Lang and Pym gratefully acknowledge the partial support of the UK EPSRC through Research Grants EP/S013008/1 and EP/R006865/1.
3. We thank the anonymous referees for their suggestions.

References

[1] Gabrielle Anderson & David J Pym (2016): *A Calculus and Logic of Bunched Resources and Processes*. *Theoretical Computer Science* 614, pp. 63–96, doi:10.1016/j.tcs.2015.11.035.

[2] Patrick Blackburn, Maarten De Rijke & Yde Venema (2001): *Modal Logic*. Cambridge University Press, doi:10.1017/CBO9781107050884.

[3] Cristiano Calcagno, Philippa Gardner & Uri Zarfaty (2005): *Context Logic and Tree Update*. *ACM SIGPLAN Notices* 40(1), pp. 271–282, doi:10.1145/1047659.1040328.

[4] Cristiano Calcagno, Philippa Gardner & Uri Zarfaty (2007): *Context Logic as Modal Logic: Completeness and Parametric Inexpressivity*. *ACM SIGPLAN Notices* 42(1), pp. 123–134, doi:10.1145/1190215.1190236.

[5] Tristan Caulfield, Marius-Constantin Ilau & David J Pym (2021): *Engineering Ecosystem Models: Semantics and Pragmatics*. In: *13th EAI International Conference on Simulation Tools and Techniques*, Springer, pp. 236–258, doi:10.1007/978-3-030-97124-3_21.

[6] Stéphane Demri & Morgan Deters (2015): *Separation logics and modalities: a survey*. *Journal of Applied Non-Classical Logics* 25(1), pp. 50–99, doi:10.1080/11663081.2015.1018801.

[7] Kosta Došen (1992): *Modal translations in substructural logics*. *Journal of Philosophical Logic* 21, pp. 283–336, doi:10.1007/BF00260931.

[8] Melvin Fitting (2012): *First-Order Logic and Automated Theorem Proving*. Springer Science & Business Media, doi:10.1007/978-1-4612-2360-3.

[9] André Fuhrmann & Edwin D. Mares (1994): *On S*. *Studia Logica* 53(1), pp. 75–91, doi:10.1007/bf01053023.

[10] Didier Galmiche, Pierre Kimmel & David J Pym (2019): *A substructural epistemic resource logic: theory and modelling applications*. Journal of Logic and Computation 29(8), pp. 1251–1287, doi:10.1093/logcom/exz024.

[11] Didier Galmiche, Timo Lang & David J Pym (2024): *Minimalistic System Modelling: Behaviours, Interfaces, and Local Reasoning*. In: 16th EAI International Conference on Simulation Tools and Techniques, Springer, pp. 48–71, doi:10.1007/978-3-031-87345-4_4.

[12] Didier Galmiche, Daniel Méry & David J Pym (2005): *The semantics of BI and resource tableaux*. Mathematical Structures in Computer Science 15(6), pp. 1033–1088, doi:10.1017/S0960129505004858.

[13] Alexander V Gheorghiu & David J Pym (2023): *Semantical Analysis of the Logic of Bunched Implications*. Studia Logica 111(4), pp. 525–571, doi:10.1007/s11225-022-10028-z.

[14] Samin S Ishtiaq & Peter W O'Hearn (2001): *BI as an Assertion Language for Mutable Data Structures*. In: Proceedings of the 28th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, pp. 14–26, doi:10.1145/360204.375719.

[15] Norihiro Kamide (2002): *Kripke Semantics for Modal Substructural Logics*. Journal of Logic, Language, and Information 11, pp. 453–470, doi:10.1023/A:1019915908844.

[16] Peter W O'Hearn & David J Pym (1999): *The Logic of Bunched Implications*. Bulletin of Symbolic Logic 5(2), pp. 215–244, doi:10.2307/421090.

[17] Hiroakira Ono (2019): *Proof Theory and Algebra in Logic*. Springer, doi:10.1007/978-981-13-7997-0.

[18] David J Pym, Peter W O'Hearn & Hongseok Yang (2004): *Possible Worlds and Resources: The Semantics of BI*. Theoretical Computer Science 315(1), pp. 257–305, doi:10.1016/j.tcs.2003.11.020.

[19] John C Reynolds (2002): *Separation Logic: A Logic for Shared Mutable Data Structures*. In: Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science, IEEE, pp. 55–74, doi:10.1109/LICS.2002.1029817.

[20] Richard Routley & Robert Meyer (1973): *The Semantics of Entailment*. In: Studies in Logic and the Foundations of Mathematics, 68, Elsevier, pp. 199–243, doi:10.1016/S0049-237X(08)71541-6.

[21] Alex K Simpson (1994): *The Proof Theory and Semantics of Intuitionistic Modal Logic*. Ph.D. thesis, School of Informatics, University of Edinburgh. Available at `https://era.ed.ac.uk/handle/1842/407`.

[22] Don Simpson & Jerry Bruckheimer (Producers) (1995): *Crimson Tide [movie]*. Available at `https://www.imdb.com/title/tt0112740/`. Directed by Tony Scott.

Modal Logic for Simulation, Refinement, and Mutual Ignorance

Hans van Ditmarsch
University of Toulouse
CNRS, IRIT, France
hansvanditmarsch@gmail.com

Tim French
University of Western Australia
Perth, Australia
tim.french@uwa.edu.au

Rustam Galimullin
University of Bergen
Norway
rustam.galimullin@uib.no

Louwe B. Kuijer
Liverpool University
UK
lbkuijer@liverpool.ac.uk

Simulation and refinement are variations of the bisimulation relation, where in the former we keep only **atoms** and **forth**, and in the latter only **atoms** and **back**. Quantifying over simulations and refinements captures the effects of information change in a multi-agent system. In the case of quantification over refinements, we are looking at all the ways the agents in a system can become more informed. Similarly, in the case of quantification over simulations, we are dealing with all the ways the agents can become less informed, or in other words, could have been less informed, as we are at liberty how to interpret time in dynamic epistemic logic. While quantification over refinements has been well explored in the literature, quantification over simulations has received considerably less attention. In this paper, we explore the relationship between refinements and simulations. To this end, we also employ the notion of mutual factual ignorance that allows us to capture the state of a model before agents have learnt any factual information. In particular, we consider the extensions of multi-modal logic with the simulation and refinement modalities, as well as modalities for mutual factual ignorance. We provide reduction-based axiomatizations for several of the resulting logics that are built extending one another in a modular fashion.

1 Introduction

To a multi-agent modal logic, and in particular to epistemic logic [35], we can add modalities that are quantifiers over arbitrary information change, called 'refinement quantifiers'. In *refinement modal logic* (RML) [20, 21, 33, 30, 14, 32, 15, 1, 44] one can intuitively say by $\langle \leftrightarrows \rangle \varphi$ that there is some public or private change of information after which the formula φ is true. It is known that it is equivalent to saying that there is an action model [7] such that after its execution, φ is true; as in arbitrary action model logic [31]. However, note that we then do not have modalities for action models in the logical language. Instead, in RML we quantify over the refinements of a given model. From the requirements **atoms**, **forth** and **back** of a bisimulation [13], a refinement relation only needs to satisfy **atoms** and **back**. The semantics of the refinement quantifier directly uses this refinement relation.

In this contribution we investigate the dual logic, called *simulation modal logic* (SML), that quantifies over the simulations of a given model. Now, from the requirements of a bisimulation, a simulation relation only needs to satisfy **atoms** and **forth**. Just as RML describes, in an epistemic setting, all the intricate ways in which agents can become more informed about the current state of information, in simulation modal logic we describe how agents can become less informed about the current state of information. Becoming less informed can be seen as a kind of forgetting [41, 36, 22, 28, 27, 37],

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 379–398, doi:10.4204/EPTCS.437.30

© H. van Ditmarsch, T. French, R. Galimullin & L.B. Kuijer
This work is licensed under the
Creative Commons Attribution License.

but possibly a more appealing intuition is to imagine a prior state of information to the current one, wherein agents were still not as informed and knowledgeable as they are now: just as refinement is a kind of *belief expansion*, simulation is therefore a kind of *belief contraction* [2]. Refinement modal logic and simulation modal logic are dynamic epistemic logics: logics with epistemic modalities as well as dynamic modalities interpreted as model updates [24, 6, 39]. Although there are various ways to deal with belief contraction in a modal epistemic logic [16, 4, 23], no general mechanism for belief contraction in dynamic epistemic logic has been proposed to our knowledge. Information updates such as public announcements amount to belief expansion [26, 3], whereas dynamic belief revision, often using Kripke models with more structure such as plausibility relations or preference relations have also found widespread use [10, 9, 17, 12]. There are two different ways to represent belief contraction in dynamic epistemic logic: logical languages with dynamic modalities referring to what was true in the past given a history-based semantics [40, 5, 8], and logical languages with dynamic modalities updating the current model with respect to some formula parametrizing it [22, 28]. Examples of the former are the dynamic epistemic logic with a 'yesterday' modality [40], the logic of what is true before a public announcement [5], and arbitrary public announcement logic with memory [8]. Examples of the latter are [22], merely forgetting atomic propositions, and [28], forgetting arbitrary modal formulas. One could say that [28] comes closest to the status quo of belief contraction in dynamic epistemic logic. With our proposed simulation modal logic we wish to further contribute to the so far quite restricted literature on belief contraction in dynamic epistemic logic in a meaningful way: it does not merely represent what was known before an announcement, a public action, but what was known before any informative action, such as also private announcements.

Despite simulation and refinement being dual, there is still a somewhat fundamental difference between the two, that can be summarized by saying that there are far more ways to become less informed than there are ways to become more informed. For example, given a publicly announced formula φ and an epistemic model M encoding what agents know, there is a unique updated model resulting from that public announcement, but there are many epistemic models such that updating them with the announcement of φ results in the model M.

In this contribution we also axiomatize SML. In order to axiomatize SML we follow two different paths. First, we propose an axiomatization that can be seen as the dual of that of refinement modal logic, however, with the difference that we have consistency requirements on formulas occurring in one of the axioms (the axiom involves knowledge of a disjunction of simulations of formulas which can be consistent even if one of the disjuncts is inconsistent). Second, we propose an integrated axiomatization of refinement modal logic and simulation modal logic avoiding this complication of the consistency requirement. The need for refinement modalities can be explained from the semantics: we compare any given information state (pointed Kripke model) M_s wherein we wish to evaluate a formula $\langle\rightrightarrows\rangle\varphi$ (there is simulation after which φ is true) to the (pointed) model M°_t of (maximal) mutual factual ignorance (also known as the epistemic model of *blissful ignorance* [19]). We typically do not have that M is a submodel of M°, and also not that φ is true in M°, but that φ is true in a pointed model $M'_{s'}$ such that M' is a refinement of M°. Model M' is so to speak 'in between M and M°': we then have that $M \rightrightarrows M' \rightrightarrows M^\circ$ (M° simulates M' and M' simulates M, or dually, M refines M' and M' refines M°). This complication also requires us to refer to truth in the mutual factual ignorance model M° by a dedicated modal construct $[\circ]\psi$, for 'ψ was *originally* true'. The consistency of constituents of φ in the above $\langle\rightrightarrows\rangle\varphi$ then translates into the requirement that these constituents of φ can become true after a refinement of the mutual factual ignorance model, meaning that they are satisfiable and therefore consistent.

In this paper we examine the relationships between the three modalities, $[\rightrightarrows]$, $[\leftrightarrows]$ and $[\circ]$, and consider their logics. We start by recalling the established results for RML and SML, and present novel

axiomatizations for each. Then we consider the combination of simulation and refinement, and in so doing also introduce *origin modal logic* OML. We axiomatize OML on its own, and furthermore in combination with simulation modalities and refinement quantifiers, in the axiomatization **ROSML**. It should be noted that all our results are for epistemic models with arbitrary accessibility relations for each agent, and not with equivalence relations. Although fair game, the intended epistemic models in dynamic epistemic logic are with equivalence relations, in which case the epistemic modality represents knowledge. The extension of our results to knowledge is left for future research.

2 Syntax and Semantics

Languages Given a non-empty countable (finite or infinite) set of *propositional variables* (*atoms*) P and a non-empty finite set of *agents* A, we consider a combined language with refinement and simulation quantifiers, as well as the origin modality. The elements of the language are the *formulas*.

$$\mathscr{L} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_a\varphi \mid [\leftrightarrows]\varphi \mid [\rightrightarrows]\varphi \mid [\circ]\varphi$$

where $p \in P$, $a \in A$. The language $\mathscr{L}^{\Box\leftrightarrows}$ is the fragment with only $\Box_a$ and $[\leftrightarrows]$ modalities, the language $\mathscr{L}^{\Box\rightrightarrows}$ is the fragment with only $\Box_a$ and $[\rightrightarrows]$ modalities, the language $\mathscr{L}^{\Box\circ}$ is the fragment with only $\Box_a$ and $[\circ]$ modalities, and $\mathscr{L}^{\Box}$ is the fragment with only the $\Box_a$ modalities. The propositional fragment is $\mathscr{L}_0$. Other propositional connectives are defined by abbreviation, and we also let $\Diamond_a\varphi := \neg\Box_a\neg\varphi$. In an epistemic setting, for $\Diamond_a\varphi$ we read 'agent a considers φ possible', and for $\Box_a\varphi$, 'agent a knows φ'. For $[\leftrightarrows]\varphi$ we read 'after any refinement, φ', and for $[\rightrightarrows]\varphi$, 'after any simulation, φ'. We also define $\langle\leftrightarrows\rangle\varphi$ as $\neg[\leftrightarrows]\neg\varphi$, and $\langle\rightrightarrows\rangle\varphi$ as $\neg[\rightrightarrows]\neg\varphi$. For $[\circ]\varphi$ we read 'originally φ (was true)'. We will call $[\circ]$ the *origin modality*. Finally, the *cover* operator $\nabla_a\Phi$, where Φ is a finite set of formulas, is $\bigwedge_{\varphi\in\Phi}\Diamond_a\varphi \wedge \Box_a\bigvee_{\varphi\in\Phi}\varphi$.

Given a formula φ, $d(\varphi)$ is the *modal depth* or $\Box$-depth of φ. The definition of $\Box$-depth is: $d(p) = 0$, $d(\neg\varphi) = d([\leftrightarrows]\varphi) = d([\rightrightarrows]\varphi) = d([\circ]\varphi) = d(\varphi)$, $d(\varphi\wedge\psi) = \max\{d(\varphi), d(\psi)\}$, $d(\Box_a\varphi) = d(\varphi)+1$.

Structures An *epistemic model* (or 'Kripke model', or just 'model') $M = (S, R, V)$ consists of a non-empty countable *domain* S of *states* (or 'worlds'; the domain is also denoted $\mathscr{D}(M)$), an *accessibility function* $R : A \to \mathscr{P}(S\times S)$, where each $R(a)$ is an *accessibility relation*, and a *valuation* $V : S \to \mathscr{P}(P)$; each $V(s)$ is the set of atoms that are true in s. For $s \in S$, a pair (M, s), for which we write M_s, is a *pointed (epistemic) model*. The model class without any restrictions is K. The class of models where all accessibility relations are equivalence relations is S5.

The *mutual factual ignorance model* M° is defined as $M = (S^\circ, R^\circ, V^\circ)$, where S° is the set of all subsets of P, $R^\circ_a = S^\circ \times S^\circ$, and $V(s) = s$ for each $s \in S^\circ$. Given that P is countably infinite, $|S^\circ|$ is not countable.

Let models $M = (S, R, V)$ and $M' = (S', R', V')$ be given. A non-empty relation $Z \subseteq S \times S'$ is a *bisimulation* between M and M', notation $Z : M \leftrightarrows M'$, if for all pairs $(s, s') \in Z$ and $a \in A$:

atoms $V(s) = V'(s')$;

forth if $R_a st$, then there is a $t' \in S'$ such that $R'_a s't'$ and Ztt';

back if $R'_a s't'$, then there is a $t \in S$ such that $R_a st$ and Ztt'.

We write $M \leftrightarrows M'$ if there is a bisimulation between M and M', and we write $M_s \leftrightarrows M'_{s'}$ if there is a bisimulation between M and M' containing pair (s, s').

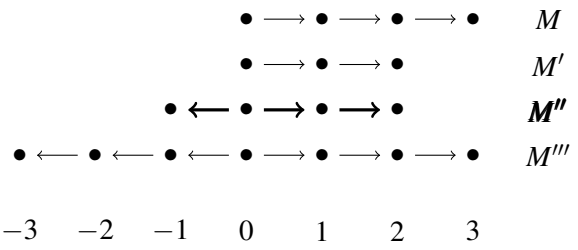
Similarly, a *simulation* satisfies **atoms** and **forth**, where we write $M \rightrightarrows M'$ if there is a simulation between M and M', and $M_s \rightrightarrows M'_{s'}$ if it contains pair (s, s'); whereas its dual, the *refinement*, satisfies

atoms and **back**, and where we write $M \leftrightarrows M'$ if there exists a refinement between M and M', and $M_s \leftrightarrows M'_{s'}$ if it contains pair (s,s'). Further, if $M \leftrightarrows M'$, the model M' is called a refinement of M as well. It is sometimes confusing that in the literature the term 'refinement' denotes both the refinement relation linking M and M', and the refined model M'.

Semantics Assume an epistemic model $M = (S,R,V)$, and let $s \in S$. We define $M_s \models \varphi$ (for: M_s *satisfies* φ, or φ is *true* in M_s) by induction.

$$
\begin{array}{lll}
M_s \models p & \text{iff} & p \in V(s) \\
M_s \models \neg\varphi & \text{iff} & M_s \not\models \varphi \\
M_s \models \varphi \wedge \psi & \text{iff} & M_s \models \varphi \text{ and } M_s \models \psi \\
M_s \models \Box_a \varphi & \text{iff} & M_t \models \varphi \text{ for all } t \in S \text{ such that } (s,t) \in R_a \\
M_s \models [\leftrightarrows]\varphi & \text{iff} & M'_{s'} \models \varphi \text{ for all } M'_{s'} \text{ such that } M_s \leftrightarrows M'_{s'} \\
M_s \models [\rightrightarrows]\varphi & \text{iff} & M'_{s'} \models \varphi \text{ for all } M'_{s'} \text{ such that } M_s \rightrightarrows M'_{s'} \\
M_s \models [\circ]\varphi & \text{iff} & M^{\circ}_{V(s)} \models \varphi
\end{array}
$$

Let us give an example. A refinement of a given model need not be a submodel and a simulation of a given model need not contain the given model. Consider the models depicted below. Assume that i and $-i$ states have the same valuation of atoms.



We note that M' is a submodel (model restriction) of M and M'' is a refinement of M that is not a submodel. However, M'' is a submodel of M''' and M''' is a bisimilar copy of M. Simulations are duals of refinement. So, M is a simulation of M'' even though M'' is not a submodel of M.

In an epistemic setting, one could imagine a two-agent situation where a and b both know that a propositional variable p is true, but are both uncertain whether the other one knows, as encoded in the model M below in state s. A simulation of this is the two-state mutual factual ignorance model M°. It is clear that M° does not contain M. But M is a refinement of bisimilar copy M'' of M°. The model M° can be seen as a previous state of information that a and b were in, where they were both informed that p is true while they remained both uncertain whether the other was also informed.

$$
\begin{array}{ll}
M: & \neg p \overset{a}{\text{———}} p \overset{b}{\text{———}} \overset{s}{p} \overset{a}{\text{———}} p \overset{b}{\text{———}} \neg p \\
M'': & \neg p \overset{ab}{\text{———}} p \overset{ab}{\text{———}} \overset{s}{p} \overset{ab}{\text{———}} p \overset{ab}{\text{———}} \neg p \\
M^{\circ}: & \qquad\qquad\qquad\quad \overset{s}{p} \overset{ab}{\text{———}} \neg p
\end{array}
$$

3 Simulation, Refinement, and Mutual Factual Ignorance

Refinement We review some results obtained for refinement modal logic over the years for which we present a novel axiomatization in the next section.

In [14] and most other works on RML, instead of the refinement relation as we defined above (for the set of all agents), the refinement relation has a parameter $G \subseteq A$, where apart from **atoms** and **back** for all agents, additionally, **forth** is required for the agents in $A \setminus G$. Here we only consider refinement for the set A of all agents, so that the only requirement is **atoms** and **back** for all agents. This simplifies the presentation.

As the refinement relation is transitive, reflexive, and confluent (for all x, y, z, if $x \leftleftarrows y$ and $x \leftleftarrows z$, there is w such that $y \leftleftarrows w$ and $z \leftleftarrows w$), and we additionally have atomicity (there is a maximal refinement: remove all relations!), we have for the refinement modality the corresponding validities:

- $[\leftleftarrows]\varphi \to \varphi$ ($\mathbf{T}^{\leftleftarrows}$)
- $\langle\leftleftarrows\rangle\langle\leftleftarrows\rangle\varphi \to \langle\leftleftarrows\rangle\varphi$ ($\mathbf{4}^{\leftleftarrows}$)
- $\langle\leftleftarrows\rangle[\leftleftarrows]\varphi \to [\leftleftarrows]\langle\leftleftarrows\rangle\varphi$ (Church-Rosser) / $\mathbf{CR}^{\leftleftarrows}$
- $[\leftleftarrows]\langle\leftleftarrows\rangle\varphi \to \langle\leftleftarrows\rangle[\leftleftarrows]\varphi$ (McKinsey) / $\mathbf{MK}^{\leftleftarrows}$

The requirement of **atoms** and **back** entails that a refinement of a given model is a restriction of a bisimilar copy of that model. Analogously, in the logical language, refinement quantification is bisimulation quantification followed by relativization. In *bisimulation quantified* logics we have explicit quantifiers $\forall p$ over propositional variables p [29, 42]. *Relativization* φ^p of a formula φ to an atomic proposition p is a syntactic way to describe model restrictions, such as the consequences of a public announcement: $(M|p)_s \models \varphi$ iff $M_s \models \varphi^p$, which in public announcement logic is the same as $M_s \models \langle p\rangle\varphi$, see [38, 11] (here $M|p$ is the restriction of the model to the state where p is true, and $\langle p\rangle\varphi$ means 'p is true and after its announcement φ is true'). In RML, given $\varphi \in \mathscr{L}^{\Box}$, we have [14, Sect. 4.3]:

$$\langle\leftleftarrows\rangle\varphi \text{ is equivalent to } \exists p\langle p\rangle\varphi$$

Furthermore, the refinement quantifier can be seen as a quantification over *action models* [7]. Already on the level of the semantics we have that:

> *Executing an action model produces a refinement of the initial model, and for every refinement of a finite model there is an action model producing it.*

On the level of the language we can, given $\langle\leftleftarrows\rangle\varphi$, synthesize an action model U with the same update effect (such that $\langle U\rangle\varphi$ is true). For details, see [31] and the related (for a different but comparable update mechanism, called 'arrow update') [25].

The axiomatization of RML reported in [14] is in terms of parametrized refinement quantifiers $\langle\leftleftarrows_a\rangle$ for $a \in A$ (interpreting a-refinement relations $\leftleftarrows_a$) with which one can also define $\langle\leftleftarrows_G\rangle$ for $G \subseteq A$. We recall that we now only consider $\leftleftarrows$, that is, $\leftleftarrows_A$). We therefore wish to report in this contribution an alternative axiomatization **RML** for RML with only refinement quantifiers $\leftleftarrows$ (Section 4). It resembles the quantifier part of the axiomatization of arbitrary arrow update model logic [25], and was reported in [18] without proof of soundness and completeness. It is simpler than the axiomatization in [14] and is also based on reduction axioms.

Simulation Perhaps the most relevant work here is [44], where the authors explore the logic that combines refinement and simulation in a single modality called covariant-contravariant refinement (modality). It is inspired by process calculi and not by epistemic modalities and updates. As in [14], these modalities are parameterized by the subgroups of the set of all agents for which the refinement (resp. simulation) relation only needs to satisfy **back** (resp. **forth**). In this work, we only refine or simulate

by the set of all agents, thus getting a more succinct formalization **SML**, presented in Section 4. In that section we furthermore present the axiomatization **ROSML**, that is truly novel.

An aspect of **SML** that might be criticized is the somewhat extra-logical (or at least uncommon in a Hilbert-style axiomatization) requirement of consistency of the formulas appearing in the cover. Proofs and algorithms are different beasts, but it is an aesthetic principle that determining whether a rule or axiom is applicable in a proof should be self-evident; at least in the sense that it can be determined in linear time. We get rid of this problem in the **ROSML** axiomatization by a more complex reduction also involving origin modalities and refinement modalities but having no extra-logical requirements.

Just as the refinement relation, the simulation relation is also transitive, reflexive, and confluent, and again we have atomicity (the maximal simulation brings you in the mutual factual ignorance model). We therefore have for the simulation modality the following validities.

Proposition 1 The following are valid:

1. $[\rightrightarrows]\varphi \to \varphi$ $(\mathbf{T}^{\rightrightarrows})$
2. $\langle\rightrightarrows\rangle\langle\rightrightarrows\rangle\varphi \to \langle\rightrightarrows\rangle\varphi$ $(\mathbf{4}^{\rightrightarrows})$
3. $\langle\rightrightarrows\rangle[\rightrightarrows]\varphi \to [\rightrightarrows]\langle\rightrightarrows\rangle\varphi$ (Church-Rosser) / $\mathbf{CR}^{\rightrightarrows}$
4. $[\rightrightarrows]\langle\rightrightarrows\rangle\varphi \to \langle\rightrightarrows\rangle[\rightrightarrows]\varphi$ (McKinsey) / $\mathbf{MK}^{\rightrightarrows}$

Proof Let $M = (S,R,V)$ and $s \in S$ be given.

1. Assume $M_s \models [\rightrightarrows]\varphi$. As $M \rightrightarrows M$ (the relation $\rightrightarrows$ is reflexive), we directly obtain that $M_s \models \varphi$. Therefore $\models [\rightrightarrows]\varphi \to \varphi$.
2. Assume $M_s \models \langle\rightrightarrows\rangle\langle\rightrightarrows\rangle\varphi$. Then there are M' and M'' with $s' \in \mathscr{D}(M')$ and $s'' \in \mathscr{D}(M'')$ such that $M_s \rightrightarrows M'_{s'}$ and $M'_{s'} \rightrightarrows M''_{s''}$, and such that $M''_{s''} \models \varphi$. From $M_s \rightrightarrows M'_{s'}$ and $M'_{s'} \rightrightarrows M''_{s''}$ we directly obtain $M_s \rightrightarrows M''_{s''}$ as the simulation relation is transitive. Therefore $M_s \models \langle\rightrightarrows\rangle\varphi$, so that $\models \langle\rightrightarrows\rangle\langle\rightrightarrows\rangle\varphi \to \langle\rightrightarrows\rangle\varphi$.
3. Assume that $M_s \models \langle\rightrightarrows\rangle[\rightrightarrows]\varphi$ Then, by the definition of semantics, there is an $M'_{s'}$ such that $M_s \rightrightarrows M'_{s'}$ and $M'_{s'} \models [\rightrightarrows]\varphi$. The latter implies, again by the definition of semantics, that $M^{\circ}_{s^{\circ}} \models \varphi$. Since M° is a simulation of every model, we have that $M''_{s''} \models \langle\rightrightarrows\rangle\varphi$ for an arbitrary $M''_{s''}$ such that $M_s \rightrightarrows M''_{s''}$. Since $M''_{s''}$ was arbitrary, the definition of the semantics implies that $M_s \models [\rightrightarrows]\langle\rightrightarrows\rangle\varphi$.
4. Assume that $M_s \models [\rightrightarrows]\langle\rightrightarrows\rangle\varphi$. From the definition of semantics and the reflexivity of the simulation relation we get $M^{\circ}_{s^{\circ}} \models \varphi$. As the mutual factual ignorance model is a simulation of itself, we trivially get $M^{\circ}_{s^{\circ}} \models [\rightrightarrows]\varphi$. Finally, since $M_s \rightrightarrows M^{\circ}_{s^{\circ}}$, we have that $M_s \models \langle\rightrightarrows\rangle[\rightrightarrows]\varphi$.

□

While refinement preserves the positive formulas, or hard knowledge [14], simulation preserves ignorance. Consider the following, *negative*, fragment $\mathscr{L}^{\Box}_{-}$ of $\mathscr{L}^{\Box}$:

$$\varphi ::= p \mid \neg p \mid \varphi \vee \psi \mid \varphi \wedge \varphi \mid \Diamond_a \varphi$$

Proposition 2 For all $\varphi \in \mathscr{L}^{\Box\rightrightarrows}_{-}$ and models M_s, if $M_s \models \varphi$ and $M_s \rightrightarrows M'_{s'}$, then $M'_{s'} \models \varphi$.

Proof The proof is by induction on the structure of φ, where the base case, $\varphi = p$ and $\varphi = \neg p$, is immediate from the fact that simulations preserve the valuation of the original states. Boolean cases follow by the induction hypothesis.

Case $\varphi = \Diamond_a \psi$. Assume that $M_s \models \Diamond_a \psi$. This means that there is an a-reachable state t such that $M_t \models \psi$. By the induction hypothesis, we have that $M'_{t'} \models \psi$. Finally, since there is a simulation $(s,s') \in Z$, we have that $(s',t') \in R'(a)$ with $M'_{t'} \models \psi$, and hence $M'_{s'} \models \Diamond_a \psi$. □

Mutual Factual Ignorance The concept of mutual factual ignorance works as an epistemic reset; it describes a state where no agent has any *factual* knowledge, and this lack of factual knowledge is *common* knowledge. Such a state is refered to as *mutual factual ignorance*. Reaching this state of information can be seen as a form of epistemic update: "If we were all to discount all our factual knowledge, then φ would be true".

As the mutual factual ignorance model M° serves as a kind of *tabula rasa* for agents, any model where agents have some factual information is a refinement of M°.

Proposition 3 Every model is a refinement of mutual factual ignorance model M°.

Proof Take an arbitrary model $M = (S, R, V)$. From the definition of M°, it is clear that we can identify every $s \in S$ with $V(s) \in S^\circ$ satisfying the same propositional variables. Having the refinement relation $Z = \{(V(s), s) \mid s \in S\}$, it is easy to check that M is a refinement of M°. □

As a direct corollary of Proposition 3, we have the fact that if a formula is true in all refinements of the mutual factual ignorance model, then the formula is valid.

Proposition 4 Given some φ, if for every $P' \subseteq P$, $M^\circ_{P'} \models [\leftrightarrows]\varphi$, then φ is valid.

All relations R°_a in the mutual factual ignorance model are total, and therefore equivalence relations. In the axiomatization **OML** of OML, this appears from the axioms involving the origin modality such as $[\circ](\Box_a\varphi \to \varphi)$, that formalizes that "given mutual factual ignorance, epistemic modalities $\Box_a$ are **S5**".

4 Axiomatizations

In this section we present the axiomatizations side by side and show their soundness and completeness.

4.1 Proof Systems

Modal Logic Let us start with the basic axiomatization for the language $\mathscr{L}^{\Box}$, which we call **ML**:

Prop	all substitution instances of tautologies of propositional logic
K	$\Box_a(\varphi \to \psi) \to (\Box_a\varphi \to \Box_a\psi)$
MP	from $\varphi \to \psi$ and φ infer ψ
N	from φ infer $\Box_a\varphi$
RE	from $\chi \leftrightarrow \psi$ infer $\varphi[\chi/p] \leftrightarrow \varphi[\psi/p]$

This forms a foundation for the axioms that follow and its soundness and completeness for the class of all Kripke models is well established (see for example [13]).

Refinement Modal Logic Let **REF** be the following set of axioms:

RQ1	$\langle\leftrightarrows\rangle\varphi_0 \leftrightarrow \varphi_0$	where $\varphi_0 \in \mathscr{L}_0$
RQ2	$\langle\leftrightarrows\rangle(\varphi \vee \psi) \leftrightarrow (\langle\leftrightarrows\rangle\varphi \vee \langle\leftrightarrows\rangle\psi)$	
RQ3	$\langle\leftrightarrows\rangle(\varphi_0 \wedge \varphi) \leftrightarrow (\varphi_0 \wedge \langle\leftrightarrows\rangle\varphi)$	where $\varphi_0 \in \mathscr{L}_0$
RQ4	$\langle\leftrightarrows\rangle \bigwedge_{a\in A} \nabla_a\Phi_a \leftrightarrow \bigwedge_{a\in A}\bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi$	

The axiomatization for refinement modal logic is then **RML = ML + REF**, where axioms and rules of **ML** are applied to $\mathscr{L}^{\Box\leftrightarrows}$. It is based on the axiomatization of arbitrary arrow update model logic in [25].

As an example, we show that rule **AR**: 'from $\varphi \to \psi$ infer $\langle\leftrightarrows\rangle\varphi \to \langle\leftrightarrows\rangle\psi$', is derivable.

1.	$\varphi \to \psi$	Given	2.	$(\varphi \vee \psi) \leftrightarrow \psi$	**Prop** and 1
3.	$\langle \leftleftarrows \rangle (\varphi \vee \psi) \leftrightarrow \langle \leftleftarrows \rangle \psi$	**RE**	4.	$\langle \leftleftarrows \rangle (\varphi \vee \psi) \leftrightarrow (\langle \leftleftarrows \rangle \varphi \vee \langle \leftleftarrows \rangle \psi)$	**RQ2**
5.	$(\langle \leftleftarrows \rangle \varphi \vee \langle \leftleftarrows \rangle \psi) \leftrightarrow \langle \leftleftarrows \rangle \psi$	**MP** and 3,4	6.	$\langle \leftleftarrows \rangle \varphi \to \langle \leftleftarrows \rangle \psi$	**Prop** and 5

Origin Modal Logic Origin modal logic is quite novel, and correspondingly, has a novel set of axioms. For origin modal logic we start with the set **ORI**, given by

O1	$[\circ]\varphi_0 \leftrightarrow \varphi_0$	where $\varphi_0 \in \mathscr{L}_0$
OT	$[\circ](\Box_a \varphi \to \varphi)$	
O5	$[\circ](\Diamond_a \varphi \to \Box_a \Diamond_a \varphi)$	
OExch	$[\circ](\Box_a \varphi \to \Box_b \varphi)$	
OFull	$[\circ]\Diamond_a \varphi$ where φ is of the form $\bigwedge_{p \in Q_1} p \wedge \bigwedge_{p \in Q_2} \neg p$ with $Q_1 \cap Q_2 = \emptyset$	
ODual	$[\circ]\neg\varphi \leftrightarrow \neg[\circ]\varphi$	
ODisj	$[\circ](\varphi \vee \psi) \leftrightarrow ([\circ]\varphi \vee [\circ]\psi)$	
OMP	from $[\circ](\varphi \to \psi)$ and $[\circ]\varphi$ infer $[\circ]\psi$	
ON	from $[\circ]\varphi$ infer $[\circ]\Box_a \varphi$	

and obtain **OML = ML + ORI**, where axioms and rules of **ML** are applied to $\mathscr{L}^{\Box\circ}$.

Observe that necessitation for the origin modality, i.e.'from φ infer $[\circ]\varphi$', is derivable in **OML**.

1.	φ	Given	2.	$\varphi \leftrightarrow (p \vee \neg p)$	1
3.	$[\circ]p \vee \neg[\circ]p$	**Prop**	4.	$[\circ]p \vee [\circ]\neg p$	**ODual** and 3
5.	$[\circ](p \vee \neg p)$	**ODisj** and 4	6.	$[\circ]\varphi$	**RE** and 5

Simulation Modal Logic We report two ways to axiomatize the logic with simulation quantifiers. First, one can use refinement quantifiers and the origin modality to ensure that all formulas in a cover set are satisfiable in a refinement of the mutual factual ignorance model (see axiom **SQ4** below). Then, let **SIM** be

SQ1	$\langle \rightrightarrows \rangle \varphi_0 \leftrightarrow \varphi_0$	where $\varphi_0 \in \mathscr{L}_0$
SQ2	$\langle \rightrightarrows \rangle (\varphi \vee \psi) \leftrightarrow (\langle \rightrightarrows \rangle \varphi \vee \langle \rightrightarrows \rangle \psi)$	
SQ3	$\langle \rightrightarrows \rangle (\varphi_0 \wedge \varphi) \leftrightarrow (\varphi_0 \wedge \langle \rightrightarrows \rangle \varphi)$	where $\varphi_0 \in \mathscr{L}_0$
SQ4	$\langle \rightrightarrows \rangle \bigwedge_{a \in A} \nabla_a \Phi_a \leftrightarrow \bigwedge_{a \in A} (\Box_a \bigvee_{\varphi \in \Phi_a} \langle \rightrightarrows \rangle \varphi \wedge [\circ] \bigwedge_{\varphi \in \Phi_a} \Diamond_a \langle \leftleftarrows \rangle \varphi)$	

and **ROSML = ML + REF + ORI + SIM**, where all of the axioms and rules are applied to $\mathscr{L}$.

Another way to axiomatize simulation quantifiers is to explicitly require all the formulas in a cover set to be consistent. In this way, we do not need refinement quantifiers and the origin modality, but we have to pay for this with the higher complexity of determining whether or not an axiom may be faithfully applied (in this case, **PSPACE** for the satisfiability problem of modal logic K [13]). So, let $\mathbf{SIM}_{\mathsf{cons}}$ be **SIM** where we replace the axiom **SQ4** by

$\mathbf{SQ4}_{\mathsf{cons}}$ $\langle \rightrightarrows \rangle \bigwedge_{a \in A} \nabla_a \Phi_a \leftrightarrow \bigwedge_{a \in A} \Box_a \bigvee_{\varphi \in \Phi_a} \langle \rightrightarrows \rangle \varphi$ where all φ are consistent $\mathscr{L}^{\Box}$ formulas

then $\mathbf{SML = ML + SIM}_{\mathsf{cons}}$, where axioms and rules of **ML** are applied to $\mathscr{L}^{\Box\rightrightarrows}$. Axiom $\mathbf{SQ4}_{\mathsf{cons}}$ is inspired by axiom **CCRKco2** in [44], that also has the consistency requirement.

Note the similarity between **SML** and **RML**. All but axiom $\mathbf{SQ4}_{\mathsf{cons}}$ simply replaced $\langle \leftleftarrows \rangle$ modalities by $\langle \rightrightarrows \rangle$ modalities. Intuitively the difference is clear: **RQ4** says that if after refinement we can cover all the formulas in Φ_a, then before the refinement all those must have been possible but it may no longer be a

cover, so we lose the $\Box$. Dually, $\mathbf{SQ4}_{\mathsf{cons}}$ says that if after simulation we can cover all the formulas in Φ_a, then before the simulation necessarily one of those must already have been reachable by a simulation. However, there is also a consistency requirement there, without which the axiom is unsound. Consider a single agent. We then get

$$\langle\rightrightarrows\rangle(\bigwedge_{\varphi\in\Phi}\Diamond\varphi\wedge\Box\bigvee_{\varphi\in\Phi}\varphi)\leftrightarrow\Box\bigvee_{\varphi\in\Phi}\langle\rightrightarrows\rangle\varphi.$$

The simplest counterexample to soundness is the set $\Phi=\{\top,\bot\}$, noting that $\Box\top\leftrightarrow\top$. We get

$$\langle\rightrightarrows\rangle(\Diamond\top\wedge\Diamond\bot)\leftrightarrow\Box(\langle\rightrightarrows\rangle\top\vee\langle\rightrightarrows\rangle\bot).$$

Since the right-hand side of the equivalence requires that just one of the elements of Φ needs to be true at every reachable state, it suffices to observe that $\top$ is true everywhere, so that the disjunction is then also true. But then this also requires the left-hand side of the equivalence to be true: there is a simulation where Φ covers the successors, so there must be an accessible world where $\bot$ is true, which is an absurdity.

4.2 Soundness

In this section, we show that each of the axiom systems is sound for its logic. Soundness of **ML** is well known, so we will not comment further on it here. Let us therefore consider the other axiom systems.

Proposition 5 The axiomatization **RML** is sound for $\mathscr{L}^{\Box\leftleftarrows}$.

Proof We have $\mathbf{RML}=\mathbf{ML}+\mathbf{REF}$, and soundness of the axioms and rules of **ML** is well known. What remains to show is that **REF** is sound.

Refinements satisfy **atoms**, and therefore do not change the truth value of propositional formulas. Hence $\models\langle\leftleftarrows\rangle\varphi_0\leftrightarrow\varphi_0$, so **RQ1** is sound.

Next, note that $\langle\leftleftarrows\rangle$ is a "diamond-like" operator, in that $\langle\leftleftarrows\rangle\varphi$ holds iff there is at least one refinement after which φ is true. Because it is diamond-like, $\langle\leftleftarrows\rangle$ distributes over disjunctions, so $\models\langle\leftleftarrows\rangle(\varphi\vee\psi)\leftrightarrow(\langle\leftleftarrows\rangle\varphi\vee\langle\leftleftarrows\rangle\psi)$, which means that **RQ2** is sound.

For **RQ3**, we start with the left-to-right direction, so suppose that $M_s\models\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)$, where $\varphi_0\in\mathscr{L}_0$. Then there is some $M'_{s'}$ such that $M_s\leftleftarrows M'_{s'}$ and $M'_{s'}\models\varphi_0\wedge\varphi$. As such, we have $M'_{s'}\models\varphi_0$ and $M'_{s'}\models\varphi$, which imply that $M_s\models\langle\leftleftarrows\rangle\varphi_0$ and $M_s\models\langle\leftleftarrows\rangle\varphi$, respectively. Furthermore, as shown above, $M_s\models\langle\leftleftarrows\rangle\varphi_0\leftrightarrow\varphi_0$. Hence $\models\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)\rightarrow(\varphi_0\wedge\langle\leftleftarrows\rangle\varphi)$.

Suppose then, for the right-to-left direction, that $M_s\models\varphi_0\wedge\langle\leftleftarrows\rangle\varphi$. Then $M_s\models\langle\leftleftarrows\rangle\varphi$, so there is some $M'_{s'}$ such that $M_s\leftleftarrows M'_{s'}$ and $M'_{s'}\models\varphi$. Because every refinement satisfies **atoms**, $M_s\leftleftarrows M'_{s'}$ together with $M_s\models\varphi_0$ implies that $M'_{s'}\models\varphi_0$. So $M'_{s'}\models\varphi_0\wedge\varphi$, and therefore $M_s\models\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)$. So $\models(\varphi_0\wedge\langle\leftleftarrows\rangle\varphi)\rightarrow\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)$. Together with the previously shown $\models\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)\rightarrow(\varphi_0\wedge\langle\leftleftarrows\rangle\varphi)$ this implies that $\models\langle\leftleftarrows\rangle(\varphi_0\wedge\varphi)\leftrightarrow(\varphi_0\wedge\langle\leftleftarrows\rangle\varphi)$, so **RQ3** is sound.

For **RQ4** we also start with the left-to-right direction, so suppose that $M_{s_1}\models\langle\leftleftarrows\rangle\bigwedge_{a\in A}\nabla_a\Phi_a$. Then there is some $M'_{s'_1}$ such that $M_{s_1}\leftleftarrows M'_{s'_1}$ and $M'_{s'_1}\models\bigwedge_{a\in A}\nabla_a\Phi_a$. This implies that, in particular, for every $a\in A$ and $\varphi\in\Phi_a$, there is some s'_2 such that $(s'_1,s'_2)\in R'_a$ and $M'_{s'_2}\models\varphi$. Because $M_{s_1}\leftleftarrows M'_{s'_1}$ and every refinement satisfies the **back** condition, there must be some s_2 such that $(s_1,s_2)\in R_a$ and $M_{s_2}\leftleftarrows M'_{s'_2}$. We then have $M_{s_2}\models\langle\leftleftarrows\rangle\varphi$, and therefore $M_{s_1}\models\Diamond_a\langle\rightrightarrows\rangle\varphi$. This holds for every $a\in A$ and $\varphi\in\Phi_a$, so $M_{s_1}\models\bigwedge_{a\in A}\bigwedge_{\varphi\in\Phi_a}\Diamond_a\langle\leftleftarrows\rangle\varphi$. We have now shown that $\models\langle\leftleftarrows\rangle\bigwedge_{a\in A}\nabla_a\Phi_a\rightarrow\bigwedge_{a\in A}\bigwedge_{\varphi\in\Phi_a}\Diamond_a\langle\leftleftarrows\rangle\varphi$.

For the right-to-left direction, suppose that $M_{s_0}\models\bigwedge_{a\in A}\bigwedge_{\varphi\in\Phi_a}\Diamond_a\langle\leftleftarrows\rangle\varphi$. So for every $a\in A$ and $\varphi\in\Phi_a$ there is some s_φ such that $(s_0,s_\varphi)\in R_a$ and $M_{s_\varphi}\models\langle\leftleftarrows\rangle\varphi$. So there are $M^\varphi=(S^\varphi,R^\varphi,V^\varphi)$ and

s'_φ such that $M_{s_\varphi} \leftrightarrows M^\varphi_{s'_\varphi}$ and $M^\varphi_{s'_\varphi} \models \varphi$. Now, let M' be the disjoint union of all M^φ, with one additional state s'_0, where we take $V'(s'_0) = V(s_0)$, and additional a-edges from s'_0 to s'_φ for all $a \in A$ and $\varphi \in \Phi_a$. Formally, $M' = (S', R', V')$, where

- $S' = \{s'_0\} \cup \bigcup_{a\in A} \bigcup_{\varphi\in\Phi_a} \{(s', a, \varphi) \mid s' \in S_\varphi\}$,
- $R'_a = \{(s'_0, (s'_\varphi, a, \varphi)) \mid a \in A, \varphi \in \Phi_A\} \cup \bigcup_{a\in A} \bigcup_{\varphi\in\Phi_a} \{((s'_1, a, \varphi), (s'_2, a, \varphi)) \mid (s'_1, s'_2) \in R^\varphi_a\}$,
- $V'(s') = \begin{cases} V(s_0) & \text{if } s' = s'_0 \\ V^\varphi(t') & \text{if } s' = (t', a, \varphi) \end{cases}$

See Figure 1 for an example of this construction.

For every $\varphi \in \Phi_a$, there is some s' such that $(s'_0, s') \in R'_a$ and $M'_{s'} \models \varphi$, namely $s' = (s'_\varphi, a, \varphi)$. Conversely, for every s' such that $(s'_0, s') \in R'_a$ there is some $\varphi \in \Phi_a$ such that $M'_{s'} \models \varphi$, because $(s'_0, s') \in R'_a$ implies that s' is of the form $s' = (s'_\varphi, a, \varphi)$. This implies that $M'_{s'_0} \models \bigwedge_{a\in A} \nabla_a \Phi_a$.

Furthermore, for every $a \in A$ and $\varphi \in \Phi_a$ we have $M_{s_\varphi} \leftrightarrows M^\varphi_{s'_\varphi} \models \varphi$. Let Z_φ be the witnessing refinement, and let Z be the disjoint union of all Z_φ plus the pair (s_0, s'_0). Formally, $Z_\varphi = (s_0, s'_0) \cup \bigcup_{a\in A} \bigcup_{\varphi\in\Phi_a} \{(s, (s', a, \varphi)) \mid (s, s') \in Z_\varphi\}$. We will show that this Z is a refinement.

Take any $(s, s') \in Z$. We consider two cases. Firstly, if $s' = s'_0$, then $s = s_0$. We chose $V'(s'_0) = V(s_0)$, so in this case **atoms** is satisfied. Secondly, if $s' \neq s'_0$ then s' is of the form $s' = (t', a, \varphi)$, and we have $(s, t') \in Z^\varphi$, which implies that $V(s) = V^\varphi(t')$ because Z^φ satisfies **atoms**. We took $V'(s') = V^\varphi(t')$, so from $V(s) = V^\varphi(t')$ it follows that $V(s) = V'(s')$, so **atoms** holds for s and s'. These two cases are exhaustive, so **atoms** is satisfied for Z.

Then take any $(s_1, s'_1) \in Z$ and $(s'_1, s'_2) \in R'_a$. Again, we consider two cases. Firstly, if $s'_1 = s'_0$, then $s_1 = s_0$ and s'_2 is of the form $s'_2 = (s'_\varphi, a, \varphi)$. We have $(s_\varphi, s'_\varphi) \in Z^\varphi$ which, by the construction of Z, implies that $(s_2, s'_2) = (s_\varphi, (s'_\varphi, a, \varphi)) \in Z$. Furthermore, $(s_1, s_2) = (s_0, s_\varphi) \in R_a$, so **back** is satisfied in this case. Secondly, if $s'_1 \neq s'_0$, then s'_1 and s'_2 are of the form $s'_1 = (t'_1, b, \varphi)$ and $s'_2 = (t'_2, b, \varphi)$, for some b and φ. We then also have, by the construction of Z and M', that $(t'_1, t'_2) \in R^\varphi$ and $(s_1, t'_1) \in Z^\varphi$. The **back** condition for Z^φ therefore implies that there is some s_2 such that $(s_1, s_2) \in R_a$ and $(s_2, t'_2) \in Z^\varphi$. This implies that $(s_1, s_2) \in R_a$ and $(s_2, s'_2) = (s_2, (t'_2, b, \varphi)) \in Z$, so **back** is satisfied in this case as well. The two cases are exhaustive, so **back** is satisfied for Z.

Z is therefore a refinement, which implies that $M_{s_0} \leftrightarrows M'_{s'_0}$. Together with the previous conclusion that $M'_{s'_0} \models \bigwedge_{a\in A} \nabla_a \Phi_a$, this implies that $M_{s_0} \models \langle\leftrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a$. We have now shown that

$$\models \bigwedge_{a\in A} \bigwedge_{\varphi\in\Phi_a} \Diamond_a \langle\leftrightarrows\rangle \varphi \to \langle\leftrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a,$$

which together with the previously shown

$$\models \langle\leftrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a \to \bigwedge_{a\in A} \bigwedge_{\varphi\in\Phi_a} \Diamond_a \langle\leftrightarrows\rangle \varphi$$

implies that

$$\models \langle\leftrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a \leftrightarrow \bigwedge_{a\in A} \bigwedge_{\varphi\in\Phi_a} \Diamond_a \langle\leftrightarrows\rangle \varphi.$$

So **RQ4** is sound. □

The soundness of **RQ4** is shown rather differently than that of the similar axiom **A4** in [25], involving quantifiers over arrow updates. However, the proof in the current setting has nice duality with that of the soundness of **SQ4**, coming up (almost) next, in Proposition 7.

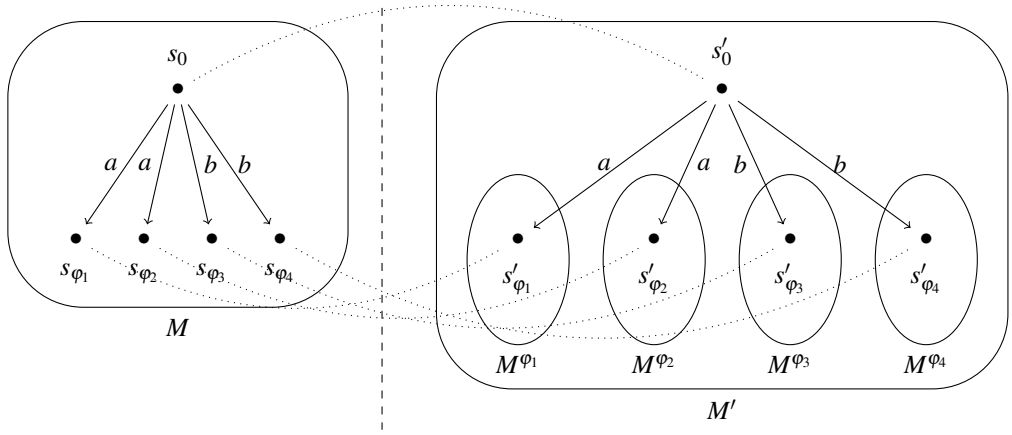


Figure 1: Schematic drawing of an example of M' as in the proof of Proposition 5. In this example, $A = \{a,b\}$, $\Phi_a = \{\varphi_1, \varphi_2\}$ and $\Phi_b = \{\varphi_3, \varphi_4\}$. Dotted lines represent the relation Z. The state s'_0 has exactly the successors $s'_{\varphi_1}, \cdots, s'_{\varphi_4}$, the state s_0 has at least the successors $s_{\varphi_1}, \cdots, s_{\varphi_4}$ but may have more. Hence Z is a refinement but not, in general, a bisimulation.

Proposition 6 The axiomatization **OML** is sound for $\mathscr{L}^{\Box\circ}$.

Proof We have $\mathbf{OML} = \mathbf{ML} + \mathbf{ORI}$, and the soundness of **ML** is well known, what remains to show is that **ORI** is sound.

To begin with, for every M_s and every $\varphi_0 \in \mathscr{L}_0$, we have $M_s \models \varphi_0 \Leftrightarrow M^\circ_{V(s)} \models \varphi_0 \Leftrightarrow M_s \models [\circ]\varphi_0$, so $\models [\circ]\varphi_0 \leftrightarrow \varphi_0$, which means that **O1** is sound.

The model M° is an S5-model, i.e., all relations in M° are equivalence relations. It follows that, for every $t \in S^\circ$, $M^\circ_t \models \Box_a\varphi \to \varphi$ and $M^\circ_t \models \Diamond_a\varphi \to \Box_a\Diamond_a\varphi$. This implies that, for all M_s, we have $M^\circ_{V(s)} \models \Box_a\varphi \to \varphi$ and $M^\circ_{V(s)} \models \Diamond_a\varphi \to \Box_a\Diamond_a\varphi$, and hence $M_s \models [\circ](\Box_a\varphi \to \varphi)$ and $M_s \models [\circ](\Diamond_a\varphi \to \Box_a\Diamond_a\varphi)$. As this is true for every M_s, we have $\models [\circ](\Box_a\varphi \to \varphi)$ and $\models [\circ](\Diamond_a\varphi \to \Box_a\Diamond_a\varphi)$, so **OT** and **O5** are sound.

Furthermore, all agents $a,b \in A$ have the same accessibility relation on M°, namely $R^\circ_a = R^\circ_b = S^\circ \times S^\circ$. This implies that $M^\circ_t \models \Box_a\varphi \to \Box_b\varphi$, for every t, and therefore $\models [\circ](\Box_a\varphi \to \Box_b\varphi)$. So **OExch** is sound.

Next, consider $Q_1, Q_2 \subseteq P$ such that $Q_1 \cap Q_2 = \emptyset$. Then there is at least one valuation t such that $M^\circ_t \models \bigwedge_{p\in Q_1} p \wedge \bigwedge_{p\in Q_2} \neg p$. Because $R^\circ_a = S^\circ \times S^\circ$, it follows that $M^\circ_{t'} \models \Diamond_a(\bigwedge_{p\in Q_1} p \wedge \bigwedge_{p\in Q_2} \neg p)$ for all t', and therefore also that $\models [\circ]\Diamond_a(\bigwedge_{p\in Q_1} p \wedge \bigwedge_{p\in Q_2} \neg p)$. So **OFull** is sound.

We have $M_s \models [\circ]\neg\varphi \Leftrightarrow M^\circ_{V(s)} \models \neg\varphi \Leftrightarrow M^\circ_{V(s)} \not\models \varphi \Leftrightarrow M_s \not\models [\circ]\varphi \Leftrightarrow M_s \models \neg[\circ]\varphi$. It follows that $\models [\circ]\neg\varphi \leftrightarrow \neg[\circ]\varphi$, so **ODual** is sound.

Similarly, $M_s \models [\circ](\varphi \vee \psi) \Leftrightarrow M^\circ_{V(s)} \models \varphi \vee \psi \Leftrightarrow (M^\circ_{V(s)} \models \varphi$ or $M^\circ_{V(s)} \models \psi) \Leftrightarrow (M_s \models [\circ]\varphi$ or $M_s \models [\circ]\psi) \Leftrightarrow M_s \models [\circ]\varphi \vee [\circ]\psi$. It follows that $\models [\circ](\varphi \vee \psi) \leftrightarrow ([\circ]\varphi \vee [\circ]\psi)$, so **ODisj** is sound.

This leaves the two rules, we show that they preserve validity. So suppose that $\models [\circ](\varphi \to \psi)$ and $\models [\circ]\psi$. Then for every M_s, we have $M_s \models [\circ](\varphi \to \psi)$ and $M_s \models [\circ]\psi$. So $M^\circ_{V(s)} \models \varphi \to \psi$ and $M^\circ_{V(s)} \models \varphi$, which implies that $M^\circ_{V(s)} \models \psi$ and therefore $M_s \models [\circ]\psi$. Since this holds for every M_s, we then have $\models [\circ]\psi$, proving the soundness of **OMP**.

Finally, suppose that $\models [\circ]\varphi$. Then for every M_s, we have $M_s \models [\circ]\varphi$ and therefore $M^\circ_{V(s)} \models \varphi$. Since this holds for every M_s, and therefore also every $V(s)$, we also have $M^\circ_{V(s')} \models \varphi$ for every $V(s')$, which

implies that $M^{\circ}_{V(s)} \models \Box_a \varphi$, and therefore $M_s \models [\circ]\Box_a \varphi$. Since this holds for every M_s, we then have $\models [\circ]\Box_a \varphi$. So **ON** is sound. □

Proposition 7 The axiomatization **SML** is sound for $\mathscr{L}^{\Box\rightrightarrows}$.

Proof We have $\mathbf{SML} = \mathbf{ML} + \mathbf{SIM_{cons}}$. Soundness of **ML** is well known, what remains to be shown is the soundness of $\mathbf{SIM_{cons}}$. Axioms **SQ1**, **SQ2**, and **SQ3** are sound for the same reasons that **RQ1**, **RQ2** and **RQ3** are: preservation of propositional logic under simulations for **SQ1** and **SQ3**, and $\langle\rightrightarrows\rangle$ being "diamond-like" for **SQ2**. This leaves $\mathbf{SQ4_{cons}}$.

For the left-to-right direction, suppose that $M_s \models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a$.[1] Then there is some $M'_{s'}$ such that $M_s \rightrightarrows M'_{s'}$ and $M'_{s'} \models \bigwedge_{a\in A} \nabla_a \Phi_a$. Take any t such that $(s,t) \in R_a$. Because $M_s \rightrightarrows M'_{s'}$, there is some t' such that $(s',t') \in R'_a$ and $M_t \rightrightarrows M'_{t'}$.

Now, because $M'_{s'} \models \bigwedge_{a\in A} \nabla_a \Phi_a$, there is some $\varphi \in \Phi_a$ such that $M'_{t'} \models \varphi$, and therefore $M_t \models \langle\rightrightarrows\rangle\varphi$. This holds for every $(s,t) \in R_a$ and every $a \in A$, so we have $M_s \models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi$. We have now shown that

$$\models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a \to \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi.$$

For right-to-left direction, suppose that $M_{s_0} \models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi$, where all φ are consistent. Take any $(s,t) \in R_a$. Because $M_s \models \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi$, there is some $\varphi \in \Phi_a$ such that $M_t \models \langle\rightrightarrows\rangle\varphi$. Then there are some model M^t and state t' of M^t such that $M_t \rightrightarrows M^t_{t'}$ and $M^t_{t'} \models \varphi$.[2]

Every $M^t_{t'}$ satisfies some $\varphi \in \Phi_a$, but there is no guarantee that every $\varphi \in \Phi_a$ is satisfied in such a model. Let $\Phi^+_a \subseteq \Phi_a$ be the set of formulas that are so satisfied, and $\Phi^-_a = \Phi_a \setminus \Phi^+_a$. Because all $\varphi \in \Phi_a$ are, by assumption, consistent, there exist models M^φ and t_φ such that $M^\varphi_{t_\varphi} \models \varphi$.

Now, let M' be the disjoint union of (i) the $M^t_{t'}$ for each $a \in A$ and $(s,t) \in R_a$ and (ii) the $M^\varphi_{t'_\varphi}$ for each $\varphi \in \Phi^-_a$. Add a single state s'_0, with $V'(s'_0) = V(s)$, and a-edges from s'_0 to every $M^t_{t'}$ where $(s,t) \in R_a$ and every $M^\varphi_{t'_\varphi}$ where $\varphi \in \Phi^-_a$. Call the resulting model $M' = (S', R', V')$. Formally,

- $S' = \{s'_0\} \cup \bigcup_{(s,t)\in R_a} \{(s',a,t) \mid s' \in S^t\} \cup \bigcup_{\varphi\in\Phi^-_a} \{(s',a,\varphi) \mid s' \in S^\varphi\}$
- $R'_a = \{(s'_0,(t',a,t)) \mid (s,t) \in R_a\} \cup \{(s'_0,(s'_\varphi,a,\varphi)) \mid \varphi \in \Phi^-_a\} \cup$
  $\bigcup_{a\in A} \bigcup_{(s,t)\in R_a} \{((s'_1,a,t),(s'_2,a,t)) \mid (s'_1,s'_2) \in R^t_a\} \cup$
  $\bigcup_{a\in A} \bigcup_{\varphi\in\Phi^-_a} \{((s'_1,a,\varphi),(s'_2,a,\varphi)) \mid (s'_1,s'_2) \in R^\varphi_a\}$
- $V'(s') = \begin{cases} V(s_0) & \text{if } s' = s'_0 \\ V^t(t') & \text{if } s' = (t',a,t) \\ V^\varphi(t') & \text{if } s' = (t',a,\varphi) \end{cases}$

See Figure 2 for an example of this construction.

For every $(s'_0,s') \in R'_a$, we have $s' = (t',a,t)$ or $s' = (s'_\varphi,a,\varphi)$. In the first case, there is some $\varphi \in \Phi^+_a$ such that $M'_{s'} \models \varphi$, since $M'_{s'}$ is a copy of $M^t_{t'}$. In the second case, there is some $\varphi \in \Phi^-_a$ such that $M'_{s'} \models \varphi$, since $M'_{s'}$ is a copy of $M^\varphi_{s'_\varphi}$. So every a-successor of s'_0 satisfies some $\varphi \in \Phi_a$.

Furthermore, for every $\varphi \in \Phi^+_a$ there is some t such that $(s_0,t) \in R_a$ and $M^t_{t'} \models \varphi$, which implies that $M'_{(t',a,t)} \models \varphi$ and therefore that $M'_{s'_0} \models \Diamond_a \varphi$. For very $\varphi \in \Phi^-_a$, we have $M^\varphi_{s'_\varphi} \models \varphi$, which implies that

[1] Note that we are not, at this point, explicitly assuming that all $\varphi \in \Phi_a$ are consistent. This is because $M_s \models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a$ implies that all φ are consistent. The separate consistency condition in $\mathbf{SQ4_{cons}}$ is only required for the right-to-left direction.

[2] Note that in $M^t_{t'}$ the superscript t refers to a state of M, but here serves only to identify a model M^t. The subscript t', meanwhile, identifies a particular state of M^t, namely the one such that $M_t \rightrightarrows M^t_{t'}$.

$M'_{(s'_\varphi,a,\varphi)} \models \varphi$ and therefore that $M'_{s'_0} \models \Diamond_a \varphi$. Taken together, this shows that every $\varphi \in \Phi_a$ is satisfied in some a-successor of s'_0. We had already shown that every a-successor of s'_0 satisfies some $\varphi \in \Phi_a$, so we have $M'_{s'_0} \models \bigwedge_{a\in A} \nabla_a \Phi_a$.

Now, we will show that $M_{s_0} \rightrightarrows M'_{s'_0}$. For each $M^t_{t'}$, we have $M_t \rightrightarrows M^t_{t'}$, let us call the witnessing simulation Z^t. Now, let Z be the disjoint union of all Z^t together with (s_0, s'_0). Formally,

$$Z = \{(s, s'_0)\} \cup \bigcup_{a\in A} \bigcup_{(s_0,t)\in R_a} \{(s,(s',a,t)) \mid (s,s') \in Z^t\}.$$

We claim that Z is a simulation.

Take any $(x,x') \in Z$. We consider two cases. Firstly, suppose that $x' = s'_0$. Then $x = s_0$. We have $V(s_0) = V'(s'_0)$ by construction, so in this case **atoms** is satisfied. Secondly, suppose $x' \neq s'_0$. Then x' must be of the form (s',a,t), since Z does not relate to any states of the form (s',a,φ). By the construction of Z, we then have $(x,s') \in Z$, which implies that $V(x) = V^t(s')$, and therefore also $V(x) = V'(x')$. So in this case **atoms** is also satisfied. These two cases are exhaustive, so **atoms** is satisfied for Z.

Then take any $(x_1,x'_1) \in Z$ and $(x_1,x_2) \in R_a$. Again, we consider two cases. Firstly, suppose $x'_1 = s'_0$. Then $x_1 = s_0$, and $x_2 = t$ for some t such that $(s,t) \in R_a$. Then there is some $M^t_{t'}$ such that $(t,t') \in Z^t$. By the construction of Z we have $(t,(t',a,t)) \in Z$, and by the construction of M' we have $(s'_0,(t',a,t)) \in R_a$. So **forth** is satisfied in this case. Secondly, suppose $x'_1 \neq s'_0$. Then x'_1 is of the form $x'_1 = (s'_1,b,t)$, where $(x_1,s'_1) \in Z^t$. Because Z^t is a simulation, there is some s'_2 such that $(s'_1,s'_2) \in R^t_a$ and $(x_2,s'_2) \in Z^t$. We then also have $(s_1,(s'_2,b,t)) \in R'_a$ and $(x_2,(s'_2,b,t)) \in Z$, so **forth** is satisfied in this case. The two cases are exhaustive, so **forth** is satisfied for Z.

Because Z satisfies **atoms** and **forth**, it is a simulation, so $M_s \rightrightarrows M'_{s'_0}$. Together with the previously shown $M'_{s'_0} \models \bigwedge_{a\in A} \nabla_a \Phi_a$, this implies that $M_{s_0} \models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a$. This completes the proof that

$$\models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi \to \bigwedge_{a\in A} \nabla_a \Phi_a,$$

if all $\varphi \in \Phi_a$ are consistent. We had already shown the other direction, so

$$\models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a \to \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi,$$

when all φ are consistent. This shows that $\mathbf{SQ4_{cons}}$ is sound, which completes the soundness proof for **SML**. □

Proposition 8 The axiomatization **ROSML** is sound for $\mathscr{L}$.

Proof $\mathbf{ROSML} = \mathbf{ML} + \mathbf{REF} + \mathbf{ORI} + \mathbf{SIM}$. Soundness of **ML** is well known, and we showed soundness of **REF**, **ORI** and **SQ1–SQ3** previously. This leaves only the soundness of **SQ4**.

For the left-to-right direction, suppose that $M_s \models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a \Phi_a$. Then each $\varphi \in \Phi_a$ is consistent, so by the soundness of $\mathbf{SQ4_{cons}}$ we have $M_s \models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a} \langle\rightrightarrows\rangle\varphi$. Furthermore, because φ is consistent, there are M^φ and s_φ such that $M^\varphi_{s_\varphi} \models \varphi$.

Every pointed model is the refinement of the state of M° with the same valuation, so we have $M^\circ_{V(s_\varphi)} \leftrightarrows M^\varphi_{s_\varphi}$. It follows that $M^\circ_{V(s_\varphi)} \models \langle\leftrightarrows\rangle\varphi$. In turn, that implies that $M^\circ_{V(s)} \models \Diamond_a\langle\leftrightarrows\rangle\varphi$, since every state of M° is an a-successor of every other state. This is true for every $\varphi \in \Phi_a$, so $M^\circ_{V(s)} \models$

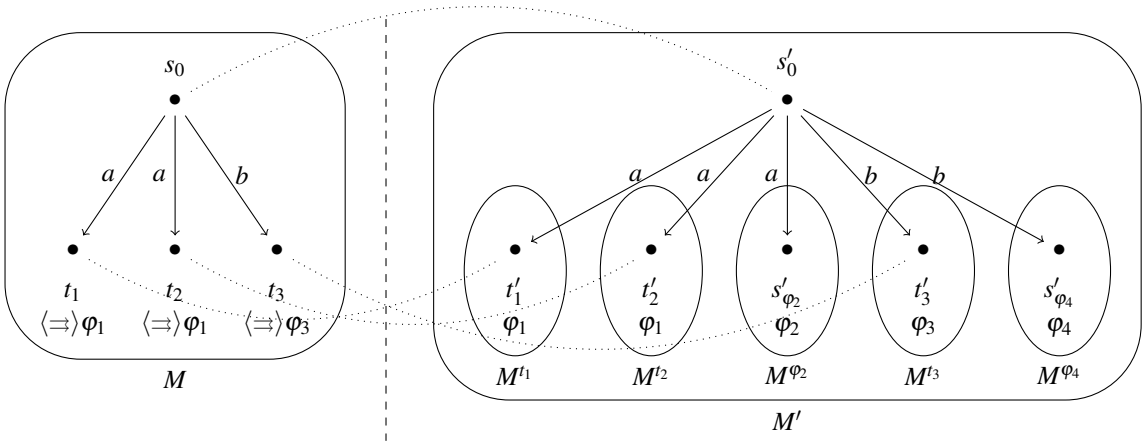


Figure 2: Schematic drawing of an example of M' as in the proof of Proposition 7. Here, we take $A = \{a,b\}$, $\Phi_a = \{\varphi_1, \varphi_2\}$ and $\Phi_b = \{\varphi_3, \varphi_4\}$. For both s_0 and s'_0, all successors are drawn. In each successor of s_0, $\langle\rightrightarrows\rangle\varphi_i$ holds for some i. Note that the same $\langle\rightrightarrows\rangle\varphi_i$ may hold in multiple successors, such as $\langle\rightrightarrows\rangle\varphi_1$ in t_1 and t_2. Each successor of s'_0 corresponds either to a successor of s_0 or to some $\varphi_i \in \Phi_a^- \cup \Phi_b^-$. Dotted lines represent the relation Z. Note that every successor of s_0 is Z-related to some successor of s'_0 but not vice versa, so Z is a simulation but not a bisimulation.

$\bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi$, which implies that $M_s \models [\circ]\bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi$. Finally, this is true for every $a \in A$, which together with the previous conclusion that $M_s \models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi$ implies that

$$M_s \models \bigwedge_{a\in A}(\Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi \wedge [\circ] \bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi).$$

We have now shown that

$$\models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a\Phi_a \to \bigwedge_{a\in A}(\Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi \wedge [\circ] \bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi).$$

For the right-to-left direction, suppose that $M_s \models \bigwedge_{a\in A}(\Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi \wedge [\circ]\bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi)$. Then $M_s \models \bigwedge_{a\in A} \Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi$ and, for every $a \in A$, $M_s \models [\circ]\bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi$. The latter implies that φ holds in some refinement of some state of M°, so, in particular, φ is satisfiable. From the soundness of $\mathbf{SQ4_{cons}}$, it then follows that $M_s \models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a\Phi_a$. We have now shown that

$$\models \bigwedge_{a\in A}(\Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi \wedge [\circ] \bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi) \to \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a\Phi_a.$$

We already showed the other direction, so we have

$$\models \langle\rightrightarrows\rangle \bigwedge_{a\in A} \nabla_a\Phi_a \leftrightarrow \bigwedge_{a\in A}(\Box_a \bigvee_{\varphi\in\Phi_a}\langle\rightrightarrows\rangle\varphi \wedge [\circ] \bigwedge_{\varphi\in\Phi_a} \Diamond_a\langle\leftrightarrows\rangle\varphi).$$

This shows that **SQ4** is sound, which completes the soundness proof for **ROSML**. □

4.3 Completeness

Having shown the soundness of the reduction axioms for **OML**, **RML**, **SML**, and **ROSML**, we can now establish the completeness of the corresponding axiomatizations. In all of the following completeness proofs, the general idea is to use the reduction axioms to translate a given formula of one of the new logics into an equivalent formula of **ML**, which is known to be complete.

Proposition 9 The axiomatization **RML** is complete for $\mathscr{L}^{\Box\leftleftarrows}$.

Proof Given $\langle\leftleftarrows\rangle\varphi$, the proof is by induction on the modal depth of φ. For the base case, $d(\varphi)=0$, and hence $\langle\leftleftarrows\rangle\varphi_0$ is equivalent to φ_0 by **RQ1**. Now assume that we have $\langle\leftleftarrows\rangle\varphi$ with $d(\varphi)=n$, and all $\langle\leftleftarrows\rangle\psi$ with $d(\psi)<n$ are equivalent to some formulas without refinement quantifiers. W.l.o.g. we may also assume that φ is a combination of propositional atoms and formulas of the type $\Box_a\chi_1$ and $\Diamond_a\chi_2$. By **RE**, we can substitute φ with an equivalent formula in the disjunctive normal form:

$$\bigvee\left(\begin{array}{l} p_{11}\wedge\ldots\wedge p_{1m}\wedge\bigwedge_{a\in ag(\varphi)}(\Diamond_a\chi_1\wedge\ldots\wedge\Diamond_a\chi_l\wedge\Box_a\theta_1\wedge\ldots\wedge\Box_a\theta_k)\\ \ldots\\ p_{n1}\wedge\ldots\wedge p_{nm}\wedge\bigwedge_{a\in ag(\varphi)}(\Diamond_a\chi_1\wedge\ldots\wedge\Diamond_a\chi_l\wedge\Box_a\theta_1\wedge\ldots\wedge\Box_a\theta_k)\end{array}\right),$$

where $ag(\varphi)$ is the set of agents appearing in φ. By **RQ2**, we have that $\langle\leftleftarrows\rangle\varphi$ is equivalent to

$$\bigvee\left(\begin{array}{l} \langle\leftleftarrows\rangle(p_{11}\wedge\ldots\wedge p_{1m}\wedge\bigwedge_{a\in ag(\varphi)}(\Diamond_a\chi_1\wedge\ldots\wedge\Diamond_a\chi_l\wedge\Box_a\theta_1\wedge\ldots\wedge\Box_a\theta_k))\\ \ldots\\ \langle\leftleftarrows\rangle(p_{n1}\wedge\ldots\wedge p_{nm}\wedge\bigwedge_{a\in ag(\varphi)}(\Diamond_a\chi_1\wedge\ldots\wedge\Diamond_a\chi_l\wedge\Box_a\theta_1\wedge\ldots\wedge\Box_a\theta_k))\end{array}\right).$$

Observe that $\Diamond_a\chi_1\wedge\ldots\wedge\Diamond_a\chi_l\wedge\Box_a\theta_1\wedge\ldots\wedge\Box_a\theta_k$ is equivalent (via **K**) to $\nabla_a\Phi_a$ with $\Phi_a=\{\chi_1\wedge\bigwedge_{i=1}^k\theta_k,\ldots,\chi_l\wedge\bigwedge_{i=1}^k\theta_k,\bigwedge_{i=1}^k\theta_k\}$. This follows from the validities $(\Box_a\varphi\wedge\Diamond_a\psi)\rightarrow\Diamond_a(\varphi\wedge\psi)$:

1.	$\Box_a(\varphi\rightarrow\neg\psi)\rightarrow(\Box_a\varphi\rightarrow\Box_a\neg\psi)$	**K**, **RE**
2.	$\Box_a(\neg\varphi\vee\neg\psi)\rightarrow(\neg\Box_a\varphi\vee\neg\Diamond_a\psi)$	**K**, **RE**, **Prop**
3.	$(\Box_a\varphi\wedge\Diamond_a\psi)\rightarrow\Diamond_a(\varphi\wedge\psi)$	**RE**, **Prop**

as well as $(\Box_a\varphi\wedge\Box_a\psi)\leftrightarrow\Box_a(\varphi\wedge\psi)$, which is a well known validity of **ML**.

Then, by **RQ3**, we have that

$$\bigvee\left(\begin{array}{l} p_{11}\wedge\ldots\wedge p_{1m}\wedge\langle\leftleftarrows\rangle\bigwedge_{a\in ag(\varphi)}\nabla_a\Phi_a\\ \ldots\\ p_{n1}\wedge\ldots\wedge p_{nm}\wedge\langle\leftleftarrows\rangle\bigwedge_{a\in ag(\varphi)}\nabla_a\Phi_a\end{array}\right),$$

which, in turn, is equivalent, by **RQ4**, to

$$\bigvee\left(\begin{array}{l} p_{11}\wedge\ldots\wedge p_{1m}\wedge\bigwedge_{a\in ag(\varphi)}\bigwedge_{\chi\in\Phi_a}\Diamond_a\langle\leftleftarrows\rangle\chi\\ \ldots\\ p_{n1}\wedge\ldots\wedge p_{nm}\wedge\bigwedge_{a\in ag(\varphi)}\bigwedge_{\chi\in\Phi_a}\Diamond_a\langle\leftleftarrows\rangle\chi\end{array}\right).$$

The modal depth of all χ is at most $n-1$, and hence, by the induction hypothesis, for each $\langle\leftleftarrows\rangle\chi$ there is an equivalent quantifier-free formula. □

With the similar approach as in the proof of Proposition 9, we can show the completeness of **SML**.

Proposition 10 The axiomatization **SML** is complete for $\mathscr{L}^{\Box\rightrightarrows}$.

Proof The proof goes, *mutatis mutandis*, exactly like the proof of Proposition 9. We use the induction on the modal depth of φ to show that φ, which may have simulation quantifiers, can be equivalently rewritten into a formula of modal logic. For the induction step $\langle\rightrightarrows\rangle\varphi$, we also substitute φ with an equivalent formula in disjunctive normal form, and then push $\langle\rightrightarrows\rangle$ inside using **SQ2**, **SQ3**, and $\mathbf{SQ4}_{\mathsf{cons}}$. As a consequence, formulas within the scope of $\langle\rightrightarrows\rangle$ have lower modal depth, and the result follows from the induction hypothesis. □

The proof of completeness for **OML** is also a proof by elimination, in that we inductively show that every formula of $\mathscr{L}^{\Box\circ}$ is provably (in **OML**) equivalent to a formula of $\mathscr{L}^{\Box}$.

Proposition 11 The axiomatization **OML** is complete for $\mathscr{L}^{\Box\circ}$.

Proof The structure of the model M° has every relation R_a° as a total relation, so it is an equivalence class. This is reflected in the axiomatization **OML** where we have the axioms **T** ($\Box_a\varphi \to \varphi$) and **5** ($\Diamond_a\varphi \to \Box_a\Diamond_a\varphi$) inside the $[\circ]$ operator as the axioms **OT** and **O5**. Similarly, we have the necessitation rule and modus ponens rule in the scope of the $[\circ]$ operator with the rules **OMP** and **ON**. Therefore, for every formula φ that is valid for multi-agent S5 frames, there is a proof in **OML** of $[\circ]\varphi$.

We proceed to the elimination argument as follows. We work by induction, where the induction hypothesis is *for all $\mathscr{L}^{\Box}$ formulas φ with modal depth less than or equal to n, $[\circ]\varphi$ is provably equivalent to a formula of $\mathscr{L}^{\Box}$*. The base follows directly from **O1**. Now we assume that the induction hypothesis holds for n.

Applying the axioms **O1**, **ODual** and **ODisj** we can move the $[\circ]$ operators up to the modalities. That is

$$[\circ]\varphi \leftrightarrow \theta[x_i \backslash [\circ]\Diamond_{a_i}\varphi_i]_{i=1}^m$$

is a validity where $\theta \in \mathscr{L}_0$ is a propositional formula. Furthermore, noting that $\Diamond_a$ commutes with disjunctions ($\Diamond_a(\varphi \vee \psi) \leftrightarrow (\Diamond_a\varphi \vee \Diamond_a\psi)$ is provable in **ML**), we can extend this result to say

$$[\circ]\varphi \leftrightarrow \theta[x_i \backslash [\circ]\Diamond_{a_i} \bigwedge_{j=1}^{k} \chi_j^i]_{i=1}^m$$

where χ_j^i is either a literal (an atomic proposition or its negation), or a modal formula of one of the forms $\Box_b\xi$, or $\Diamond_b\xi$.

So, via some basic **ML** reasoning, it is now sufficient for us to show that formulas of the type

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell_b}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k_b}\xi_j^b) \tag{1}$$

are equivalent to $\mathscr{L}^{\Box}$ formulas, where π is a conjunction of literals, and χ_i^b and ξ_i^b are $\mathscr{L}^{\Box}$ formulas of modal depth at most n.

By S5 reasoning, we have:

$$\Diamond_a(\pi \wedge \textstyle\bigwedge_{b\in A}\bigwedge_{i=1}^{\ell_b}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k_b}\xi_j^b)$$
$$\leftrightarrow$$
$$\Diamond_a(\pi \wedge \textstyle\bigwedge_{b\in A\setminus\{a\}}\bigwedge_{i=1}^{\ell_b}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k_b}\xi_j^b) \wedge \bigwedge_{i=1}^{\ell_a}\Diamond_a\chi_i^a \wedge \Box_a\bigwedge_{j=1}^{k_a}\xi_j^a$$

Therefore, by the above observation that all S5 validities are derivable in the scope of the $[\circ]$ operator, and **ODual** with **ODisj**, we have:

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b)$$
$$\leftrightarrow$$
$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A\setminus\{a\}}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \wedge \bigwedge_{i=1}^{\ell}[\circ]\Diamond_a\chi_i^a \wedge [\circ]\Box_a\bigwedge_{j=1}^{k}\xi_j^a$$

As for all i, $\Diamond_a\chi_i^a$ and $\Box_a\xi_i^a$ have modal depth less than or equal to n, we can apply the induction hypothesis, so that they are provably equivalent to $\mathscr{L}^{\Box}$ formulas. Therefore we have

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b)$$
$$\leftrightarrow$$
$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A\setminus\{a\}}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \wedge \varphi_a$$

where φ_a is a $\mathscr{L}^{\Box}$ formula. We can now apply **OExch**, noting that is is an equivalence, to derive:

$$\begin{array}{c}[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b)\\ \leftrightarrow\\ [\circ]\Diamond_b(\pi \wedge \bigwedge_{b\in A\setminus\{a\}}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \wedge \varphi_a\end{array} \tag{2}$$

Noting the similarity of the first line of (2) and the first conjunct of the second line of (2), where $b \in A\setminus\{a\}$, we can recursively apply this argument for each modality $a \in A$ to derive

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \leftrightarrow [\circ]\Diamond_a\pi \wedge \bigwedge_{a\in A}\varphi_a$$

Here, the a in $[\circ]\Diamond_a\pi$ is arbitrary. It will be the last modality left in the set A, but via the **OExch** axiom it can be changed to any modality. As π is a conjunction of literals, either it contains an atom and its negation, in which case we have:

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \leftrightarrow \bot$$

or it matches the condition for **OFull**, in which case we can derive $[\circ]\Diamond_a\pi$ so

$$[\circ]\Diamond_a(\pi \wedge \bigwedge_{b\in A}\bigwedge_{i=1}^{\ell}\Diamond_b\chi_i^b \wedge \Box_b\bigwedge_{j=1}^{k}\xi_j^b) \leftrightarrow \bigwedge_{a\in A}\varphi_a$$

This is sufficient to show $\mathscr{L}^{\Box\circ}$ formulas of the kind in equation (1) are equivalent to $\mathscr{L}^{\Box}$ formulas which concludes the completeness proof. □

Finally, having established the way to eliminate refinement and simulation quantifiers, as well as the origin modality, in proofs of Propositions 9, 10, and 11, the completeness of **ROSML** follows.

Proposition 12 The axiomatization **ROSML** is complete for $\mathscr{L}$.

Proof The proof is similar to those of Propositions 9, 11, and 10. Yet again, we recursively eliminate quantifiers and the origin modality from the inside-out, starting from the innermost occurrence of the operator in question. Hence, for the induction case $\langle\leftrightharpoons\rangle\varphi$ we proceed exactly as in the proof of Proposition 9. For the case $\langle\rightrightarrows\rangle\varphi$, we follow the strategy in the proof of Proposition 10. Finally, for the origin modality we follow the proof of Proposition 11. □

5 Conclusion and Discussion

We have presented several novel logics for reasoning about information growth and information loss. Refinement modal logic RML quantifies over information growth. We gave a novel axiomatization for that logic, slightly different from [14]. Simulation modal logic SML quantifies over information loss. We presented two axiomatizations for (or containing) SML, **SML** and **ROSML**. The first, SML, has consistency requirements on formulas in one of the axioms and is a variation of [44]. The second, **ROSML**, does not have that consistency requirement but apart from simulation quantifiers also has refinement quantifiers and the novel origin modality. The origin modality checks whether the formula bound by it was true in the model before any of the agents received any factual information: the mutual factual ignorance model. We also gave an axiomatization **OML** for origin modal logic OML with only this origin modality but without refinement and simulation quantifiers.

There is ample scope for further research. The semantics of the refinement and simulation quantifiers depend on the class of models it quantifies over. In our work, we have considered the class K of all models. Hence, an immediate research problem is the axiomatization of *simulation epistemic logic*, i.e. SML with simulation modalities quantifying over S5 models, that is, epistemic models with equivalence accessibility relations. The axiomatization of a corresponding *refinement epistemic logic* where refinement modalities quantify over S5 models has been presented in [34, 33, 32]; these works also contain refinement modal logics over other classes of models, such as KD45 encoding consistent belief.

Finally, it would be of clear interest to determine the complexity of satisfiability of SML. Let us survey such results for RML. The satisfiability problem of single-agent RML is AEXP_{pol} and conjectured to be non-elementary for multi-agent RML [15]. It is slightly unclear if the latter also holds for the RML of our contribution with refinement quantifiers for the set of all agents only. In [1] a PSPACE upper bound is given for complexity of the existential fragment of single-agent RML (a negation normal form wherein only diamonds $\langle \leftrightarrows \rangle$ may occur). Complexities have also been reported for the logic μ-RML extending RML with fixpoints [14], and in work in progress building on the fixpoint extension [43] of covariant-contravariant refinement modal logic [44] (with modalities combining refinement and simulation).

Acknowledgements We thank the TARK reviewers for their comments. Although based in the past of [14], this research really got off the ground during Tim's visit to Hans in Toulouse, in April 2024, followed by subsequent one-on-one meetings of two collaborators, and also involving Rustam and Louwe, and both in Bergen and Prague, while the four of us were never together in the same place and time. Hans van Ditmarsch acknowledges the support of a grant *CIMI Mobilité Internationale* from Université Paul Sabatier to visit Bergen, and support from the *Wolfgang Pauli Institute* (WPI) in Vienna where he was a Fellow during the completion of this research.

References

[1] A. Achilleos & M. Lampis (2013): *Closing a Gap in the Complexity of Refinement Modal Logic*. *CoRR* abs/1309.5184. Available at `http://arxiv.org/abs/1309.5184`.

[2] C.E. Alchourrón, P. Gärdenfors & D. 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`.

[3] G. Aucher (2008): *Perspectives on Belief and Change*. Ph.D. thesis, University of Otago & Institut de Recherche en Informatique de Toulouse, New Zealand & France. Available at `https://tel.archives-ouvertes.fr/tel-00556089`.

[4] G. Aucher (2012): *DEL-sequents for Regression and Epistemic Planning*. Journal of Applied Non-Classical Logics 22(4), pp. 337–367, doi:10.1080/11663081.2012.736703.

[5] P. Balbiani, H. van Ditmarsch & A. Herzig (2016): *Before Announcement*. In L.D. Beklemishev, S. Demri & A. Maté, editors: Proc. of 11th AiML, College Publications, pp. 58–77.

[6] A. Baltag, H. van Ditmarsch & L.S. Moss (2008): *Epistemic Logic and Information Update*. In J. van Benthem & P. Adriaans, editors: Handbook on the Philosophy of Information, Elsevier, Amsterdam, pp. 361–456, doi:10.1016/b978-0-444-51726-5.50015-7.

[7] A. Baltag, L.S. Moss & S. Solecki (1998): *The Logic of Public Announcements, Common Knowledge, and Private Suspicions*. In I. Gilboa, editor: Proc. of 7th TARK, Morgan Kaufmann, pp. 43–56.

[8] A. Baltag, A. Özgün & A.L. Vargas Sandoval (2022): *Arbitrary Public Announcement Logic with Memory*. Journal of Philosophical Logic, doi:10.1007/s10992-022-09664-6.

[9] A. Baltag & S. Smets (2008): *A Qualitative Theory of Dynamic Interactive Belief Revision*. In: Proc. of 7th LOFT, Texts in Logic and Games 3, Amsterdam University Press, pp. 11–58.

[10] 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.

[11] J. van Benthem (2011): *Logical Dynamics of Information and Interaction*. Cambridge University Press, doi:10.1017/CBO9780511974533.

[12] J. van Benthem & S. Smets (2015): *Dynamic Logics of Belief Change*. In H. van Ditmarsch, J.Y. Halpern, W. van der Hoek & B. Kooi, editors: Handbook of epistemic logic, College Publications, pp. 313–393.

[13] P. Blackburn, M. de Rijke & Y. Venema (2001): *Modal Logic*. Cambridge University Press, doi:10.1017/CBO9781107050884.

[14] L. Bozzelli, H. van Ditmarsch, T. French, J. Hales & S. Pinchinat (2014): *Refinement Modal Logic*. Information and Computation 239, pp. 303–339, doi:10.1016/J.IC.2014.07.013.

[15] L. Bozzelli, H. van Ditmarsch & S. Pinchinat (2015): *The Complexity of One-Agent Refinement Modal Logic*. Theoretical Computer Science 603, pp. 58–83, doi:10.1016/j.tcs.2015.07.015.

[16] R. Demolombe, A. Herzig & I. Varzinczak (2003): *Regression in Modal Logic*. Journal of Applied Non-Classical Logics 13(2), pp. 165–185, doi:10.3166/JANCL.13.165-185.

[17] H. van Ditmarsch (2005): *Prolegomena to Dynamic Logic for Belief Revision*. Synthese 147, pp. 229–275, doi:10.1007/s11229-005-1349-7.

[18] H. van Ditmarsch (2023): *To be announced*. Information and Computation 292, p. 105026, doi:10.1016/J.IC.2023.105026.

[19] H. van Ditmarsch, J. van Eijck & R. Verbrugge (2009): *Common Knowledge and Common Belief*. In J. van Eijck & R. Verbrugge, editors: Discourses on Social Software, Amsterdam University Press, pp. 99–122.

[20] H. van Ditmarsch & T. French (2009): *Simulation and Information: Quantifying over Epistemic Events*. In J. Broersen & J.-J. Meyer, editors: Proc. of 1st KRAMAS, LNCS 5605, Springer, pp. 51–65, doi:10.1007/978-3-642-05301-6_4.

[21] H. van Ditmarsch, T. French & S. Pinchinat (2010): *Future Event Logic - Axioms and Complexity*. In L. Beklemishev, V. Goranko & V. Shehtman, editors: Proc. of 8th AiML, College Publications, pp. 77–99.

[22] H. van Ditmarsch, A. Herzig, J. Lang & P. Marquis (2009): *Introspective Forgetting*. Synthese 169(2), pp. 405–423, doi:10.1007/S11229-009-9554-4.

[23] H. van Ditmarsch, A. Herzig & T. De Lima (2007): *Optimal Regression for Reasoning about Knowledge and Actions*. In: Proc. of 22nd AAAI, AAAI Press, pp. 1070–1075.

[24] H. van Ditmarsch, W. van der Hoek & B. Kooi (2007): *Dynamic Epistemic Logic*. Synthese Library 337, Springer, doi:10.1007/978-1-4020-5839-4.

[25] 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.

[26] H. van Ditmarsch, W. van der Hoek & B. Kooi (2005): *Public Announcements and Belief Expansion*. In R. Schmidt, I. Pratt-Hartmann, M. Reynolds & H. Wansing, editors: *Proc. of 5th AiML*, King's College Publications, pp. 335–346.

[27] L. Fang, Y. Liu & H. van Ditmarsch (2019): *Forgetting in Multi-Agent Modal Logics*. *Artificial Intelligence* 266, pp. 51–80, doi:10.1016/J.ARTINT.2018.08.003.

[28] D. Fernández-Duque, Á. Nepomuceno-Fernández, E. Sarrión-Morillo, F. Soler-Toscano & F.R. Velázquez-Quesada (2015): *Forgetting Complex Propositions*. *Logic Journal of the IGPL* 23(6), pp. 942–965, doi:10.1093/JIGPAL/JZV049.

[29] T. French (2006): *Bisimulation Quantifiers for Modal Logic*. Ph.D. thesis, University of Western Australia. Available at https://research-repository.uwa.edu.au/en/publications/bisimulation-quantifiers-for-modal-logics.

[30] J. Hales (2011): *Refinement Quantifiers for Logics of Belief and Knowledge*. Honours Thesis, University of Western Australia.

[31] J. Hales (2013): *Arbitrary Action Model Logic and Action Model Synthesis*. In: *Proc. of 28th LICS*, IEEE, pp. 253–262, doi:10.1109/LICS.2013.31.

[32] J. Hales (2016): *Quantifying Over Epistemic Updates*. Ph.D. thesis, School of Computer Science & Software Engineering, University of Western Australia. Available at https://research-repository.uwa.edu.au/en/publications/quantifying-over-epistemic-updates.

[33] J. Hales, T. French & R. Davies (2011): *Refinement Quantified Logics of Knowledge*. *Electronic Notes in Theoretical Computer Science* 278, pp. 85–98, doi:10.1016/J.ENTCS.2011.10.008.

[34] J. Hales, T. French & R. Davies (2012): *Refinement Quantified Logics of Knowledge and Belief for Multiple Agents*. In Thomas Bolander, Torben Braüner, Silvio Ghilardi & Lawrence S. Moss, editors: *Proc. of 9th AiML*, College Publications, pp. 317–338.

[35] J. Hintikka (1962): *Knowledge and Belief*. Cornell University Press.

[36] J. Lang & P. Marquis (2010): *Reasoning Under Inconsistency: A Forgetting-Based Approach*. *Artificial Intelligence* 174(12-13), pp. 799–823, doi:10.1016/J.ARTINT.2010.04.023.

[37] X. Liang & Y. N. Wáng (2024): *Epistemic Skills: Logical Dynamics of Knowing and Forgetting*. In Antonis Achilleos & Adrian Francalanza, editors: *Proc. of 15th GandALF 2024*, *EPTCS* 409, pp. 120–137, doi:10.4204/EPTCS.409.12.

[38] J.S. Miller & L.S. Moss (2005): *The Undecidability of Iterated Modal Relativization*. *Studia Logica* 79(3), pp. 373–407, doi:10.1007/S11225-005-3612-9.

[39] L.S. Moss (2015): *Dynamic Epistemic Logic*. In H. van Ditmarsch, J.Y. Halpern, W. van der Hoek & B. Kooi, editors: *Handbook of epistemic logic*, College Publications, pp. 261–312.

[40] J. Sack (2010): *Logic for Update Products and Steps Into the Past*. *Annals of Pure and Applied Logic* 161(12), pp. 1431–1461, doi:10.1016/J.APAL.2010.04.011.

[41] 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.

[42] A. Visser (1996): *Bisimulations, Model Descriptions and Propositional Quantifiers*. Logic Group Preprint Series 161, Department of Philosophy, Utrecht University.

[43] H. Xing (2022): *Covariant-Contravariant Refinement Modal μ-calculus*. *CoRR* abs/2208.02989, doi:10.48550/ARXIV.2208.02989. Available at http://arxiv.org/abs/2208.02989.

[44] H. Xing, Zh. Zhu & J. Zhang (2019): *Covariant-Contravariant Refinement Modal Logic*. *FLAP* 6(1), pp. 1–30.

Impure Simplicial Complex and Term-Modal Logic with Assignment Operators

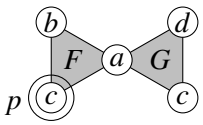
Yuanzhe Yang
Department of Philosophy and Religious Studies
Peking University
Beijing, China
1900014924@pku.edu.cn

Impure simplicial complexes are a powerful tool to model multi-agent epistemic situations where agents may die, but it is difficult to define a satisfactory semantics for the ordinary propositional modal language on such models, since many conceptually dubious expressions involving dead agents can be expressed in this language. In this paper, we introduce a term-modal language with assignment operators, in which such conceptually dubious expressions are syntactically excluded. We define both simplicial semantics and first-order Kripke semantics for this language, characterize their respective expressivity through notions of bisimulation, and show that the two semantics are equivalent when we consider a special class of first-order Kripke models called local epistemic models. We also offer a complete axiomatization for the epistemic logic based on this language, and show that our language has a notion of assignment normal form. Finally, we discuss the behavior of a kind of intensional distributed knowledge that can be naturally expressed in our language.

1 Introduction

Simplicial Complexes have been a useful technical tool to study distributed computing since the 1990s,[1] but it was only recently that logicians also started using them to study multi-agent epistemic logic [19, 9]. Roughly, a simplicial complex is a set of vertices colored by names of agents, equipped with a family of downward-closed subsets of vertices called *faces* or *simplexes*. Intuitively, the colored vertices represent the local epistemic states of the agents, while the maximal faces, called *facets*, represent global epistemic states, which are epistemically possible for the agents they contain.

A simplicial complex is called *impure* if the vertices in different facets can be colored by different sets of agents. Intuitively, this means that an agent can be *alive* in some global states, while *dead* in the others. Though this is a rather natural idea, it turns out to be quite tricky to define a reasonable semantics for the ordinary propositional modal language (with agent-labeled propositional letters and modal operators) on impure simplicial complexes. This is mainly because it is sometimes difficult to decide the truth value of formulas involving possibly dead agents. For example, consider the sentence $\mathsf{K}_a\mathsf{K}_b p_c$, which intuitively says that a knows that b knows that c has property p. Should it be true at facet F in the following simplicial complex $\mathscr{C}$?



[1] Earliest works on distributed computing using simplicial complexes include [8, 15, 2], etc. One may refer to [13] for more details on this topic.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 399–416, doi:10.4204/EPTCS.437.31

© Y. Yang
This work is licensed under the
Creative Commons Attribution License.

Our intuition here is not very clear: since a can imagine the global state G where b is dead, we need to decide the truth value of $\mathsf{K}_b p_c$ at G; but conceptually, it is hard to tell what a dead agent should know.

Hitherto, there are two kinds of impure simplicial semantics: one is the two-valued semantics proposed in [10], in which dead agents also have concrete properties — in particular, a dead agent always "knows" everything, including the contradiction; the other is the three-valued semantics proposed in [18, 6, 1], in which formulas involving dead agents might be neither true nor false, but *undefined*.[2] It is worth noting that according to both semantics, in the above example, $\mathsf{K}_a\mathsf{K}_b p_c$ is *true* at F.

None of the two semantics seems completely satisfactory: in the two-valued semantics, we are forced to assign true or false to arbitrary expressions involving possibly dead agents, which might be conceptually dubious, [3] and it sounds absurd to suggest that a dead agent *knows* the contradiction; on the other hand, the three-valued semantics relies on a rather complicated notion of definability, making it a bit difficult to study its behavior on the technical level, and the corresponding axiomatization is also a non-classical non-normal modal logic [18, 6]. It is also interesting to note that neither semantics validates the standard T-axiom for knowledge: in the two-valued semantics, $\mathsf{K}_a\phi \to \phi$ can simply be *false* when a is dead, while in the three-valued semantics, the truth of $\mathsf{K}_a\phi$ only implies that ϕ is true *if defined*.

In this paper, we propose an alternative way to do epistemic logic on impure simplicial complexes, which seems more well-behaved both conceptually and technically. Our strategy is to *avoid* talking about dead agents in any non-trivial way. For example, let us reconsider the example $\mathsf{K}_a\mathsf{K}_b p_c$ — a knows that b knows that c satisfies p. For simplicity, we assume that a is a living agent (as in the above model), so that it makes perfect sense to talk about a's epistemic states. Then, while the truth value of $\mathsf{K}_a\mathsf{K}_b p_c$ is difficult to tell, the following four "clarifications" of it all have obvious truth values:

(i) a knows that (*if b is alive, then b* knows that (*if c is alive, then c* satisfies p)).
(ii) a knows that (*if b is alive, then b* knows that (c *is alive and c* satisfies p)).
(iii) a knows that (b *is alive and b* knows that (*if c is alive, then c* satisfies p)).
(iv) a knows that (b *is alive and b* knows that (c *is alive and c* satisfies p)).

Unlike the original expression, these clarifications involve only properties and epistemic states of *living* agents, and it is quite clear that at $\mathscr{C}, F$, (i), (ii) are true, while (iii), (iv) are false. Then, instead of the ordinary propositional modal language which allows dubious expressions like $\mathsf{K}_a\mathsf{K}_b p_c$, if we can work in a more delicate language, in which we are only allowed to talk about properties and epistemic states of living agents, then we should be able to define a fairly intuitive semantics without entering any serious philosophical controversy concerning the states of dead agents, since philosophically controversial expressions would be excluded *syntactically* from our language.

This is our motivation to introduce a special kind of *term-modal language*, which is a *first-order modal language* with unary predicates (viewed as labeled propositional letters), modal operators labeled by variables, and *assignment operators* of the form $[x{:=}a]\phi$ for each agent a. Intuitively, $[x{:=}a]\phi$ says that *if a is alive*, then after assigning a to x, ϕ is true; its dual, $\langle x{:=}a\rangle\phi$, says that *$a$ is alive and* after assigning a to x, ϕ is true. In this language, unary predicates and modal operators are labeled only by *variables*, which refer to living agents only; on the other hand, names of agents appear only in assignment operators. Thus, whenever we want to talk about an agent directly, we are forced to use the corresponding assignment operators, which clarify the existential state of the agent in question before we start talking about any of their concrete properties and epistemic states. As a result, in such a language, sentences like (i) to (iv) above can still be expressed: for example, assuming the existence of a, (ii) can be formalized

[2] The relation between this two kinds of semantic is studied in [7].

[3] In [10], the authors indeed offer an conceptual explanation for assigning non-trivial truth values of propositional letters to dead agents: as the authors explain, propositional letters might describe what the dead agents did *before they died*. However, it seems that such an explanation only works when we interpret *all* propositional letters in this rather specific way.

as $\langle x{:=}a\rangle \mathsf{K}_x[y{:=}b]\mathsf{K}_y\langle z{:=}c\rangle p_z$, while (iii) can be formalized as $\langle x{:=}a\rangle \mathsf{K}_x\langle y{:=}b\rangle \mathsf{K}_y[z{:=}c]p_z$; in contrast, dubious expressions like $\mathsf{K}_a\mathsf{K}_b p_c$ are no longer inexpressible.

It should be noted that in the literature, assignment operators are usually of the form $[x{:=}t]\phi$, where t is a *term*. This kind of assignment operators are firstly introduced in [14] as a part of a more expressive first-order dynamic language; more closely related to our work, a quantifier-free term-modal logic with assignment operators is studied in [20]. The semantics of $[x{:=}t]\phi$ is that, after assigning the value of t to x, ϕ is true. As noted in [14], this kind of assignment operators are also closely related to the notion of *predicate abstraction* introduced in [17]: $[x{:=}t]\phi$ is equivalent with $\langle \lambda x.\phi\rangle(t)$.

Our assignment operators differ from the above kind of assignment operators in many important ways. Conceptually, assignment operators of the above kind (as well as predicate abstraction) usually work a tool to capture the *de re* / *de dicto* distinction in first-order modal logic; but we use assignment operators to clarify the existential states of agents. Technically, in $[x{:=}t]\phi$, t is also a term of the language, which can be used to label predicates or modal operators; but in our operator $[x{:=}a]\phi$, a is *not* a legitimate term, and can only appear within the assignment operator. Also, there is no difference between $[x{:=}t]\phi$ and its dual $\langle x{:=}t\rangle\phi$, but in our language, there is an essential difference between $[x{:=}a]\phi$ and $\langle x{:=}a\rangle\phi$: only the latter entails the existence of a.

In this paper, then, we will develop an epistemic logic for impure simplicial complexes based on the more delicate language with assignment operators of the form $[x{:=}a]\phi$.

The structure of the paper is as follows. In Section 2, we formally introduce the language and its semantics. Since we are working with a kind of first-order modal language, we will also give it a natural *first-order Kripke semantics*, which should help us better understand the simplicial semantics both conceptually and technically. We will also introduce two notions of bisimulation to characterize the expressivity of the two semantics respectively. Next, in Section 3, we show that for a special class of first-order Kripke models called *local epistemic models*, the first-order Kripke semantics is equivalent to the simplicial semantics; this resembles the correspondence between simplicial complexes and propositional Kripke models proved in [9, 18, 6]. Then, in Section 4, we offer a complete axiomatization of the epistemic logic based on our language, and show that our language has a notion of *assignment normal form*. Then, in Section 5, we study a kind of *intensional* distributed knowledge that can be naturally expressed in our language. Finally, in Section 6, we summarize our results and consider directions for future work.

2 Language, Semantics and Expressivity

We first introduce the language and semantics. Fix a non-empty finite set $\mathbf{A}$ of agents and a countable set $\mathbf{X}$ of variables (let $\mathbf{A}\cap\mathbf{X}=\emptyset$). Also let $\mathbf{P}$ be a countable set of unary predicates / propositional letters to be labeled. Then, we define the language as follows.

Definition 2.1 ($\mathscr{L}^{[:=]}$) *Formulas in $\mathscr{L}^{[:=]}$, together with the notion of* free variables *of a formula, are defined recursively as follows:*[4]

$$\phi ::= p_x \mid \top \mid \neg\phi \mid (\phi\wedge\phi) \mid [x{:=}a]\phi \mid \mathsf{K}_X\alpha$$

where $p\in\mathbf{P}$, $x\in\mathbf{X}$, $a\in\mathbf{A}$, $X\subseteq\mathbf{X}$ is finite (X can be empty), and α is an $\mathscr{L}^{[:=]}$-formula s.t. $FV(\alpha)=\emptyset$.

$$FV(p_x)=\{x\};\ FV(\top)=\emptyset;\ FV(\neg\phi)=FV(\phi);\ FV(\phi\wedge\psi)=FV(\phi)\cup FV(\psi);$$
$$FV([x{:=}a]\phi)=FV(\phi)\setminus\{x\};\ FV(\mathsf{K}_X\alpha)=X.$$

[4] We need to define them at the same time, since we need to use the notion of free variables when forming modal formulas.

Other Boolean connectives are defined in the usual way.

$\langle x:=a\rangle\phi$ is the abbreviation for $\neg[x:=a]\neg\phi$, and $\widehat{\mathsf{K}}_X\alpha$ is the abbreviation for $\neg\mathsf{K}_X\neg\alpha$.

We use $\phi[y/x]$ to denote the formula obtained by replacing all free occurrences of x in ϕ with y, and we say $\phi[y/x]$ is admissible *if x has no free occurrence within the scope of $[y:=a]$ for any $a\in\mathbf{A}$.*

*Let $\mathscr{L}_0^{[:=]}$ denote the set of all formulas in $\mathscr{L}^{[:=]}$ with no free variable — as usual, we also call them $\mathscr{L}^{[:=]}$-*sentences*. (In this paper, we usually use ϕ,ψ,χ to denote $\mathscr{L}^{[:=]}$-formulas in general, and use α,β,γ to denote $\mathscr{L}^{[:=]}$-sentences.)*

Note that the primitive notion of knowledge in our language is *distributed knowledge*: intuitively, $\mathsf{K}_X\alpha$ says that the group of agents denoted by X knows that α after putting their information together. [5] As we shall see, the distributed knowledge operators enable us to express certain model properties in a more natural way.

Next, we introduce the simplicial complex semantics and first-order Kripke semantics for this language. The notion of simplicial models and first-order Kripke models are defined as follows:

Definition 2.2 (Simplicial Model) *A* simplicial model *is a 4-tuple $\mathscr{C}=(\mathscr{V},C,\chi,\ell)$, where $\mathscr{V}\neq\emptyset$ is the set of* vertices*, $C\subseteq\wp(\mathscr{V})$ is the set of* faces *(or* simplexes*) satisfying that (i) $\emptyset\notin C$, (ii) if $\emptyset\neq Y\subseteq X\in C$, then $Y\in C$, and (iii) for all $v\in\mathscr{V}$, $\{v\}\in C$; $\chi:\mathscr{V}\to\mathbf{A}$ is the coloring function s.t. for all $X\in C$, $\chi_{|X}$ is an injection, and $\ell:\mathbf{P}\to\wp(\mathscr{V})$ is the labeling function.*

Given a simplicial model $\mathscr{C}=(\mathscr{V},C,\chi,\ell)$, let $\mathscr{F}(C)=\{F\in C\mid \text{for all } X\in C\setminus\{F\}, F\not\subseteq X\}$. Elements in $\mathscr{F}(C)$ are called the facets *in C.*

For convenience, we will sometimes use notations like $\mathscr{V}^{\mathscr{C}}$, $C^{\mathscr{C}}$, $\chi^{\mathscr{C}}$ and $\ell^{\mathscr{C}}$ to denote the corresponding components of a simplicial model $\mathscr{C}$.

Definition 2.3 (First-Order Kripke Model) *In this paper, by a* first-order Kripke model*, we mean a 4-tuple $\mathscr{M}=(W,\delta,\{R_a\}_{a\in\mathbf{A}},\rho)$, where $W\neq\emptyset$ is the set of possible worlds, $\delta:W\to\wp(\mathbf{A})\setminus\{\emptyset\}$ assigns each $w\in W$ a local domain of agents, for each $a\in\mathbf{A}$, $R_a\subseteq W\times W$ is the accessibility relation for a s.t. for all $w\in W$, if $a\notin\delta(w)$, then $R_a(w)=\emptyset$, and $\rho:\mathbf{P}\times W\to\wp(\mathbf{A})$ is the interpretation function that assigns each pair of $p\in\mathbf{P}$ and $w\in W$ a subset of $\delta(w)$.*[6]

For convenience, we will sometimes use notations like $W^{\mathscr{M}}$, $\delta^{\mathscr{M}}$, $R_a^{\mathscr{M}}$ and $\rho^{\mathscr{M}}$ to denote the corresponding components of a first-order Kripke model $\mathscr{M}$.

Here, we do not set any further constraint on the accessibility relations in a first-order Kripke model, so they are in a sense more *free* than simplicial models. In fact, as we shall see in Section 3, only a special class of first-order Kripke models corresponds to simplicial models. But our results concerning first-order Kripke models presented in this section do not rely on any such special properties.

In order to define the semantics for open formulas, we also need the following notion of *admissible assignments*. Intuitively, such assignments ensure that an open formula only mentions *existent* agents.

Definition 2.4 (Admissible Assignment) *An* assignment *is a function from $\mathbf{X}$ to $\mathbf{A}$.*

For a simplicial model $\mathscr{C}=(\mathscr{V},C,\chi,\ell)$ and $F\in\mathscr{F}(C)$, $\phi\in\mathscr{L}^{[:=]}$, we say an assignment σ is admissible *for $\mathscr{C},F$ and ϕ, if $\sigma[FV(\phi)]\subseteq\chi[F]$.*

Similarly, for a first-order Kripke model $\mathscr{M}=(W,\delta,R,\rho)$, and $w\in W$, $\phi\in\mathscr{L}^{[:=]}$, we say an assignment σ is admissible *for $\mathscr{M},w$ and ϕ, if $\sigma[FV(\phi)]\subseteq\delta(w)$.*

[5] A similar kind of distributed knowledge operator labeled by a set of variables is introduced in [16].

[6] Note that according to our definition, in every world, only existing agents have non-trivial epistemic relations, and the extension of a predicate only includes existing agents. We make these assumptions here because we do not want to decide the properties and epistemic states of non-existing agents on the conceptual level, and also because they help us establish a cleaner correspondence between simplicial models and first-order Kripke models.

Then, the two semantics are defined as follows.

Definition 2.5 (The Semantics) *The satisfaction relation $\Vdash$ for simplicial models and the satisfaction relation $\vDash$ for first-order Kripke models are defined recursively as follows when the corresponding assignment is* admissible *for the pointed model and formula in question (the parts for $\top$ and Boolean connectives are omitted):*[7]

Simplicial Model			*First-Order Kripke Model*		
$\mathscr{C},F,\sigma \Vdash p_x$	*iff*	$\sigma(x) \in \chi[F \cap \ell(p)]$	$\mathscr{M},w,\sigma \vDash p_x$	*iff*	$\sigma(x) \in \rho(p,w)$
$\mathscr{C},F,\sigma \Vdash [x{:=}a]\phi$	*iff*	*if* $a \in \chi[F]$, *then* $\mathscr{C},F,\sigma[x \mapsto a] \Vdash \phi$	$\mathscr{M},w,\sigma \vDash [x{:=}a]\phi$	*iff*	*if* $a \in \delta(w)$, *then* $\mathscr{M},w,\sigma[x \mapsto a] \vDash \phi$
$\mathscr{C},F,\sigma \Vdash \mathsf{K}_X\alpha$	*iff*	*for all* $G \in \mathscr{F}(C)$ *s.t.* $\sigma[X] \subseteq \chi[F \cap G]$, $\mathscr{C},G,\sigma \Vdash \alpha$	$\mathscr{M},w,\sigma \vDash \mathsf{K}_X\alpha$	*iff*	*for all* $v \in W$ *s.t.* $v \in \bigcap_{a\in\sigma[X]} R_a(w)$ $\mathscr{M},v,\sigma \vDash \alpha$

Note that the truth value of an $\mathscr{L}^{[:=]}$-sentence α has nothing to do with the assignment. Thus, for $\alpha \in \mathscr{L}_0^{[:=]}$, we may simply write $\mathscr{C},F \Vdash \alpha$ (resp. $\mathscr{M},w \vDash \alpha$) when $\mathscr{C},F,\sigma \Vdash \alpha$ (resp. $\mathscr{M},w,\sigma \vDash \alpha$) for some assignment σ.

It is not hard to see that the semantics of the assignment operators is exactly the one we described in Section 1. Also note that $\langle x{:=}a\rangle\top$ simply says that *a exists*, so in this sense, we can express the *existential predicate* in our language.

Now, we introduce two notions of bisimulation to characterize the expressivity of the above two semantics, respectively. Note that our notion of bisimulation for simplicial models is essentially the same as the kind of bisimulation introduced in [1].

Definition 2.6 (Bisimulations) *Let $\mathscr{C}$, $\mathscr{D}$ be simplicial models, and let $\mathscr{M}$, $\mathscr{N}$ be first-order Kripke models. A relation $Z \subseteq \mathscr{F}(C^{\mathscr{C}}) \times \mathscr{F}(C^{\mathscr{D}})$ (resp. $Z \subseteq W^{\mathscr{M}} \times W^{\mathscr{N}}$) is a* bisimulation *between $\mathscr{C}$ and $\mathscr{D}$ (resp. $\mathscr{M}$ and $\mathscr{N}$), if the following conditions are satisfied for any $(F,G) \in Z$ (resp. $(w,v) \in Z$):*

	Simplicial Models	*First-Order Kripke Models*
(Inv)	$\chi^{\mathscr{C}}[F] = \chi^{\mathscr{D}}[G]$, *and for all* $p \in \mathbf{P}$, $\chi^{\mathscr{C}}[F \cap \ell^{\mathscr{C}}(p)] = \chi^{\mathscr{D}}[F \cap \ell^{\mathscr{D}}(p)]$	$\delta^{\mathscr{M}}(w) = \delta^{\mathscr{N}}(v)$, *and for all* $p \in \mathbf{P}$, $\rho^{\mathscr{M}}(p,w) = \rho^{\mathscr{N}}(p,v)$
(Zig)	*For all* $A \subseteq \mathbf{A}$ *and* $F' \in \mathscr{F}(C^{\mathscr{C}})$ *s.t.* $A \subseteq \chi^{\mathscr{C}}[F \cap F']$, *there is* $G' \in \mathscr{F}(C^{\mathscr{D}})$ *s.t.* $A \subseteq \chi^{\mathscr{D}}[G \cap G']$ *and* $(F',G') \in Z$	*For all* $A \subseteq \mathbf{A}$ *and* $w' \in W^{\mathscr{M}}$ *s.t.* $w' \in \bigcap_{a\in A} R_a^{\mathscr{M}}(w)$, *there is* $v' \in W^{\mathscr{N}}$ *s.t.* $v' \in \bigcap_{a\in A} R_a^{\mathscr{N}}(v)$ *and* $(w',v') \in Z$
(Zag)	*For all* $A \subseteq \mathbf{A}$ *and* $G' \in \mathscr{F}(C^{\mathscr{D}})$ *s.t.* $A \subseteq \chi^{\mathscr{D}}[G \cap G']$, *there is* $F' \in \mathscr{F}(C^{\mathscr{C}})$ *s.t.* $A \subseteq \chi^{\mathscr{C}}[F \cap F']$ *and* $(F',G') \in Z$	*For all* $A \subseteq \mathbf{A}$ *and* $v' \in W^{\mathscr{N}}$ *s.t.* $v' \in \bigcap_{a\in A} R_a^{\mathscr{N}}(v)$, *there is* $w' \in W^{\mathscr{M}}$ *s.t.* $w' \in \bigcap_{a\in A} R_a^{\mathscr{M}}(w)$ *and* $(w',v') \in Z$

We say $\mathscr{C},F$ and $\mathscr{D},G$ (resp. $\mathscr{M},w$ and $\mathscr{N},v$) are bisimlar*, if there is a bisimulation between $\mathscr{C}$ and $\mathscr{D}$ (resp. $\mathscr{M}$ and $\mathscr{N}$) s.t. $(F,G) \in Z$ (resp. $(w,v) \in Z$).*

For any simplicial models $\mathscr{C},F$, $\mathscr{D},G$ and any first-order Kripke models $\mathscr{M},w$, $\mathscr{N},v$, we say $\mathscr{C},F$ and $\mathscr{D},G$ (resp. $\mathscr{M},w$ and $\mathscr{N},v$) are $\mathscr{L}^{[:=]}$-*logically equivalent*, denoted by $\mathscr{C},F \equiv_{\mathscr{L}^{[:=]}} \mathscr{D},G$ (resp. $\mathscr{M},w \equiv_{\mathscr{L}^{[:=]}} \mathscr{N},v$), if for any $\alpha \in \mathscr{L}_0^{[:=]}$, $\mathscr{C},F \Vdash \alpha$ iff $\mathscr{D},G \Vdash \alpha$ (resp. $\mathscr{M},w \vDash \alpha$ iff $\mathscr{N},v \vDash \alpha$). Then, it is rather routine to prove the following theorem:

[7] Some remarks: (i) For a *sentence* α, any assignment is admissible. This guarantees that the semantics for K_X is well-defined. (ii) In the first-order Kripke semantics for $\mathsf{K}_X\alpha$, we make the convention that $\bigcap_{a\in\emptyset} R_a(w) = W$, so $\mathsf{K}_\emptyset$ works as the universal modality.

Theorem 2.7 *For any simplicial models $\mathscr{C},F$, $\mathscr{D},G$ and any first-order Kripke models $\mathscr{M},w$, $\mathscr{N},v$,*

$$\mathscr{C},F \leftrightarrows \mathscr{D},G \Longrightarrow \mathscr{C},F \equiv_{\mathscr{L}^{[:=]}} \mathscr{D},G \quad \textit{and} \quad \mathscr{M},w \leftrightarrows \mathscr{N},v \Longrightarrow \mathscr{M},w \equiv_{\mathscr{L}^{[:=]}} \mathscr{N},v$$

PROOF. (Sketch) If Z is a bisimulation between two simplicial models $\mathscr{C}$ and $\mathscr{D}$, then we can show by induction that for any $\phi \in \mathscr{L}^{[:=]}$, any $(F,G) \in Z$ and any assignment σ, σ is admissible for $\mathscr{C},F$ and ϕ iff it is admissible for $\mathscr{D},G$ and ϕ, and when it is admissible, $\mathscr{C},F,\sigma \Vdash \phi$ iff $\mathscr{D},G,\sigma \Vdash \phi$. The case for first-order Kripke models is basically the same. □

In order to show that both our semantics have the Hennessy-Milner property, we define the following notions of saturation:

Definition 2.8 (Saturation) *For any simplicial model $\mathscr{C} = (\mathscr{V},C,\chi,\ell)$ and any first-order Kripke model $\mathscr{M} = (W,\delta,\{R_a\}_{a\in\mathbf{A}},\rho)$, we say $\mathscr{C}$ (resp. $\mathscr{M}$) is* saturated*, if the following hold:*

Simplicial Model	*First-Order Kripke Model*
For any $A \subseteq \mathbf{A}$, $F \in \mathscr{F}(C)$ and $\Delta \subseteq \mathscr{L}_0^{[:=]}$, *if Δ is finitely satisfiable in $\{G \in \mathscr{F}(C) \mid A \subseteq \chi[F \cap G]\}$,* *then there is $G \in \mathscr{F}(C)$ s.t. $A \subseteq \chi[F \cap G]$ and $\mathscr{C},G \Vdash \Delta$*	*For any $A \subseteq \mathbf{A}$, $w \in W$ and $\Delta \subseteq \mathscr{L}_0^{[:=]}$,* *if Δ is finitely satisfiable in $\bigcap_{a\in A} R_a(w)$,* *then there is $v \in \bigcap_{a\in A} R_a(w)$ s.t. $\mathscr{M},v \vDash \Delta$*

Then, we can formulate the following theorem (the proof can be found in Appendix A):

Theorem 2.9 *For any saturated simplicial models $\mathscr{C},F$, $\mathscr{D},G$ and any saturated first-Kripke models $\mathscr{M},w$, $\mathscr{N},v$,*

$$\mathscr{C},F \leftrightarrows \mathscr{D},G \Longleftrightarrow \mathscr{C},F \equiv_{\mathscr{L}^{[:=]}} \mathscr{D},G \quad \textit{and} \quad \mathscr{M},w \leftrightarrows \mathscr{N},v \Longleftrightarrow \mathscr{M},w \equiv_{\mathscr{L}^{[:=]}} \mathscr{N},v$$

3 Equivalence of the Two Semantics

In [9], a correspondence between *pure* simplicial models and a certain class of Kripke models is proved, and such a correspondence is later generalized to *impure* simplicial models in [18, 6]. In this section, we prove a similar result between impure simplicial models and a certain class of *first-order* Kripke models. Though we are considering first-order Kripke models rather than propositional ones, and the simplicial complex semantics we are working with is also different from the one employed in [18, 6] (which is three-valued), the idea and formal construction are still essentially the same as [18, 6].

Imitating the terminology in [18, 6], we introduce the following notion of *local epistemic models*, which characterizes the first-order Kripke models that correspond to simplicial models.

Definition 3.1 (Local Epistemic Model) *For a first-order Kripke model $\mathscr{M} = (W,\delta,\{R_a\}_{a\in\mathbf{A}},\rho)$, we say $\mathscr{M}$ is a* local epistemic model*, if the following are satisfied:*

- *(Local S5) For all $a \in \mathbf{A}$, the restriction of R_a to $\{w \in W \mid a \in \delta(w)\}$ is an equivalence relation;*
- *(Individually Increasing Domain) For all $w,v \in W$ and $a \in \delta(w)$ s.t. wR_av, $a \in \delta(v)$;*
- *(Local Predicates) For all $p \in \mathbf{P}$, $w,v \in W$ and $a \in \rho(p,w)$ s.t. wR_av, $a \in \rho(p,v)$;*
- *(Collectively Decreasing Domain) For all $w,v \in W$ s.t. $v \in \bigcap_{a\in\delta(w)} R_a(w)$, $\delta(v) \subseteq \delta(w)$.*

We say $\mathscr{M}$ is a proper *local epistemic model, if it has the above properties plus the following one:*

- *(Properness) For all $w \in W$, $\bigcap_{a\in\delta(w)} R_a(w) = \{w\}$.*

All the above properties have rather intuitive meanings: for example, (Individually Increasing Domain) says that every agent knows their own existence, (Local Predicates) says that every agent knows their own atomic properties, and (Collectively Decreasing Domain) says that at every world, all existing agents have the distributed knowledge that no other agent exists. As we shall see in Section 4, the first four properties also correspond to axioms in our axiomatization in a natural way.

Before showing the correspondence between local epistemic models and simplicial models, we first show that our language cannot really distinguish local epistemic models and *proper* local epistemic models, in the sense that every local epistemic model has a logically equivalent "properization":

Definition 3.2 *For any local epistemic model $\mathscr{M} = (W, \delta, \{R_a\}_{a\in\mathbf{A}}, \rho)$ and $w \in W$, for convenience, let $[w]_\delta = \bigcap_{a\in\delta(w)} R_a(w)$. Then, the properization of $\mathscr{M}$ is $\mathscr{M}^{pr} = (W^{pr}, \delta^{pr}, \{R_a^{pr}\}_{a\in\mathbf{A}}, \rho^{pr})$, where $W^{pr} = \{[w]_\delta \mid w \in W\}$, $\delta^{pr}([w]_\delta) = \{([w]_\delta, \delta(w)) \mid w \in W\}$, $R_a^{pr} = \{([w]_\delta, [v]_\delta) \mid wR_av\}$ for all $a \in \mathbf{A}$, and $\rho^{pr} = \{((p, [w]_\delta), \rho(p, w)) \mid p \in \mathbf{P}, w \in W\}$.*

Proposition 3.3 *For any local epistemic model $\mathscr{M} = (W, \delta, \{R_a\}_{a\in\mathbf{A}}, \rho)$, $\mathscr{M}^{pr}$ is a proper local epistemic model s.t. for any $w \in W$, $\mathscr{M}, w \leftrightarrows \mathscr{M}^{pr}, [w]_\delta$ (and thus $\mathscr{M}, w \equiv_{\mathscr{L}^{[:=]}} \mathscr{M}^{pr}, [w]_\delta$).*

In particular, if $\mathscr{M}$ is already a proper local epistemic model, then $\mathscr{M} \cong \mathscr{M}^{pr}$.

Now, we prove the correspondence between simplicial models and local epistemic models.

The direction from simplicial models to local epistemic models is rather straightforward: we just take the facets as worlds, and the colors in a facet as its local domain. The formal construction is as follows:

Definition 3.4 *For any simplicial complex $\mathscr{C} = (\mathscr{V}, C, \chi, \ell)$, let $\mathtt{LEM}(\mathscr{C}) = (W^{\mathscr{C}}, \delta^{\mathscr{C}}, \{R_a^{\mathscr{C}}\}_{a\in\mathbf{A}}, \rho^{\mathscr{C}})$, where $W^{\mathscr{C}} = \mathscr{F}(C)$, $\delta^{\mathscr{C}} = \{(F, \chi[F]) \mid F \in W^{\mathscr{C}}\}$, $R_a^{\mathscr{C}} = \{(F, G) \in W^{\mathscr{C}} \times W^{\mathscr{C}} \mid a \in \chi[F \cap G]\}$ for all $a \in \mathbf{A}$, and $\rho^{\mathscr{C}} = \{((p, F), \chi[F \cap V(p)]) \mid p \in \mathbf{P}, F \in W^{\mathscr{C}}\}$.*

The converse direction is a bit more complicated, since we need to first construct a set of vertices from the first-order Kripke model, and then construct the facets from the vertices. The idea is to use indistinguishability cells labeled by agents as vertices. The formal construction is as follows:

Definition 3.5 *For any local epistemic model $\mathscr{M} = (W, \delta, \{R_a\}_{a\in\mathbf{A}}, \rho)$ and $w \in W$, $a \in \mathbf{A}$, for convenience, let $[w]_a = R_a(w)$, and let $F_w^{\mathscr{M}} = \{(a, [w]_a) \mid a \in \delta(w)\}$. Then, let $\mathtt{SC}(\mathscr{M}) = (\mathscr{V}^{\mathscr{M}}, C^{\mathscr{M}}, \chi^{\mathscr{M}}, \ell^{\mathscr{M}})$, where $\mathscr{V}^{\mathscr{M}} = \{(a, [w]_a) \mid w \in W, a \in \delta(w)\}$, $C^{\mathscr{M}} = \{X \subseteq \mathscr{V}^{\mathscr{M}} \mid X \subseteq F_w^{\mathscr{M}}$ for some $w \in W\} \setminus \{\emptyset\}$, $\chi^{\mathscr{M}}(a, [w]_a) = a$ for all $(a, [w]_a) \in \mathscr{V}^{\mathscr{M}}$, and $\ell^{\mathscr{M}}(p) = \{(a, [w]_a) \mid a \in \rho(p, w)\}$ for all $p \in \mathbf{P}$.*

Proposition 3.6 *For any simplicial model $\mathscr{C} = (\mathscr{V}, C, \chi, \ell)$, $\mathtt{LEM}(\mathscr{C})$ is a* proper *local epistemic model s.t. for any $F \in \mathscr{F}(C)$ and $\alpha \in \mathscr{L}_0^{[:=]}$, $\mathscr{C}, F \Vdash \alpha$ iff $\mathtt{LEM}(\mathscr{C}), F \vDash \alpha$.*

For any local epistemic model $\mathscr{M} = (W, \delta, \{R_a\}_{a\in\mathbf{A}}, \rho)$, $\mathtt{SC}(\mathscr{M})$ is a simplicial model with $\mathscr{F}(C^{\mathscr{M}}) = \{F_w^{\mathscr{M}} \mid w \in W\}$ s.t. for any $w \in W$ and $\alpha \in \mathscr{L}_0^{[:=]}$, $\mathscr{M}, w \vDash \alpha$ iff $\mathtt{SC}(\mathscr{M}), F_w^{\mathscr{M}} \Vdash \alpha$.

Finally, it is also interesting to note the following properties of LEM and SC:

Proposition 3.7 *For any simplicial model $\mathscr{C}$, $\mathscr{C} \cong \mathtt{SC}(\mathtt{LEM}(\mathscr{C}))$, and for any local epistemic model $\mathscr{M}$, $\mathscr{M}^{pr} \cong \mathtt{LEM}(\mathtt{SC}(\mathscr{M}))$ (so for any* proper *local epistemic model $\mathscr{M}$, $\mathscr{M} \cong \mathtt{LEM}(\mathtt{SC}(\mathscr{M}))$).*

PROOF. (Sketch) Given a simplicial complex $\mathscr{C} = (\mathscr{V}, C, \chi, \ell)$, we can check that the function which maps each $v \in \mathscr{V}$ to $(\chi(v), \{F \in \mathscr{F}(C) \mid v \in F\})$ is an isomorphism from $\mathscr{C}$ to $\mathtt{SC}(\mathtt{LEM}(\mathscr{C}))$.

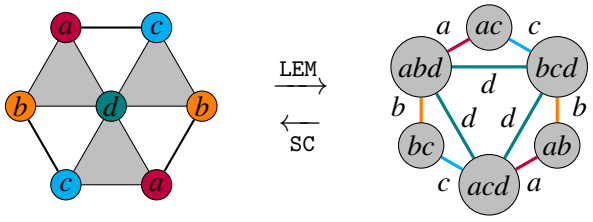
On the other hand, given a local epistemic model $\mathscr{M} = (W, \delta, \{R_a\}_{a\in\mathbf{A}}, \rho)$, we can check that if $[w]_\delta = [v]_\delta$, then $F_w^{\mathscr{M}} = F_v^{\mathscr{M}}$, so the function which maps each $[w]_\delta$ to $F_w^{\mathscr{M}}$ is a well-defined isomorphism from $\mathscr{M}^{pr}$ to $\mathtt{LEM}(\mathtt{SC}(\mathscr{M}))$. □

Proposition 3.8 *Let $\mathscr{C},F$, $\mathscr{D},G$ be arbitrary simplicial models, and let $\mathscr{M},w$, $\mathscr{N},v$ be arbitrary local epistemic models. Then, we have the following*

$$\mathscr{C},F \leftrightarrows \mathscr{D},G \iff \mathtt{LEM}(\mathscr{C}),F \leftrightarrows \mathtt{LEM}(\mathscr{D}),G \quad \text{and} \quad \mathscr{M},w \leftrightarrows \mathscr{N},v \iff \mathtt{SC}(\mathscr{M}),F_w^{\mathscr{M}} \leftrightarrows \mathtt{SC}(\mathscr{N}),F_v^{\mathscr{N}}$$

PROOF. (Sketch) We can check that if Z is a bisimulation between two simplicial models $\mathscr{C}$ and $\mathscr{D}$, then Z itself is also a bisimulation between $\mathtt{LEM}(\mathscr{C})$ and $\mathtt{LEM}(\mathscr{D})$; conversely, if Z is a bisimulation between two local epistemic models $\mathscr{M}$ and $\mathscr{N}$, then $\{(F_w^{\mathscr{M}},F_v^{\mathscr{N}}) \mid (w,v)\in Z\}$ is a bisimulation between $\mathtt{SC}(\mathscr{M})$ and $\mathtt{SC}(\mathscr{N})$. Then, the conclusion can be proved using Proposition 3.3 and 3.7. □

Below is a diagram of a simplicial model and a proper local epistemic model which correspond to each other via LEM and SC (up to isomorphism). The reflexive arrows are omitted in the diagram of the local epistemic model.



4 Axiomatization and Normal Form

In this section, we offer a strongly complete axiomatization for both semantics. Our axiomatization is different from the existing ones in interesting ways: on the one hand, our axiomatization is based on a *normal modal logic*, which distinguishes it from the one presented in [6]; on the other hand, our axiomatization is *not* based on **KB4** (it is worth noting that we have the T-axiom for knowledge in our axiomatization), which distinguishes it from the one presented in [10]. Below is our system:[8]

The System $\mathbf{LEL}^{[:=]}$

Axiom Set I		**Axiom Set II**	
TAUT	All propositional tautologies	$\mathtt{T}^{\mathsf{K}}$	$\mathsf{K}_X\alpha \to \alpha$
$\mathtt{K}^{\mathsf{K}}$	$\mathsf{K}_X(\alpha\to\beta)\to(\mathsf{K}_X\alpha\to\mathsf{K}_X\beta)$	KNI	$[\vec{x}:=\vec{a}](\neg\mathsf{K}_{\vec{x}}\alpha\to\mathsf{K}_{\vec{x}}[\vec{x}:=\vec{a}]\neg\mathsf{K}_{\vec{x}}\alpha)$
$\mathtt{MONO}^{\mathsf{K}}$	$\mathsf{K}_X\alpha\to\mathsf{K}_Y\alpha \quad (X\subseteq Y)$	EPI	$[x:=a]\mathsf{K}_x\langle x:=a\rangle\top$
$\mathtt{K}^{[:=]}$	$[x:=a](\phi\to\psi)\to([x:=a]\phi\to[x:=a]\psi)$	API	$[x:=a](p_x\to\mathsf{K}_x[x:=a]p_x)$
$\mathtt{DET}^{[:=]}$	$\langle x:=a\rangle\phi\to[x:=a]\phi$	ENI	$[\vec{x}:=\vec{a}](\bigwedge_{b\in B}[x:=b]\bot\to\mathsf{K}_{\vec{x}}\bigwedge_{b\in B}[x:=b]\bot)$ (where $B=\mathbf{A}\setminus\{\vec{a}\}$)
$\mathtt{TR}^{[:=]}$	$\phi\to[x:=a]\phi \quad (x\notin FV(\phi))$	**Rules**	
$\mathtt{SUB}^{[:=]}$	$[y:=a]([x:=a]\phi\to\phi[y/x])$ ($\phi[y/x]$ is admissible)	MP	$\dfrac{\phi,\phi\to\psi}{\psi}$ NEC$^{\mathsf{K}}$ $\dfrac{\alpha}{\mathsf{K}_\emptyset\alpha}$ NEC$^{[:=]}$ $\dfrac{\phi}{[x:=a]\phi}$
$\mathtt{COM}^{[:=]}$	$[x:=a][y:=b]\phi\to[y:=b][x:=a]\phi \quad (x\neq y)$		
$\mathtt{UI}^{[:=]}$	$\bigwedge_{a\in\mathbf{A}}[x:=a]\phi\to\phi$		

Note that **Axiom Set I** contains the basic axioms for the distributed knowledge operator and the assignment operator, among which $\mathtt{SUB}^{[:=]}$ is probably the most interesting: it says that in an "environment" where a is already assigned to y, assigning the same agent a to x *semantically* is equivalent as substituting

[8] Some notes on notations: for a string of variables $\vec{x}=x_1,\ldots,x_n$ and a string of agents $\vec{a}=a_1,\ldots,a_n$ of the same length, let $\{\vec{x}\}$, $\{\vec{a}\}$ denote $\{x_1,\ldots,x_n\}$ and $\{a_1,\ldots,a_n\}$ respectively, and let $[\vec{x}:=\vec{a}]\phi$ be the abbreviation for $[x_1:=a_1]\cdots[x_n:=a_n]\phi$. In particular, when $\vec{x}$ is the empty string, $[\vec{x}:=\vec{a}]\phi$ is just ϕ. Also let $\mathsf{K}_{\vec{x}}\alpha$ (resp. $\widehat{\mathsf{K}}_{\vec{x}}\alpha$) be the abbreviation for $\mathsf{K}_{\{\vec{x}\}}\alpha$ (resp. $\widehat{\mathsf{K}}_{\{\vec{x}\}}\alpha$).

y for x *syntatically* (which should remind the reader of the so-called *Substitution Lemma* in first-order logic). On the other hand, axioms in **Axiom Set II** characterize the special properties of local epistemic models — $\mathtt{T}^{\mathsf{K}}$ and KNI correspond to (Local $S5$), EPI corresponds to (Individually Increasing Domain), API corresponds to (Local Predicates), and ENI corresponds to (Collectively Decreasing Domain). It should be noted that $\mathbf{LEL}^{[:=]}$ is *not* closed under uniform substitution for labeled propositional letters: this is because API only holds for atomic propositions.

Below are some useful theorems of $\mathbf{LEL}^{[:=]}$ (the proof can be found in Appendix A):

Proposition 4.1 *The following are* $\mathbf{LEL}^{[:=]}$*-theorems:*

$\mathtt{R}^{[:=]}$ $[x:=a]\phi \leftrightarrow [y:=a]\phi[y/x]$ *(y not in* ϕ*)*

KPI $[\vec{x}:=\vec{a}](\mathsf{K}_{\vec{x}}\alpha \rightarrow \mathsf{K}_{\vec{x}}[\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha)$

ANI $[x:=a](\neg p_x \rightarrow \mathsf{K}_x[x:=a]\neg p_x)$

Using techniques of first-order modal logic, we can prove the following completeness theorem (again, the proof can be found in Appendix A):

Theorem 4.2 *Under first-order Kripke semantics,* $\mathbf{LEL}^{[:=]}$ *is sound and strongly complete w.r.t. the class of all (proper) local epistemic models.*

Then, with the help of the correspondence result we prove in Section 3, it follows that

Corollary 4.3 *Under simplicial complex semantics,* $\mathbf{LEL}^{[:=]}$ *is sound and strongly complete w.r.t. the class of all simplicial models.*

In the rest of this section, we show that there is a notion of *normal form* for sentences in $\mathbf{LEL}^{[:=]}$: every $\mathscr{L}^{[:=]}$-sentence is equivalent to a sentence in which the assignment operators only appear right before atomic propositions, $\bot$ and distributed knowledgeusu operators. (It should be noted that such a notion of normal form is *not* used in the completeness proof presented in Appendix A.)

To put it more formally, we define the following notion of *assignment normal form*:

$$\mathscr{L}^{[:=]}_{\mathtt{ANF}} \ni \gamma ::= [x:=a]p_x \mid [x:=a]\bot \mid \top \mid \neg\gamma \mid (\gamma \wedge \gamma) \mid [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\gamma$$

Then, we have the following proposition (the proof can be found in Appendix A):

Proposition 4.4 *For all* $\alpha \in \mathscr{L}^{[:=]}_0$*, there is* $\gamma \in \mathscr{L}^{[:=]}_{\mathtt{ANF}}$ *s.t.* $\vdash \alpha \leftrightarrow \gamma$.

Note that $[x:=a]p_x$ can simply be viewed as a propositional letter labeled by a, $[\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\gamma$ can be viewed as a distributed knowledge formula labeled by the set $\{\vec{a}\}$, and $[x:=a]\bot$ works as a *non-existence* predicate for a, which can be expressed by $\mathsf{K}_a\bot$ in the two-valued approach (or more directly, by the negation of the global atom $\neg a$ employed in some works of the three-valued approach [7, 1]). In this way, $\mathscr{L}^{[:=]}_{\mathtt{ANF}}$ can be viewed as an ordinary propositional modal language equipped with a special version of two-valued simplicial semantics, in which p_a is automatically *true* when a does *not* exist.[9]

This connects our approach and the two-valued approach in a natural way: the two-valued language can be viewed as a fragment of our language, which has the same expressivity as the latter. Under this perspective, formulas like $\mathsf{K}_{\vec{a}}\alpha$ in the two-valued language become abbreviations for $[\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha$, so for example, $\mathsf{K}_a\alpha$ actually says that "*if a exists*, then the a knows that α", rather than simply "a knows that α". Thus, some peculiarities of two-valued approach are no longer peculiar: for example, for any

[9] This is somewhat like the "opposite" of *negative free logic*, in which an atomic formula is automatically *false* when it mentions non-existing object(s). But note that as long as we can express the existential predicate in our language, in terms of expressivity, there is no essential difference between them.

ϕ, we have $\mathsf{K}_a\phi$ when a does not exist, not because a dead agent "knows" the everything, but because $[x{:=}a]\bot$ implies $[x{:=}a]\mathsf{K}_x\phi$; and the T-axiom for K_a fails because $[x{:=}a]\mathsf{K}_x\alpha \to \alpha$ is true only under the assumption $\langle x{:=}a\rangle\top$, but not because knowledge does not imply truth.

It is also worth noting that the connection between our approach and the two-valued one gives us the *decidability* of our logic. The satisfiability problem for $\mathscr{L}^{[:=]}_{\mathtt{ANF}}$ is decidable, since it can be viewed as a kind of two-valued language with an essentially **KB4**-logic, which is clearly decidable; moreover, computing the normal form of an $\mathscr{L}^{[:=]}$-sentence is also a decidable matter (as one may check in the proof of Proposition 4.4 in Appendix A). Thus, we have

Theorem 4.5 *The satisfiability problem for $\mathscr{L}^{[:=]}$-sentences is decidable.*

5 Intensional Distributed Knowledge

Since we have open formulas in our language, it is natural to define groups of agents with them. Formally, for a formula $\phi(x)$ with at most one free variable x (let $\mathscr{L}^{[:=]}_x$ denote the set of all such formulas), the group defined by $\phi(x)$ at a pointed simplicial model $\mathscr{C},F$ is $\phi(\mathscr{C},F) = \{a \in \chi^{\mathscr{C}}[F] \mid \mathscr{C},F,\sigma[x\mapsto a] \Vdash \phi(x)\}$, while the group it defines at a pointed first-order Kripke model $\mathscr{M},w$ is $\phi(\mathscr{M},w) = \{a \in \delta^{\mathscr{M}}(w) \mid \mathscr{M},w,\sigma[x\mapsto a] \vDash \phi(x)\}$ (our choice of σ does not matter here). Note that such groups are *intensional*, in the sense that the same formula $\phi(x)$ might define different groups at different facets / worlds.[10]

Then, it seems interesting to consider a kind of *intensional distributed knowledge* based on such intensional groups. Formally, for each $\phi \in \mathscr{L}^{[:=]}_x$, the intended truth conditions for $\mathsf{K}_\phi\alpha$ are the following:

$\mathscr{C},F \Vdash \mathsf{K}_\phi\alpha$	iff	for all $G \in \mathscr{F}^{\mathscr{C}}(C)$ s.t. $\phi(\mathscr{C},F) \subseteq \chi^{\mathscr{C}}[F\cap G]$, $\mathscr{C},G \Vdash \alpha$	$\mathscr{M},w \vDash \mathsf{K}_\phi\alpha$	iff	for all $v \in W^{\mathscr{M}}$ s.t. $v \in \bigcap_{a\in\phi(\mathscr{M},w)} R^{\mathscr{M}}_a(w)$, $\mathscr{M},v \vDash \alpha$

In our language, we can define K_ϕ as follows: first, for each $\phi \in \mathscr{L}^{[:=]}_x$, let $\phi!(A) = \bigwedge_{a\in A}\langle x{:=}a\rangle\phi \wedge \bigwedge_{b\in\mathbf{A}\setminus A}[x{:=}b]\neg\phi$. Intuitively, $\phi!(A)$ says that A is exactly the group defined by ϕ at the present world. Then, $\mathsf{K}_\phi\alpha$ can be defined as follows:

$$\mathsf{K}_\phi\alpha = \bigwedge_{\{\vec{a}\}\subseteq\mathbf{A}} (\phi!(\{\vec{a}\}) \to [\vec{x}{:=}\vec{a}]\mathsf{K}_{\vec{x}}\alpha)$$

It is not hard to check that the truth conditions of $\mathsf{K}_\phi\alpha$ are exactly the ones above.

The logical behavior of K_ϕ is quite interesting. The logic of K_ϕ is at least a normal modal logic with the T-axiom $\mathsf{K}_\phi\alpha \to \alpha$, but both $\mathsf{K}_\phi\alpha \to \mathsf{K}_\phi\mathsf{K}_\phi\alpha$ and $\neg\mathsf{K}_\phi\alpha \to \mathsf{K}_\phi\neg\mathsf{K}_\phi\alpha$ are *invalid* in general, even though for any rigid group $A = \{\vec{a}\}$, if we define $\mathsf{K}_A\alpha$ as $[\vec{x}{:=}\vec{a}]\mathsf{K}_{\vec{x}}\alpha$, then both $\mathsf{K}_A\alpha \to \mathsf{K}_A\mathsf{K}_A\alpha$ and $\neg\mathsf{K}_A\alpha \to \mathsf{K}_A\neg\mathsf{K}_A\alpha$ are valid (by KPI, KNI and EPI): essentially, this is because an intensional group might not know its *extension*. Nevertheless, *positive introspection* is still relatively easy to achieve for intensional groups: the following lemma tells us that when $\phi(x)$ expresses a property that can be positively introspected by the agent possessing it, K_ϕ satisfies the principle of positive introspection:

Lemma 5.1 *For any $\phi \in \mathscr{L}^{[:=]}_x$, if $\vdash [x{:=}a](\phi \to \mathsf{K}_x[x{:=}a]\phi)$ for all $a \in \mathbf{A}$, then $\vdash \mathsf{K}_\phi\alpha \to \mathsf{K}_\phi\mathsf{K}_\phi\alpha$.*

[10] The idea of introducing intensional groups to epistemic logic dates back to [12, 11], and intensional distributed knowledge has been studied in [3, 4], for example. Unlike these works, however, the intensional groups here are denoted by *open formulas* in our language, rather than a distinct set of *group names*.

PROOF. (Sketch) If $\vdash [x:=a](\phi \to \mathsf{K}_x[x:=a]\phi)$ for all $a \in \mathbf{A}$, then for any $\{\vec{a}\} \subseteq \mathbf{A}$, we can check $\vdash \phi!(\{\vec{a}\}) \to [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}} \bigvee_{\{\vec{a}\} \subseteq B \subseteq \mathbf{A}} \phi!(B)$. Moreover, $\vdash [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha \to [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}[\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha$, and for any $\{\vec{a}\} \subseteq \{\vec{b}\} \subseteq \mathbf{A}$, $\vdash [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha \to [\vec{y}:=\vec{b}]\mathsf{K}_{\vec{y}}\alpha$. Thus, $\vdash \mathsf{K}_\phi\alpha \to (\phi!(\{\vec{a}\}) \to [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\mathsf{K}_\phi\alpha)$ for all $\{\vec{a}\} \subseteq \mathbf{A}$. Then, by definition, $\vdash \mathsf{K}_\phi\alpha \to \mathsf{K}_\phi\mathsf{K}_\phi\alpha$. □

Then, we can show that for a natural kind of intensional groups, the corresponding notion of intensional distributed knowledge indeed satisfies positive introspection:

Proposition 5.2 *Fix a variable x, and let χ be a formula generated by the following scheme:*[11]

$$\chi ::= p_x \mid \mathsf{K}_x\alpha \mid \neg\chi \mid (\chi \wedge \chi)$$
$$\text{where } \alpha \in \mathscr{L}_0^{[:=]}$$

Then, for any $\beta \in \mathscr{L}_0^{[:=]}$, $\vdash \mathsf{K}_\chi\beta \to \mathsf{K}_\chi\mathsf{K}_\chi\beta$.

PROOF. (Sketch) By Lemma 5.1, it suffices to check that for any such χ, $\vdash [x:=a](\chi \to \mathsf{K}_x[x:=a]\chi)$ for any $a \in \mathbf{A}$. In order to do so, notice that every such χ can be written in *disjunctive normal form*, and when χ is p_x, $\neg p_x$, $\mathsf{K}_x\alpha$ or $\neg\mathsf{K}_x\alpha$, by API, ANI, KPI and KNI, $\vdash [x:=a](\chi \to \mathsf{K}_x[x:=a]\chi)$ for any $a \in \mathbf{A}$. □

On the other hand, *negative introspection* is much more difficult to achieve. Some rather boring intensional groups indeed have distributed knowledge with negative introspection, e.g. the group defined by $\bot$ and the group defined by $\top$. The former is too weak, since $\mathsf{K}_\bot\alpha$ is just $\mathsf{K}_\emptyset\alpha$; and the latter is too strong: $\mathsf{K}_\top\alpha$ expresses the distributed knowledge of *all existing agents*, so $\alpha \to \mathsf{K}_\top\alpha$ is actually valid for any sentence α.[12] More generally, for any sentence α, the distributed knowledge of the intensional group defined by $\mathsf{K}_\emptyset\alpha$ also have negative introspection, since in each model, it either always defines the empty group, or always defines the group of all existing agents. Unfortunately, we have not yet found any interesting $\phi \in \mathscr{L}_x^{[:=]}$ that validates $\neg\mathsf{K}_\phi\alpha \to \mathsf{K}_\phi\neg\mathsf{K}_\phi\alpha$.

6 Conclusion

In this paper, we introduced a term-modal language with assignment operators to study epistemic logic on impure simplicial complexes. We characterized the expressivity of the simplicial semantics and first-order Kripke semantics of this language, showed the correspondence between simplicial complexes and local epistemic models, offered a complete axiomatization for the epistemic logic based on our language, showed that the our language has assignment normal form, and also studied the behavior of the intensional distributed knowledge operators, which can be defined in our language.

There are a number of interesting directions for future work. For example, it is interesting to further study the connection between our simplicial semantics and the existing approaches, especially the *three-valued* one. Though sentences in our language are two-valued, there might be some natural way to generate a three-valued semantics from our semantics. If this can be done, then it should also be interesting to see whether such a three-valued semantics coincides with the existing ones.

[11] It is interesting to note the truth value of a formula χ generated this way can be defined on *vertices* instead of *facets*: notice that for any $\mathscr{C} = (V, C, \chi, \ell)$ and $v \in V$, $F, G \in \mathscr{F}(C)$, if $v \in F \cap G$ and $\sigma(x) = \chi(v)$, then $\mathscr{C}, F, \sigma \Vdash \chi$ iff $\mathscr{C}, G, \sigma \Vdash \chi$; thus, we are justified to simply write $\mathscr{C}, v \Vdash \chi$ instead.

[12] This is easy to check by the semantics, especially by the simplicial complex semantics. However, deducing $\alpha \to \mathsf{K}_\top\alpha$ in $\mathbf{LEL}^{[:=]}$ is not easy. We need to make use of the fact that every formula in $\mathscr{L}^{[:=]}$ has an assignment normal form, and then show that for each γ in normal form, we indeed have $\vdash \gamma \to \mathsf{K}_\top\gamma$.

We can also use our language to do epistemic logic on more general kinds of structures. On the one hand, properties of local epistemic models clearly can be relaxed; on the other hand, it is also interesting to consider *face-semantics* rather than *facet-semantics*, like the one studied in [7], or more general simplicial structures based on *semi-simplicial sets*, like the one studied in [5].

The notion of intensional distributed knowledge also seems interesting. Our results concerning such a notion of knowledge presented in this paper are rather primitive, and understanding the behavior of such intensional distributed knowledge operators might help us better understand the structure of the definable groups in such models.

Finally, it should also be interesting to generalize the kind of assignment operators we introduced in this paper to study varying-domain first-order modal logic in general. This might provide us with an interesting first-order modal logic alternative to free logic-based varying-domain first-order modal logic.

Acknowledgement

I would like to thank Yanjing Wang and Hans van Ditmarsch for their advice on the early versions of this paper. I would also like to thank three anonymous reviewers from TARK 2025 for their insightful comments on this paper.

References

[1] Marta Bílková, Hans van Ditmarsch, Roman Kuznets & Rojo Randrianomentsoa (2024): *Bisimulation for Impure Simplicial Complexes*. In Agata Ciabattoni, David Gabelaia & Igor Sedlár, editors: *Advances in Modal Logic, AiML 2024, Prague, Czech Republic, August 19-23, 2024*, College Publications, pp. 225–248.

[2] Ofer Biran, Shlomo Moran & Shmuel Zaks (1990): *A combinatorial characterization of the distributed 1-solvable tasks*. *Journal of Algorithms* 11(3), pp. 420–440, doi:10.1016/0196-6774(90)90020-F.

[3] Marta Bílková, Zoé Christoff & Olivier Roy (2021): *Revisiting Epistemic Logic with Names*. *Electronic Proceedings in Theoretical Computer Science* 335, p. 39–54, doi:10.4204/eptcs.335.4.

[4] Marta Bílková & Igor Sedlár (2023): *Epistemic Logics of Structured Intensional Groups*. *Electronic Proceedings in Theoretical Computer Science* 379, p. 113–130, doi:10.4204/eptcs.379.11.

[5] Christian Cachin, David Lehnherr & Thomas Studer (2025): *Synergistic knowledge*. *Theoretical Computer Science* 1023, p. 114902, doi:10.1016/j.tcs.2024.114902.

[6] Hans van Ditmarsch & Roman Kuznets (2024): *Wanted dead or alive: epistemic logic for impure simplicial complexes*. *Journal of Logic and Computation*, p. exae055, doi:10.1093/logcom/exae055.

[7] Hans van Ditmarsch, Roman Kuznets & Rojo Randrianomentsoa (2023): *On two- and three-valued semantics for impure simplicial complexes*. In: *Electronic Proceedings in Theoretical Computer Science*, 390, pp. 50–66, doi:10.4204/EPTCS.390.4.

[8] Michael Fischer, Nancy Lynch & Mike Paterson (1985): *Impossibility of Distributed Consensus with One Faulty Process*. *Journal of the ACM* 32, pp. 374–382, doi:10.1145/3149.214121.

[9] Éric Goubault, Jérémy Ledent & Sergio Rajsbaum (2021): *A simplicial complex model for dynamic epistemic logic to study distributed task computability*. *Information and Computation* 278, p. 104597, doi:10.1016/j.ic.2020.104597. Special Issue on Ninth International Symposium on Games, Automata, Logics and Formal Verification (GandALF 2018).

[10] Éric Goubault, Jérémy Ledent & Sergio Rajsbaum (2022): *A Simplicial Model for $KB4_n$: Epistemic Logic with Agents that May Die*. arXiv:2108.10293.

[11] Adam J. Grove (1995): *Naming and Identity in Epistemic Logic Part II: A First-Order Logic for Naming*. *Artif. Intell.* 74, pp. 311–350, doi:10.1016/0004-3702(95)98593-D.

[12] 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. arXiv:https://academic.oup.com/logcom/article-pdf/3/4/345/9875391/3-4-345.pdf.

[13] Maurice Herlihy, Dmitry Kozlov & Sergio Rajsbaum (2013): *Distributed Computing Through Combinatorial Topology*, 1st edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, doi:10.1016/B978-0-12-404578-1.00017-6.

[14] Barteld Kooi (2008): *Dynamic term-modal logic*. In: *van Benthem, J., Ju, S. and Veltman, F. (eds.). A meeting of the minds. Proceedings of the workshop on Logic, Rationality and Interaction, Beijing , 2007 Texts in Computing Computer Science 8. pp. 173-185 College publications , London*, 8, pp. 173–185.

[15] Michael C Loui & Hosame H Abu-Amara (1987): *Memory requirements for agreement among unreliable asynchronous processes*. Advances in Computing research 4(163-183), p. 31.

[16] PAVEL NAUMOV & JIA TAO (2019): *EVERYONE KNOWS THAT SOMEONE KNOWS: QUANTIFIERS OVER EPISTEMIC AGENTS*. The Review of Symbolic Logic 12(2), p. 255–270, doi:10.1017/S1755020318000497.

[17] Robert C. Stalnaker & Richmond H. Thomason (1968): *Abstraction in First-Order Modal Logic*. Theoria 34(3), pp. 203–207, doi:10.1111/j.1755-2567.1968.tb00351.x.

[18] Hans van Ditmarsch (2021): *Wanted Dead or Alive: Epistemic Logic for Impure Simplicial Complexes*. In: *Logic, Language, Information, and Computation: 27th International Workshop, WoLLIC 2021, Virtual Event, October 5–8, 2021, Proceedings*, Springer-Verlag, Berlin, Heidelberg, p. 31–46, doi:10.1007/978-3-030-88853-4_3.

[19] Hans van Ditmarsch, Éric Goubault, Marijana Lazić, Jérémy Ledent & Sergio Rajsbaum (2021): *A dynamic epistemic logic analysis of equality negation and other epistemic covering tasks*. Journal of Logical and Algebraic Methods in Programming 121, p. 100662, doi:10.1016/j.jlamp.2021.100662.

[20] Yanjing Wang, Yu Wei & Jeremy Seligman (2022): *Quantifier-free epistemic term-modal logic with assignment operator*. Ann. Pure Appl. Log. 173(3), p. 103071, doi:10.1016/J.APAL.2021.103071.

A Appendix

A.1 Proof of Theorem 2.9

PROOF. We only consider the case for simplicial models.

Let $\mathscr{C}$ and $\mathscr{D}$ be arbitrary saturated simplicial models, and let $Z = \{(F,G) \in \mathscr{F}(S^{\mathscr{C}}) \times \mathscr{F}(C^{\mathscr{D}}) \mid \mathscr{C},F \equiv_{\mathscr{L}^{[:=]}} \mathscr{D},G\}$. We show that Z is a bisimulation between $\mathscr{C}$ and $\mathscr{D}$: Let $(F,G) \in Z$ be arbitrary. We only check that (Inv) and (Zig) hold.

For (Inv): Let $A = \chi^{\mathscr{C}}[F]$, and consider the sentence $\alpha_A := \bigwedge_{a\in A}\langle x:=a\rangle\top \wedge \bigwedge_{b\in \mathbf{A}\setminus A}[x:=b]\bot$. Clearly $\mathscr{C},F \Vdash \alpha_A$, so $\mathscr{D},G \Vdash \alpha_A$, and thus $\chi^{\mathscr{D}}[G] = A = \chi^{\mathscr{C}}[F]$. Then, for any $p \in \mathbf{P}$, let $A_p = \chi^{\mathscr{C}}[F \cap \ell^{\mathscr{C}}(p)]$, and consider the sentence $\alpha_{A_p} := \bigwedge_{a\in A_p}\langle x:=a\rangle p_x \wedge \bigwedge_{b\in \mathbf{A}\setminus A_p}[x:=b]\neg p_x$. Again, $\mathscr{C},F \Vdash \alpha_{A_p}$, so $\mathscr{D},G \Vdash \alpha_{A_p}$, and thus $\chi^{\mathscr{D}}[G \cap \ell^{\mathscr{D}}(p)] = A_p = \chi^{\mathscr{C}}[F \cap \ell^{\mathscr{C}}(p)]$.

For (Zig): Suppose (towards a contradiction) that there are $A \subseteq \mathbf{A}$, $F' \in \mathscr{F}(C^{\mathscr{C}})$ s.t. $A \subseteq \chi^{\mathscr{C}}[F \cap F']$, but for any $G' \in \mathscr{F}(C^{\mathscr{D}})$ s.t. $A \subseteq \chi^{\mathscr{D}}[G \cap G']$, $(F',G') \notin Z$. Then, by the definition of Z, for every $G' \in \mathscr{F}(C^{\mathscr{D}})$ s.t. $A \subseteq \chi^{\mathscr{D}}[G \cap G']$, there is a sentence $\alpha_{G'}$ s.t. $\mathscr{C},F' \Vdash \alpha_{G'}$ but $\mathscr{D},G' \nVdash \alpha_{G'}$. Then, since $\mathscr{D}$ is saturated, there is some sentence β s.t. $\mathscr{C},F' \Vdash \beta$ but $\mathscr{D},G' \nVdash \alpha$ for all $G' \in \mathscr{F}(C^{\mathscr{D}})$ s.t. $A \subseteq \chi^{\mathscr{D}}[G \cap G']$. Assume that $A = \{a_1,...,a_n\}$, and let $X = \{x_1,...,x_n\}$. Then, $\mathscr{D},G \nVdash \langle x_1:=a_1\rangle \cdots \langle x_n:=a_n\rangle \widehat{\mathsf{K}}_X\beta$ but $\mathscr{C},F \Vdash \langle x_1:=a_1\rangle \cdots \langle x_n:=a_n\rangle \widehat{\mathsf{K}}_X\beta$, causing a contradiction. □

A.2 Proof of Proposition 4.1

PROOF. $\mathtt{R}^{[:=]}$: Assume that y is not in ϕ. By $\mathtt{SUB}^{[:=]}$, $\vdash [y:=a]([x:=\phi]\phi \to \phi[y/x])$. Then, by $\mathtt{K}^{[:=]}$, $\vdash [y:=a][x:=a]\phi \to [y:=a]\phi[y/x]$. Finally, by $\mathtt{TR}^{[:=]}$, $\vdash [x:=a]\phi \to [y:=a][x:=a]\phi$, so $\vdash [x:=a]\phi \to [y:=a]\phi[y/x]$. The converse direction is similar.

KPI: First, by $\mathtt{SUB}^{[:=]}$ and $\mathtt{COM}^{[:=]}$, we have $\vdash [\vec{x}:=\vec{a}](\mathsf{K}_{\vec{x}}\alpha \to \langle \vec{x}:=\vec{a}\rangle \mathsf{K}_{\vec{x}}\alpha)$. Then, by $\mathtt{T}^{\mathsf{K}}$, we have $\vdash [\vec{x}:=\vec{a}](\mathsf{K}_{\vec{x}}\alpha \to \widehat{\mathsf{K}}_{\vec{x}}\langle \vec{x}:=\vec{a}\rangle \mathsf{K}_{\vec{x}}\alpha)$. Hence, by KNI, $\vdash [\vec{x}:=\vec{a}](\mathsf{K}_{\vec{x}}\alpha \to \mathsf{K}_{\vec{x}}[\vec{x}:=\vec{a}]\widehat{\mathsf{K}}_{\vec{x}}\langle \vec{x}:=\vec{a}\rangle \mathsf{K}_{\vec{x}}\alpha)$. Finally, by KNI and $\mathtt{K}^{[:=]}$, $\vdash [\vec{x}:=\vec{a}]\widehat{\mathsf{K}}_{\vec{x}}\langle \vec{x}:=\vec{a}\rangle \mathsf{K}_{\vec{x}}\alpha \to [\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha$, so $\vdash [\vec{x}:=\vec{a}](\mathsf{K}_{\vec{x}}\alpha \to \mathsf{K}_{\vec{x}}[\vec{x}:=\vec{a}]\mathsf{K}_{\vec{x}}\alpha)$.

ANI: By API and $\mathtt{K}^{[:=]}$, $\vdash \langle x:=a\rangle p_x \to \langle x:=a\rangle \mathsf{K}_x p_x$, so $\vdash [x:=a](\widehat{\mathsf{K}}_x\langle x:=a\rangle p_x \to \widehat{\mathsf{K}}_x\langle x:=a\rangle \mathsf{K}_x p_x)$. By KNI and $\mathtt{T}^{\mathsf{K}}$, $\vdash [x:=a](\widehat{\mathsf{K}}_x\langle x:=a\rangle \mathsf{K}_x p_x \to p_x)$. Thus, we have $\vdash [x:=a](\widehat{\mathsf{K}}_x\langle x:=a\rangle p_x \to p_x)$, so $\vdash [x:=a](\neg p_x \to \mathsf{K}_x[x:=a]\neg p_x)$ by contraposition. □

A.3 Proof of Theorem 4.2

Now, we offer a proof for Theorem 4.2. The strategy is to first construct a kind of canonical *quasi* model, and then use tree-unraveling to turn it into a real model.

As preparation, we first add all agents to our language as new variables: let $\mathscr{L}_{\mathbf{A}}^{[:=]}$ be the language based on the variable set $\mathbf{X} \cup \mathbf{A}$. In the following constructions, the agents will work as the canonical names of themselves. Then, in order to construct the canonical model, we introduce following notions:

Definition A.1 *A set Δ of $\mathscr{L}_{\mathbf{A}}^{[:=]}$-formulas is an FV-*maximal consistent set *(FV-*MCS *for short), if Δ is consistent, and for all $\phi \in \mathscr{L}_{\mathbf{A}}^{[:=]}$ s.t. $FV(\phi) \subseteq FV(\Delta)$, either $\phi \in \Delta$ or $\neg\phi \in \Delta$.*

For an FV-MCS Δ of $\mathscr{L}_{\mathbf{A}}^{[:=]}$-formulas and an assignment $\sigma : FV(\Delta) \to \mathbf{A}$, we say (Δ, σ) is coherent, *if (i) for all $a \in \mathbf{A} \cap FV(\Delta)$, $\sigma(a) = a$, (ii) for all $x \in FV(\Delta)$, $\langle x:=\sigma(x)\rangle\Delta$ is consistent,*[13] *and (iii) if $\langle x:=a\rangle\top \in \Delta$, then $a \in FV(\Delta)$.*

We can show that a coherent pair (Δ, σ) has the following good properties:

Lemma A.2 *For any FV-MCS Δ and assignment $\sigma : FV(\Delta) \to \mathbf{A}$ s.t. (Δ, σ) is coherent,*

(i) If $\phi \in \Delta$ and $x \in FV(\Delta)$, then $\langle x:=\sigma(x)\rangle\phi \in \Delta$;

(ii) If $[x:=a]\phi \in \Delta$, $a \in FV(\Delta)$ and $\phi[a/x]$ is admissible, then $\phi[a/x] \in \Delta$.

PROOF. Assume that (Δ, σ) is coherent.

(i): Assume that $\phi \in \Delta$ and $x \in FV(\Delta)$, and suppose (towards a contradiction) that $\langle x:=\sigma(x)\rangle\phi \notin \Delta$. Note that $FV(\langle x:=\sigma(x)\rangle\phi) \subseteq FV(\phi) \subseteq \Delta$, so $\neg\langle x:=\sigma(x)\rangle\phi \in \Delta$. Then, $\langle x:=\sigma(x)\rangle(\phi \wedge \neg\langle x:=\sigma(x)\rangle\phi)$ is consistent by coherence, contradicting $\mathtt{SUB}^{[:=]}$.

For (ii): Assume that $[x:=a]\phi \in \Delta$, $a \in FV(\Delta)$ and $\phi[a/x]$ is admissible, and suppose (towards a contradiction) that $\phi[a/x] \notin \Delta$. Since $a \in FV(\Delta)$ and $\phi[a/x] \notin \Delta$, $\neg\phi[a/x] \in \Delta$. Then, by coherence, $\langle a:=a\rangle([x:=a]\phi \wedge \neg\phi[a/x])$ is consistent, contradicting $\mathtt{SUB}^{[:=]}$. □

Then, we show that every consistent set of $\mathscr{L}_{\mathbf{A}}^{[:=]}$-formulas containing no free occurrence of elements in $\mathbf{A}$ can be extended into an FV-MCS Δ paired with an assignment σ s.t. (Δ, σ) is coherent.

Lemma A.3 *For any consistent $\Gamma \subseteq \mathscr{L}_{\mathbf{A}}^{[:=]}$ s.t. $\mathbf{A} \cap FV(\Gamma) = \emptyset$, there is an FV-MCS Δ and a assignment σ s.t. $\Gamma \subseteq \Delta$ and (Δ, σ) is coherent.*

[13] By "$\langle x:=a\rangle\Delta$ is consistent", we mean that for any finite $\Delta_0 \subseteq \Delta$, $\langle x:=a\rangle \bigwedge \Delta_0$ is consistent.

PROOF. Let $\Gamma \subseteq \mathscr{L}_{\mathbf{A}}^{[:=]}$ be arbitrary, and assume that Γ is consistent and $\mathbf{A} \cap FV(\Gamma) = \emptyset$. We first extend Γ into a consistent set $\Delta_0 \subseteq \mathscr{L}^{[:=]}$ s.t. for all $a \in \mathbf{A}$, either $[x:=a]\bot \in \Delta_0$ or $\langle x:=a\rangle\top \in \Delta_0$. For convenience, let $\{a \in \mathbf{A} \mid \langle x:=a\rangle\top \in \Delta_0\} = \{a_0, ..., a_m\}$, and let $\vec{a} = a_0, ..., a_m$.

Then, enumerate all variables in $FV(\Delta_0)$ as $\{x_k \mid k \in \omega\}$ (possibly with repetitions since $FV(\Delta_0)$ might be finite). We construct a sequence $(\sigma_k)_{k\in\omega}$ s.t. for all $k \in \omega$, σ_k is a function from $\{x_0, ..., x_{k-1}\}$ to $\mathbf{A}$ and $\langle \vec{a}:=\vec{a}\rangle\langle x_0 := \sigma_k(x_0)\rangle \cdots \langle x_{k-1} := \sigma_k(x_{k-1})\rangle\Delta_0$ is consistent.

As the starting point, Let $\sigma_0 = \emptyset$. We show that it has the properties we need.

Assume that $\langle \vec{a}:=\vec{a}\rangle \bigwedge \Theta$ is inconsistent for some finite $\Theta \subseteq \Delta_0$, i.e. $\vdash [\vec{a}:=\vec{a}]\neg \bigwedge \Theta$. Then, since $\langle x:=a_0\rangle\top, ..., \langle x:=a_m\rangle\top \in \Delta_0$, $\Delta_0 \vdash \langle \vec{a}:=\vec{a}\rangle\neg \bigwedge \Theta$. However, since $a_0, ..., a_m$ are not in $FV(\Delta_0)$, by $\mathtt{TR}^{[:=]}$, $\vdash \langle \vec{a}:=\vec{a}\rangle\neg \bigwedge \Theta \to \neg \bigwedge \Theta$. Hence, $\Delta_0 \vdash \neg \bigwedge \Theta$, contradicting that Δ_0 is consistent and $\Theta \subseteq \Delta$.

Next, given σ_k, we show how to construct σ_{k+1}. If $x_k \in \mathrm{dom}(\sigma_k)$, then just let $\sigma_{k+1} = \sigma_k$. Otherwise, we show that there is some $b \in \mathbf{A}$ s.t. $\sigma_k \cup \{(x_k, b)\}$ has the property we need.

Suppose not. For simplicity, let $\vec{x} = x_0, ..., x_{k-1}$. Then, for all $b \in \mathbf{A}$, there is some finite $\Theta_b \subseteq \Delta_0$ s.t. $\langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle\langle x_k := b\rangle \bigwedge \Theta_b$ is inconsistent. By $\mathtt{COM}^{[:=]}$, it follows that $\langle x_k := b\rangle\langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle \bigwedge \Theta_b$ is inconsistent. Then, let $\Theta = \bigcup_{b\in\mathbf{A}} \Theta_b$. Clearly $\vdash \bigvee_{b\in\mathbf{A}} \langle x_k := b\rangle\langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle \bigwedge \Theta \to \bot$. Moreover, by $\mathtt{UI}^{[:=]}$, $\vdash \langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle \bigwedge \Theta \to \bigvee_{b\in\mathbf{A}} \langle x_k := b\rangle\langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle \bigwedge \Theta$. Thus, $\langle \vec{a}:=\vec{a}\rangle\langle \vec{x}:=\sigma_k(\vec{x})\rangle \bigwedge \Theta$ is inconsistent, causing a contradiction.

Therefore, we can find some $b \in \mathbf{A}$ with the intended property. Then, let $\sigma_{k+1} = \sigma_k \cup \{(x_k, b)\}$.

Next, let $\sigma = \{(a,a) \mid a \in \mathbf{A}_0\} \cup \bigcup_{k\in\omega} \sigma_k$. We then construct a sequence $(\Delta_k)_{k\in\omega}$ of $\mathscr{L}_{\mathbf{A}}^{[:=]}$-formula sets s.t. for all $k \in \omega$, $FV(\Delta_k) \subseteq FV(\Delta_0) \cup \mathbf{A}_0$, and for all $\vec{x} \in FV(\Delta_0) \cup \mathbf{A}_0$, $\langle \vec{x}:=\sigma(\vec{x})\rangle\Delta_k$ is consistent. As preparation, we first enumerate all ϕ in $\mathscr{L}_{\mathbf{A}}^{[:=]}$ s.t. $FV(\phi) \subseteq FV(\Delta_0) \cup \mathbf{A}_0$ as $\{\phi_k \mid k \in \omega\}$. Δ_0 is already constructed, and $(\Delta_k)_{k>0}$ is constructed recursively as follows:

Given Δ_k with the intended property, we show that at least one of $\Delta_k \cup \{\phi_k\}$ and $\Delta_k \cup \{\neg\phi_k\}$ has the intended property. Suppose not. Then, there are $\vec{x}, \vec{x}' \in FV(\Delta_k)$ and some finite $\Theta \subseteq \Delta_k$, $\Theta' \subseteq \Delta_k$ s.t. $\langle \vec{x}:=\sigma(\vec{x})\rangle(\bigwedge \Theta \wedge \neg\phi_k)$ and $\langle \vec{x}':=\sigma(\vec{x}')\rangle(\bigwedge \Theta' \wedge \phi_k)$ are inconsistent, so $\vdash [\vec{x}:=\sigma(\vec{x})][\vec{x}':=\sigma(\vec{x}')](\bigwedge \Theta \to \phi_k)$ and $\vdash [\vec{x}:=\sigma(\vec{x})][\vec{x}':=\sigma(\vec{x}')](\bigwedge \Theta' \to \neg\phi_k)$ by $\mathtt{NEC}^{[:=]}$ and $\mathtt{COM}^{[:=]}$. Thus, $\langle \vec{x}:=\sigma(\vec{x})\rangle\langle \vec{x}':=\sigma(\vec{x}')\rangle \bigwedge(\Theta \cup \Theta')$ is inconsistent, causing a contradiction. Thus, if $\Delta_k \cup \{\phi_k\}$ has the property we need, then let $\Delta_{k+1} = \Delta_k \cup \{\phi_k\}$; otherwise, let $\Delta_{k+1} = \Delta_k \cup \{\phi_{k+1}\}$.

Finally, let $\Delta = \bigcup_{k\in\omega} \Delta_k$. It is easy to check that Δ is an FV-MCS and (Δ, σ) is coherent. □

Then, we define the canonical *quasi* model as follows:

Definition A.4 (Canonical Quasi-Model) *$\mathscr{M}^c = (W^c, \delta^c, \{R_A^c\}_{A\subseteq\mathbf{A}}, \rho^c)$ is defined as follows:*

- *$W^c = \{(\Delta, \sigma) \mid \Delta$ is an FV-MCS in $\mathscr{L}_{\mathbf{A}}^{[:=]}, \sigma : FV(\Delta) \to \mathbf{A}$ and (Δ, σ) is coherent$\}$;*
- *For all $(\Delta, \sigma) \in W^c$, $\delta^c(\Delta, \sigma) = \sigma[FV(\Delta)] = \mathbf{A} \cap FV(\Delta)$ (by coherence);*
- *For all $A \subseteq \mathbf{A}$ and $(\Delta, \sigma), (\Theta, \tau) \in \Theta$, $(\Delta, \sigma)R_A^c(\Theta, \tau)$ iff $A \subseteq FV(\Delta)$ and $\mathsf{K}_A\alpha \in \Delta$ implies $\alpha \in \Theta$;*
- *For all $p \in \mathbf{P}$ and $(\Delta, \sigma) \in W^c$, $\rho^c(p, (\Delta, \sigma)) = \{a \mid p_a \in \Delta\}$.*

Note that by $\mathtt{UI}^{[:=]}$, $\bigvee_{a\in\mathbf{A}} \langle x:=a\rangle\top$ is an $\mathbf{LEL}^{[:=]}$-theorem, so for all $(\Delta, \sigma) \in W^c$, $\delta^c(\Delta, \sigma) \neq \emptyset$.

Then, we can prove the *Existence Lemma* for K_X in the usual way:

Lemma A.5 (Existence) *For all $(\Delta, \sigma) \in W^c$, if $\widehat{\mathsf{K}}_A\alpha \in \Delta$, then there is $(\Theta, \tau) \in W^c$ s.t. $(\Delta, \sigma)R_A^c(\Theta, \tau)$ and $\alpha \in \Theta$.*

PROOF. (Sketch) For any $(\Delta, \sigma) \in W^c$ and $\widehat{\mathsf{K}}_A\alpha \in \Delta$, it is routine to show that $\{\alpha\} \cup \{\beta \mid \mathsf{K}_A\beta \in \Delta\}$ is consistent and has no free variable. Thus, by Lemma A.3, it can be extended into some $(\Theta, \tau) \in W^c$,

and clearly $(\Delta,\sigma)R^c_A(\Theta,\tau)$. □

Using the axioms $\mathtt{MONO}^{\mathsf{K}}$, $\mathtt{T}^{\mathsf{K}}$ $\mathtt{KNI}$, $\mathtt{EPI}$, $\mathtt{API}$ and $\mathtt{ENI}$, we can also check the following properties of $\mathscr{M}^c$, which will be useful when we construct the tree-unraveling of $\mathscr{M}^c$:

Lemma A.6 *For all $(\Delta,\sigma),(\Theta,\tau)\in W^c$, $A\subseteq B\subseteq \mathbf{A}$ and $a\in\mathbf{A}$, the following hold:*

(i) If $(\Delta,\sigma)R^c_B(\Theta,\tau)$, then $(\Delta,\sigma)R^c_A(\Theta,\tau)$;

(ii) The restriction of R^c_A to $\{(\Delta,\sigma)\in W^c \mid A\subseteq FV(\Delta)\}$ is an equivalence relation;

(iii) If $(\Delta,\sigma)R^c_{\{a\}}(\Theta,\tau)$, then $a\in FV(\Theta)$;

(iv) If $(\Delta,\sigma)R^c_{\{a\}}(\Theta,\tau)$ and $p_a\in\Delta$, then $p_a\in\Theta$.

(v) If $(\Delta,\sigma)R^c_{\delta^c(\Delta,\sigma)}(\Theta,\tau)$, then $\delta^c(\Theta,\tau)\subseteq\delta^c(\Delta,\sigma)$.

PROOF. (i) is quite routine to prove.

(ii): For simplicity, let $A=\{\vec{a}\}$. By Lemma A.2 and the axiom $\mathtt{KNI}$, for any $(\Delta,\sigma)\in W^c$ s.t. $A\subseteq FV(\Delta)$ and any sentence α, $\widehat{\mathsf{K}}_A\langle\vec{x}:=\vec{a}\rangle\mathsf{K}_{\vec{x}}\alpha\to\mathsf{K}_A\alpha\in\Delta$. Thus, for any $(\Delta,\sigma),(\Delta',\sigma'),(\Delta'',\sigma'')\in\{(\Delta,\sigma)\in W^c\mid A\subseteq FV(\Delta)\}$ s.t. $(\Delta,\sigma)R^c_A(\Delta',\sigma')$ and $(\Delta,\sigma)R^c_A(\Delta'',\sigma'')$, if $\mathsf{K}_A\alpha\in\Delta'$, then $\langle\vec{a}:=\vec{a}\rangle\mathsf{K}_A\alpha\in\Delta'$ (by Lemma A.2), so $\widehat{\mathsf{K}}_A\langle\vec{a}:=\vec{a}\rangle\mathsf{K}_A\alpha\in\Delta$, and thus $\mathsf{K}_A\alpha\in\Delta$, which implies that $\alpha\in\Delta''$. Therefore, $(\Delta',\sigma')R^c_A(\Delta'',\sigma'')$. Moreover, by $\mathtt{T}^{\mathsf{K}}$, for any $(\Delta,\sigma)\in W^c$ s.t. $A\subseteq\Delta$, clearly $(\Delta,\sigma)R^c_A(\Delta,\sigma)$. Hence, the restriction of R^c_A to $\{(\Delta,\sigma)\in W^c\mid A\subseteq FV(\Delta)\}$ is an equivalence relation.

(iii): If $a\in FV(\Delta)$, then by Lemma A.2 and $\mathtt{EPI}$, $\mathsf{K}_{\{a\}}\langle x:=a\rangle\top\in\Delta$, so for any $(\Theta,\tau)\in W^c$ s.t. $(\Delta,\sigma)R^c_{\{a\}}(\Theta,\tau)$, we have $\langle x:=a\rangle\top\in\Theta$, and thus $a\in FV(\Theta)$ by coherence.

(iv): If $p_a\in\Delta$, then $a\in FV(\Delta)$, so by Lemma A.2 and $\mathtt{API}$, $p_a\in\mathsf{K}_{\{a\}}[x:=a]p_x\in\Delta$, and thus $\mathsf{K}_{\{a\}}[x:=a]p_x\in\Delta$. Hence, for any $(\Theta,\tau)\in W^c$ s.t. $(\Delta,\sigma)R^c_{\{a\}}(\Theta,\tau)$, $[x:=a]p_x\in\Theta$; moreover, by (iii), $a\in FV(\Theta)$, so $p_a\in\Theta$ by Lemma A.2.

(v): Let $A=\delta^c(\Delta,\sigma)$ and $B=\mathbf{A}\setminus\delta^c(\Delta,\sigma)$. By coherence, for all $b\in B$, $[x:=b]\bot\in\Delta$. Thus, by Lemma A.2 and $\mathtt{ENI}$, clearly $\mathsf{K}_A\bigwedge_{b\in B}[x:=b]\bot\in\Delta$. Hence, for any $(\Theta,\tau)\in W^c$ s.t. $(\Delta,\sigma)R^c_{\delta^c(\Delta,\sigma)}(\Theta,\tau)$, $\bigwedge_{b\in B}[x:=b]\bot\in\Theta$, so $B\cap FV(\Theta)=\emptyset$ by Lemma A.2. Hence, $\delta^c(\Theta,\tau)\subseteq\delta^c(\Delta,\sigma)$. □

Now, we introduce the canonical unraveling of the canonical quasi-model.

Definition A.7 (Canonical Unraveling) *The canonical unraveling of the canonical quasi model is $\mathscr{U}^c=(W^{uc},\delta^{uc},\{R^{uc}_a\}_{a\in\mathbf{A}},\rho^{uc})$, where*

- *$W^{uc}\subseteq(W^c\times\wp(\mathbf{A}))^{<\omega}\times W^c$, and $(\Delta_0,\sigma_0)A_0\cdots A_{n-1}(\Delta_n,\sigma_n)\in W^c_{uc}$ iff for all $k<n$, the following hold: $A_k\subseteq\mathbf{A}$ and $(\Delta_k,\sigma_k)R^c_{A_k}(\Delta_{k+1},\sigma_{k+1})$.*
  For convenience, for any $s=(\Delta_0,\sigma_0)A_0\cdots A_{n-1}(\Delta_n,\sigma_n)\in W^{uc}$, we use Δ_s and σ_s to denote Δ_n and σ_n, respectively.
- *For all $s\in W^{uc}$, $\delta^{uc}(s)=\delta^c(\Delta_s,\sigma_s)$.*
- *For all $s,t\in W^{uc}$, $sR^{uc}_a t$ iff s and t are of the following forms:*
  - *$s=(\Delta_0,\sigma_0)A_0\cdots A_{n-1}(\Delta_n,\sigma_n)B_0(\Theta_1,\tau_1)B_1\cdots B_{m-1}(\Theta_m,\tau_m)$*
  - *$t=(\Delta_0,\sigma_0)A_0\cdots A_{n-1}(\Delta_n,\sigma_n)B'_0(\Theta'_1,\tau'_1)B'_1\cdots B'_{m'-1}(\Theta'_{m'},\tau'_{m'})$*

  where $a\in\delta^c((\Delta_n,\sigma_n))\cap\bigcap_{i<m}B_i\cap\bigcap_{j<m'}B'_j$ (we allow n, m and m' to be 0);
- *For all $p\in\mathbf{P}$ and $s\in W^{uc}$, $\rho^{uc}(p,s)=\rho^c(p,(\Delta_s,\sigma_s))$.*

Then, for any $(\Gamma,\pi)\in W^c$*, let* $\mathscr{U}^c_{(\Gamma,\pi)}$ *be the submodel of* $\mathscr{U}^c$ *generated from* $(\varepsilon,(\Gamma,\pi))$*, where* ε *is the empty string.*

By the definition of the canonical unraveling and Lemma A.5,A.6, we can easily check the following:

Lemma A.8 *For any* $(\Gamma,\pi)\in W^c$*, the following hold:*

- $\mathscr{U}^c_{(\Gamma,\pi)}$ *is a local epistemic model;*
- *For all* s,t *in* $\mathscr{U}^c_{(\Gamma,\pi)}$*, if* $t\in\bigcap_{a\in A}R^{uc}_a(s)$*, then* $(\Delta_s,\sigma_s)R^c_A(\Delta_t,\sigma_t)$*;*
- *For all* s *in* $\mathscr{U}^c_{(\Gamma,\pi)}$*, if* $\widehat{\mathsf{K}}_A\alpha\in\Delta_s$*, then there is* $t\in\bigcap_{a\in A}R^{uc}_a(s)$ *s.t.* $\alpha\in\Delta_t$*.*

Then, we can prove the following *Truth Lemma*:

Lemma A.9 (Truth) *For any* $(\Gamma,\pi)\in W^c$*, any* s *in* $\mathscr{U}^c_{(\Gamma,\pi)}$ *and any* $\phi\in\mathscr{L}_{\mathbf{A}}^{[:=]}$ *s.t.* $FV(\phi)\subseteq FV(\Delta_s)$*,*

$$\mathscr{U}^c_{(\Gamma,\pi)},s,\sigma_s\vDash\phi\iff\phi\in\Delta_s$$

PROOF. As usual, the above Truth Lemma can be proved by induction on the structure of $\mathscr{L}_{\mathbf{A}}^{[:=]}$-formulas. We only consider the case for p_x and $[x:=a]\phi$ here, since the other cases are rather routine (Only note that when proving the case for $\mathsf{K}_X\alpha$, we need to make use of the fact that α is a sentence: the truth value of a sentence is irrelevant to the assignment, so we can change the assignment to apply IH.)

For p_x: Let $\sigma_s(x)=a$. First, assume that $p_x\in\Delta$. Then, since (Δ_s,σ_s) is coherent, by Lemma A.2, $\langle x:=a\rangle p_x\in\Delta_s$. Then, by $\mathtt{DET}^{[:=]}$, $[x:=a]p_x\in\Delta_s$; and we also have $\langle x:=a\rangle\top\in\Delta_s$. Note that the latter implies that $a\in FV(\Delta_s)$ by coherence. Then, by Lemma A.2, $p_a\in\Delta_s$, so $a\in\rho^c(\Delta_s,\sigma_s)=\rho^{uc}(s)$.

Conversely, if $p_x\notin\Delta$, then $\neg p_x\in\Delta$. Then, by the same reasoning as above, we have $\neg p_a\in\Delta_s$, so $p_a\notin\Delta$, and thus $a\notin\rho^{uc}(s)$.

For $[x:=a]\phi$: By $\mathtt{R}^{[:=]}$, we may assume without loss of generality that $\phi[a/x]$ is admissible.

First, assume that $[x:=a]\phi\in\Delta_s$. Also assume that $a\in\delta^{uc}(s)=\mathbf{A}\cap FV(\Delta_s)$. Then, by Lemma A.2, we have $\phi[a/x]\in\Delta_s$. Hence, by IH, $\mathscr{U}^c_{(\Gamma,\pi)},s,\sigma_s\vDash\phi[a/x]$, and thus $\mathscr{U}^c_{(\Gamma,\pi)},s,\sigma_s[x\mapsto a]\vDash\phi$.

Conversely, assume that $[x:=a]\phi\notin\Delta_s$. Then, $\langle x:=a\rangle\neg\phi\in\Delta_s$, so $[x:=a]\neg\phi\in\Delta_s$ by $\mathtt{DET}^{[:=]}$, and also $\langle x:=a\rangle\top\in\Delta_s$. By coherence, the latter implies that $a\in FV(\Delta_s)$, so by Lemma A.2, $\neg\phi[a/x]\in\Delta$, and thus $\phi[a/x]\notin\Delta$. Hence, by IH, $\mathscr{U}^c_{(\Gamma,\pi)},s,\sigma_s\nvDash\phi[a/x]$, and thus $\mathscr{U}^c_{(\Gamma,\pi)},s,\sigma_s[x\mapsto a]\nvDash\phi$. □

Now, with the help of the above lemmas, we can easily prove the Completeness Theorem.

PROOF OF THEOREM 4.2. Let Γ_0 be an arbitrary consistent set of $\mathscr{L}^{[:=]}$ formulas. By Lemma A.3, there is $(\Gamma,\pi)\in W^c$ s.t. $\Gamma_0\subseteq\Gamma$. Then, by Lemma A.9, we have $\mathscr{U}^c_{(\Gamma,\pi)},(\varepsilon,(\Gamma,\pi)),\pi\vDash\Gamma_0$, so Γ_0 is satisfiable. Thus, $\mathbf{LEL}^{[:=]}$ is strongly complete w.r.t. the class of all local epistemic model. Then, by Proposition 3.3, $\mathbf{LEL}^{[:=]}$ is also strongly complete w.r.t. the class of all *proper* local epistemic model. □

A.4 Proof of Proposition 4.4

In order to prove Proposition 4.4, first notice that we have the following:

Lemma A.10 *The following are* $\mathbf{LEL}^{[:=]}$*-theorems:*

$$\begin{array}{ll}
\mathtt{ELM}_a^{\mathtt{TR}} & [x{:=}a]\chi \leftrightarrow ([x{:=}a]\bot \vee \chi) \quad (x \notin FV(\chi)) \\
\mathtt{ELM}_a^{\neg} & [x{:=}a]\neg\phi \leftrightarrow ([x{:=}a]\bot \vee \neg[x{:=}a]\phi) \\
\mathtt{ELM}_a^{\wedge} & [x{:=}a](\phi \wedge \psi) \leftrightarrow ([x{:=}a]\phi \wedge [x{:=}a]\psi) \\
\mathtt{ELM}_a^{[:=]} & [x{:=}a][y{:=}b]\phi \leftrightarrow [y{:=}b][x{:=}a]\phi \quad (y \neq x)
\end{array}$$

Then, with the help of the $\mathbf{LEL}^{[:=]}$-theorems listed above, we can prove Proposition 4.4.

PROOF OF PROPOSITION 4.4. For convenience, for an assignment operator $[x{:=}a]$ and a formula ϕ, we say ϕ is in $[x{:=}a]$-*normal form*, if for any subformula of ϕ of the form $[x{:=}a]\theta$, θ is one of p_x, $\bot$ and $[\vec{y}{:=}\vec{b}]\mathsf{K}_Z\alpha$, where $p \in \mathbf{P}$, $\vec{y} \in Z$ and $x \in Z \setminus \{\vec{y}\}$. Then, given $[x{:=}a]$, consider the following translation:

When the formula is not $[x{:=}a]\chi$:

$$\begin{array}{lcl}
\mathtt{NF}_x^a(p_y) & = & p_y \\
\mathtt{NF}_x^a(\neg\phi) & = & \neg\mathtt{NF}_x^a(\phi) \\
\mathtt{NF}_x^a(\phi \wedge \psi) & = & \mathtt{NF}_x^a(\phi) \wedge \mathtt{NF}_x^a(\psi) \\
\mathtt{NF}_x^a([y{:=}b]\phi) & = & [y{:=}b]\mathtt{NF}_x^a(\phi) \\
\mathtt{NF}_x^a(\mathsf{K}_X\alpha) & = & \mathsf{K}_X\mathtt{NF}_x^a(\alpha) \\
\mathtt{NF}_x^a(\top) & = & \top
\end{array}$$

When the formula is $[x{:=}a]\chi$ and $x \in FV(\chi)$:

$$\begin{array}{lcl}
\mathtt{NF}_x^a([x{:=}a]p_x) & = & [x{:=}a]p_x \\
\mathtt{NF}_x^a([x{:=}a]\neg\phi) & = & [x{:=}a]\bot \vee \neg\mathtt{NF}_x^a([x{:=}a]\phi) \\
\mathtt{NF}_x^a([x{:=}a](\phi \wedge \psi)) & = & \mathtt{NF}_x^a([x{:=}a]\phi) \wedge \mathtt{NF}_x^a([x{:=}a]\psi) \\
\mathtt{NF}_x^a([x{:=}a][y{:=}b]\phi) & = & [y{:=}b]\mathtt{NF}_x^a([x{:=}a]\phi) \\
\mathtt{NF}_x^a([x{:=}a]\mathsf{K}_X\alpha) & = & [x{:=}a]\mathsf{K}_X\mathtt{NF}_x^a(\alpha)
\end{array}$$

When the formula is $[x{:=}a]\chi$ and $x \notin FV(\chi)$:

$$\mathtt{NF}_x^a([x{:=}a]\chi) = [x{:=}a]\bot \vee \mathtt{NF}_x^a(\chi)$$

Then, using Lemma A.10, it is rather routine to check by induction that for any $\phi \in \mathscr{L}^{[:=]}$, $\mathtt{NF}_x^a(\phi)$ is in $[x{:=}a]$-normal form and $\vdash \phi \leftrightarrow \mathtt{NF}_x^a(\phi)$; moreover, if ϕ is already in $[y{:=}b]$-normal form, then $\mathtt{NF}_x^a(\phi)$ is still in $[y{:=}b]$-normal form.

Then, given $\alpha \in \mathscr{L}_0^{[:=]}$, we first list all the assignment operators occurring in α as $[x_0{:=}a_0], [x_1{:=}a_1]$, ..., $[x_{n-1}{:=}a_{n-1}]$; then, let $\alpha_0 = \alpha$, and given α_k, let $\alpha_{k+1} = \mathtt{NF}_{x_k}^{a_k}(\alpha_k)$. In this way, the formula α_n we obtained is an assignment normal form of α. □

AI-Generated Compromises for Coalition Formation: Modeling, Simulation, and a Textual Case Study

Eyal Briman
Ben-Gurion University of the Negev
Beer Sheva, Israel
briman@post.bgu.ac.il

Ehud Shapiro
Weizmann Institute of Science
Rehovot, Israel
udi.shapiro@gmail.com

Nimrod Talmon
Ben-Gurion University of the Negev
Beer Sheva, Israel
nimrodtalmon77@gmail.com

The challenge of finding compromises between agent proposals is fundamental to AI sub-fields such as *argumentation* [27], *mediation* [22], and *negotiation* [20]. Building on this tradition, Elkind et al. [9] introduced a process for coalition formation that seeks majority-supported proposals preferable to the status quo, using a metric space where each agent has an ideal point. The crucial step in this iterative process involves identifying *compromise proposals* around which agent coalitions can unite. How to effectively find such compromise proposals, however, remains an open question. We address this gap by formalizing a holistic model that encompasses agent bounded rationality and uncertainty and developing AI models to generate such compromise proposals.

We focus on the domain of collaboratively writing text documents – e.g., to enable the democratic creation of a community constitution. We apply NLP (Natural Language Processing [7]) techniques and utilize LLMs (Large Language Models [35]) to create a semantic metric space for text and develop algorithms to suggest suitable compromise points. To evaluate the effectiveness of our algorithms, we simulate various coalition formation processes and demonstrate the potential of AI to facilitate large-scale democratic text editing, such as collaboratively drafting a constitution—an area where traditional tools are limited.

1 Introduction

We propose a framework for iterative *compromise-based coalition formation* that enables a set of agents to collaboratively develop a single text document. Each agent starts with an *ideal document*, modeled as a point in a (potentially high-dimensional) metric space. At each step, certain agents may collectively switch to a newly proposed *compromise document*—also represented as a point in the metric space.

Our work generalizes the model of Elkind et al. [9], who examine an iterative coalition formation process wherein agents only move to a new coalition (i) if its compromise document is closer to their ideal points than the status quo, and (ii) if the new coalition is at least as large as their current one. These two conditions ensure certain theoretical guarantees (e.g., on the convergence of the process; see footnote 5 for a formal description of the generalization). However, in many realistic scenarios, agents may not behave strictly according to these criteria. E.g., an agent might rationally move to a smaller coalition if it yields a document that more strongly aligns with its preferences, or it might stochastically deviate from strict rationality due to partial information, uncertainty, or other behavioral considerations.

Generalizing Agent Behavior. In contrast to Elkind et al. [9], we allow for more flexible coalition formation. Agents may:

- Move to a new coalition even if that coalition is smaller than their current one,
- Take actions probabilistically, representing bounded rationality or incomplete knowledge,

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 417–432, doi:10.4204/EPTCS.437.32

© E. Briman, E. Shapiro & N. Talmon
This work is licensed under the
Creative Commons Attribution License.

These relaxed conditions capture a broader range of real-world behaviors. Thus, our modeling goal is to develop a framework robust enough to accommodate both purely rational and partially rational agents, while still facilitating majority-supported text generation.

Collaborative Text Editing and the Mediator Concept. Although Elkind et al. [9] discuss the theoretical dynamics of forming coalitions via compromise proposals, they do not specify *how* such proposals are generated. We address this gap by introducing the notion of an *mediator* that systematically produces compromise documents. Specifically, we embed potential texts in a semantic metric space and employ modern natural language processing (NLP) techniques to measure distances between documents, thereby identifying compromise points that better align with multiple agents' preferences. This approach is particularly relevant for large-scale, democratic text editing tasks—such as drafting a constitution for a decentralized autonomous organization (DAO) [13]—which existing collaborative platforms (e.g., Google Docs, Notion, Wikipedia) do not address in a structured, consensus-driven manner. By modeling documents as points in a high-dimensional embedding space, the mediator can propose new drafts that balance diverse viewpoints, thus paving the way for a more democratic editing process.

Main Contributions. Our contributions can be summarized as follows:

1. **Generalizing the Coalition-Formation Model.** We extend the work of Elkind et al. [9] by permitting less restrictive movement rules, thus supporting bounded rationality and agents who may move to smaller coalitions.
2. **AI-Mediated Proposal Generation.** We introduce algorithms that employ large language models (LLMs) and other NLP tools to embed and manipulate text documents in a semantic metric space, enabling the discovery of meaningful compromise drafts.
3. **Empirical Evaluation in Euclidean and Textual Domains.** We present simulations in both a simplified 2D Euclidean space and a more realistic text-editing environment. Our findings show that—even under relaxed decision rules—agents converge to a majority-supported document that improves upon the status quo.

For space considerations, some text is deferred to the appendix: a more detailed exposition of certain related work; application of the model to the Euclidean setting; more examples; and some details regarding the simulation results.

1.1 Model State

The model is defined by the following fixed components[1]:

- A metric space X.
- A distance function $\mathbf{d} : X \times X \to \mathbb{R}_{\geq 0}$ defining the metric on X.
- A point $r \in X$, representing the fixed status quo.
- A set $V = \{v_1, \ldots, v_n\}$ of n agents. Each agent $v \in V$ is associated with an ideal point $p^v \in X$ and has *Euclidean preferences* [4], meaning that preferences are determined by distance from the ideal point.

[1]This is a centralized description for ease of presentation. Conceptually, we envision a decentralized setting, where the mediator operates as a non-centralized tool available to individual coalitions. That is, coalitions may grow in a bottom-up manner, each using an instance of the mediator independently.

The *state* of the process is given by a coalition structure:

- A set $D = \{d_1, \ldots, d_z\}$, where each $d_i = (C_i, p_i) \in D$ for $i \in [z] := \{1, \ldots, z\}$, is a *coalition*. Here, $C_i \subseteq V$ denotes the set of agents in the coalition and $p_i \in X$ the compromise point around which the coalition is formed. The coalition structure D is a partition of the agents: for all $i \neq j \in [z]$, we have $C_i \cap C_j = \emptyset$, and $\bigcup_{i \in [z]} C_i = V$.

  We assume $z \in [n] := \{1, \ldots, n\}$; that is, the number of coalitions does not exceed the number of agents. The notation $[z]$ refers to the index set of the current coalition structure.

1.2 Initialization, Iterative Process, and Halting Conditions

Next, we describe a specific modeling and configuration. This approach allows us to present the capabilities of the model in a clear, specific, and traceable manner, making it easier to understand. By focusing on a concrete example, we aim to illustrate the potential applications and advantages of the model, while leaving room to discuss its broader possibilities in the outlook section.

Initialization Initially, the process starts with each agent forming its own singleton coalition: namely, $D = \{d_1, \ldots, d_n\} = \{(C_1, p_1), \ldots, (C_n, p_n)\}$ with $C_i = \{v_i\}$ and $p_i \in X$.

Process The model contains an entity—the **mediator**—which is the workhorse of the process.

Definition 1 (mediator)**.** *A* mediator *M is a function that gets as input a coalition structure D and returns as output a tuple (d_i, d_j, p) with $d_i, d_j \in D$ and p a point in the metric space.*

Intuitively, the mediator suggests that two coalitions, d_i and d_j, merge around a compromise point p. Given the current coalition structure D, the mediator returns a triple (d_i, d_j, p), where p is proposed as a new joint position.

Each coalition responds to this suggestion according to a predefined *constitution*, which governs how agents decide whether to join the new coalition. Specifically, agents in d_i and d_j vote on whether they prefer the proposed compromise p over remaining in their current coalition. Based on these votes and the constitution, some agents may transition to the new coalition while others remain.

We first define the voting behavior of an agent before specifying the constitutions that aggregate these votes.

Definition 2 (Agent, vote)**.** *An* agent *v corresponds to some ideal point p^v; and, furthermore, a* vote *of agent v regarding some point p is $vote(v, p) \in \{0, 1\}$ (where $vote(v, p) = 1$ means that v accepts the suggestion to move to a coalition to be formed around p).*

Now, a constitution *const* is defined as follows.

Definition 3 (Constitution)**.** *A* constitution *const gets as input a tuple (d_i, d_j, p) by the mediator and, when applied on d_i – and based on the votes of the agents in d_i, as described by $\{vote(v, p) : v \in d_i\}$ – returns an assignment to a coalition for each $v \in d_i$, namely $const(v) \in \{d_i, d^p\}$, where d^p describes the coalition to be possibly-formed around the suggested compromise point p.*

Following a suggestion of (d_i, d_j, p) and an application of the constitution *const* on d_i and d_j (which internally depends on the votes of the agents in both coalitions), the resulting Markov state contains a new coalition structure D' that is defined as follows:[2] $D' := D \setminus \{d_i, d_j\} \cup \{d'_i, d'_j, d^p\}$, where $d'_i := (\{v \in d_i : const(v) = d_i\}, p^i)$; $d'_j := (\{v \in d_j : const(v) = d_j\}, p^j)$; $d^p := (\{v \in d_i \cup d_j : const(v) = d^p\}, p)$.

[2] A coalition with no members can safely be removed from a coalition structure (such that d'_i, d'_j, and d^p may be empty).

That is, agents from d_i whom the constitution assigns to d_i remain in it; agents from d_j whom the constitution assigns to d_j remains in it; and agents from $d_i \cup d_j$ whom the constitution assigns to the new coalition around p are being moved there.

A halting condition The process halts whenever a coalition that contains an agent majority is being formed; i.e., whenever some $d \in D$, $d = (C, p)$, exists with $|C|/|V| \geq \mathcal{Q}$, where the fraction $\mathcal{Q} \in [0,1]$ can be set to be majority, super majority, or consensus (in our simulation we implement a simple majority).

2 Concrete Model Realizations

We provide concrete realizations of the following ingredients: agent models (in Section 2.1), coalition constitutions (in Section 2.2), and mediators (in Section 2.3). These concrete realizations are used later, for the 2D Euclidean setting presented in the appendix and the setting that involves text documents. Furthermore, some of the details next are needed for the computer-based simulations that follow. [3]

2.1 One Concrete Agent Model

As abstractly stated above, an agent v corresponds to an ideal point p^v and shall have the ability to vote on a proposal p by returning a binary answer in the form of $vote(v,p) \in \{0,1\}$ – if $vote(v,p) = 1$, then we say that *v approves p*. Naturally, various realizations of agent models are possible. Below we describe the agent model we use in our theoretical realization (later, in Section 4 we use a different, LLM-based agent model). Let us first define a simple, deterministic agent model.

Definition 4 (A deterministic agent model)**.** *Under the* deterministic agent model*, an agent v within previous coalition d_i with ideal point p^v votes as follows: $vote(v,p) = 1$ if $\mathbf{d}(p^v,p) < \mathbf{d}(p^v,r)$.*

That is, an agent approves a proposal p if p is closer to its ideal point than the status quo r, and it disapproves of a proposal p otherwise. Next, as we are interested in modeling agent altruism and flexibility in the process in a naive and intuitive manner (influenced by [17]) we use a probabilistic generalization of the simple model, as described next.

In particular, given the status quo r, a proposal p, and an agent v with ideal point p^v, we define a function $F(r,p,p^v)$ that returns the probability of the agent approving p. Specifically, $F(r,p,p^v) \in [0,1]$. (It may be helpful to note that the deterministic agent model corresponds to the probabilistic model if $F(r,p,p^v) = 1$ whenever $\mathbf{d}(p^v,p) < \mathbf{d}(p^v,r)$.)

Specifically, to model different types of non-deterministic agents, we introduce a parameter $\sigma \geq 0$, where larger values of σ results in a more altruistic agent behavior as compared to the simplest agent model described above. Mathematically, we use a half (positive) Gaussian distribution to "enlarge" a bit the region for which the agent approves the proposal (i.e., so that an agent will approve a proposal even if it is farther away from its ideal point, compared to the status quo; but with diminishing probability of doing so); formally, we have the following definition of F (note that the *else* case is 0 in case of $\sigma = 0$):

$$F(r,p,p^v) = \begin{cases} 1, & \text{if } \mathbf{d}(p^v,r) \geq_v \mathbf{d}(p^v,p) \\ \frac{2}{\sigma_v\sqrt{2\pi}} e^{-\frac{(\mathbf{d}(p^v,p))^2}{2\sigma_v^2}} & \text{else} \end{cases}$$

[3] We consider realizations of the model in which all agents share the same agent model; all coalitions share the same constitution; and there is one mediator throughout the process. We discuss other options in Section 5.

Definition 5 (A probabilistic agent model)**.** *Under the* probabilistic agent model*, an agent v with ideal point p^v votes as follows: $vote(v,p) = 1$ with probability $F(r,p,p^v)$.*[4]

Remark 6. *The current agent model assumes that voting behavior depends only on the distance between the proposed point p, the status quo r, and the agent's ideal point p^v. A natural extension is to allow votes to depend on the anticipated composition of the new coalition. For instance, agents may approve p only if sufficiently many others are expected to join.*

2.2 Two Concrete Constitutions

As abstractly stated above, given a proposal for a coalition $d_i = (C_i, p_i)$ to move to a new coalition around a compromise point p, a constitution takes the votes of the agents and determines whether any of the coalition members shall move to the new coalition, and, if so, who. We explore two options for such constitutions.

- **Coalition Discipline:** A new coalition is formed only if at least $Q \in [0, |C_i|]$ members of C_i approve the proposal. Formally:

$$\text{if } |\{v \in C_i : \text{vote}(v,p) = 1\}| \geq Q, \text{ then for each } v \in C_i : \quad \text{const}(v) := \begin{cases} d_p & \text{if vote}(v,p) = 1 \\ d_i & \text{otherwise} \end{cases}$$

  otherwise, $\text{const}(v) := d_i$ for all $v \in C_i$.[5]

- **No Coalition Discipline** is a special case of the above with $Q = 0$, where each agent independently decides whether to join the new coalition:

$$\text{const}(v) := \begin{cases} d_p & \text{if vote}(v,p) = 1 \\ d_i & \text{otherwise} \end{cases} \quad \text{for each } v \in C_i.$$

Remark 7. *We assume the coalition size $|C_i|$ is known to its members at the time of voting.*

Remark 8. *Agent preferences depend only on distance to their ideal point. Coalition discipline imposes coordination constraints but does not affect individual utility.*

2.3 Several Concrete Mediators

Recall that an mediator takes as input a coalition structure D and returns two coalitions, d_i and d_j, and a compromise point p. It is convenient to break the description of our realizations into the two main tasks of mediators, namely: (1) choosing the coalitions d_i and d_j to suggest p to; and (2) choosing the compromise point p to suggest to d_i and d_j. Note that the role of the AI in our design is in the implementation of such mediators.

[4] Indeed, for $\sigma = 0$, the probabilistic agent model and the deterministic agent mode coincide.

[5] Our model builds upon and generalizes aspects of the framework proposed by Elkind et al. In the case of coalition discipline with deterministic agents and unanimous approval ($Q = |C_i|$), we recover their convergence results under the constraint that agents only transition to strictly preferred larger coalitions.

Choosing the coalitions d_i, d_j Our mediators proceed by first computing the centroid of the coalitions' ideal points, weighted by the coalition sizes. Formally: $centroid(D) = \arg\min_{x \in X} \frac{1}{n} \cdot \sum_{i \in |D|} |C_i| \cdot \mathbf{d}(x, p_i)$. Using the centroid, the mediators consider the distance of each coalition from the centroid, denoted by $\mathbf{d}(p_i, centroid)$. The selection process is guided by a parameter $\alpha \in [-1, 1]$, intuitively ranging from whether the closest coalitions to the centroid are preferred ($\alpha = -1$), the furthest ones are preferred ($\alpha = 1$), or there is no significance ($\alpha = 0$) to the distance from the centroid.

Each coalition i is assigned a score S_i based on its distance from the centroid using the following scoring function $S_i = e^{\alpha \cdot d'(p_i, centroid)}$, where $d'(p_i, centroid) \in [0, 1]$ is the normalized distance; formally:

$$d'(p_i, centroid) = \frac{\mathbf{d}(p_i, centroid)}{\arg\max_{j \in |D|} \mathbf{d}(p_j, centroid)} .$$

Subsequently, the mediator assigns a probability $prob(d_i)$ to each coalition d_i, proportionate to its scores: $prob(d_i) = \frac{S_i}{\sum_{i=1}^{|D|} S_i}$. In practice, the mediator probabilistically chooses one coalition based on these probabilities and then selects the closest coalition to the initially chosen one.

Choosing the Compromise Point p Given coalitions $d_i = (C_i, p_i)$ and $d_j = (C_j, p_j)$, the mediator selects a compromise point $p \in X$ that minimizes the weighted sum of distances to p_i and p_j, with weights proportional to coalition sizes:

$$p = \arg\min_{x \in X} \left(\frac{|C_i|}{|C_i| + |C_j|} \cdot \mathbf{d}(p_i, x) + \frac{|C_j|}{|C_i| + |C_j|} \cdot \mathbf{d}(p_j, x) \right).$$

In Euclidean space, this reduces to the standard weighted average.

Remark 9. *The mediator is assumed to know all agents' ideal points. This may result from voluntary disclosure or be treated as a modeling assumption. While this enables computation of globally optimal coalitions under a defined objective, the current mediator applies local, myopic merges. Designing optimal, forward-looking mediators is left for future work.*

3 Related Work

Coalition formation in a metric space has been studied from a multiagent system context [5, 36, 30]. We build upon the theoretical framework of Elkind et al. [9], which introduced a model for deliberative coalition formation in metric spaces. Their work presents a transition system to capture the dynamics of the coalition formation process. While Elkind et al. describe the formation process in detail, they assume that compromise points are provided by an external source (an oracle), without specifying how these points should be determined. Our contribution addresses this gap by introducing *mediators* that algorithmically suggest compromise points, allowing coalitions to unite around majority-supported proposals. We implement and optimize these mediators to make the coalition formation process both practical and efficient. We also demonstrate that, under a specific configuration of our model, it aligns with Elkind's model, thereby inheriting their theoretical results for that configuration thus our model generalizes Elkind's model. For the general case we show simulations that show convergence rates are very good even for large instances. In developing mediators, we utilize NLP techniques and LLMs; this relates to NLP-based recommendation systems [2, 32], where models suggest content based on user preferences, and to recent work in *Generative Social Choice* [12], which explores the use of LLMs to generate representative statements for social choice tasks. However, our work differs by focusing on

identifying compromise points in the coalition formation process, where the goal is to find a majority-supported text or proposal. Another relevant line of work is by Bakker et al. [1], who study how machines can assist in finding agreements among individuals with diverse preferences. Their approach fine-tunes LLMs to generate statements that maximize the expected approval of a group, which is conceptually similar to our use of AI for proposing compromise points. However, our model incorporates an iterative coalition formation process, making it distinct in its operational dynamics. We also mention Yang et al. [34] that investigate how GPT-4 and LLaMA-2 behave in voting scenarios compared to human voters. They show that voting methods, presentation order, and temperature settings can significantly influence LLM choices, often reducing preference diversity and risking bias.

In the context of *Dynamic Coalition Formation*, there is significant prior work on how agents with diverse preferences form and adapt coalitions to achieve consensus [31, 19]. This is relevant, as coalition formation plays a key role in decision-making processes, especially when agents aim to form majority-supported agreements [26]. Our approach to coalition formation in metric spaces also draws on existing research in spatial voting models [10, 15]. We are also motivated by psychological research on the ability of agents to objectively evaluate proposals. Mikhaylovskaya et al. [23] provide evidence that AI-based mediators can mitigate human biases, making AI a promising tool for generating neutral, data-driven compromise points. This motivates our use of AI-mediated coalition formation, where agents can evaluate AI-suggested compromise points to find collective agreements. Our work also draws inspiration from negotiation-based approaches to coalition formation [25, 14, 28, 9, 16, 33, 3, 11], which offer valuable insights into how agents with divergent preferences negotiate and form coalitions. These approaches further reinforce the relevance of AI based mediation in improving the efficiency and effectiveness of coalition formation in multi-agent systems.

4 Mediators in a Textual Space

As our ultimate goal relates to text aggregation – i.e., to enable an agent community to converge towards a majority-supported textual document. So, we wish to utilize the mediators framework (demonstrated also in a Euclidean space in the appendix) to a setting in which the metric space contains textual documents and coalitions form around different texts, until a majority-supported textual document is identified. We describe our specific model; and then report on computer-based simulations.

4.1 Modeling Mediators in a Textual Space

Our general solution works as follows. We use word embedding (a standard NLP technique) to translate texts into numerical-valued vectors; this is crucial as, after applying such a word embedding, we are then able to compute distances between the embedded coalition ideal points and use the mediators of the Euclidean space. In this work, we use Google's Universal Sentence Encoder [6]: this is a pre-trained model that converts sentences into fixed-size vectors, capturing their semantic meanings. (The Universal Sentence Encoder is designed to generate 512-dimensional embedding vectors, providing a semantic representation of sentences.) Thus our embedded metric space contains as elements all those 512-length vectors that can be the output of the Universal Sentence Encoder. As a distance measure in this space, we use the square root of the cosine-based dissimilarity [29], a commonly used pseudo metric in NLP.[6]

[6]Formally, we define

$$d_{\cos}(A,B) := \sqrt{2 - 2 \cdot \frac{A \cdot B}{\|A\| \cdot \|B\|}} \in [0,2],$$

The mediator guides the coalition formation by proposing sentences to two coalitions within the given word limit (in our simulations, 15 words). In each iteration, the mediator's goal is to find two coalitions to suggest a sentence that minimizes the *squared cosine similarity* pseudo metric between the embedding vector of the chosen sentence and the weighted average of the embedding vectors of the two coalition points.

Our approach to coalition formation in the domain of text relies on the integration of OpenAI's GPT-3.5-turbo-1106 model (`https://openai.com/blog/chatgpt`) with a temperature parameter (responsible for the randomness of results) of 0.75 as recommended in some of the documentation to provide a good trade-off for applications like ours where the output should be coherent but still allow for some diversity and creativity. This LLM takes 3 key roles within our framework:

1. It generates sentences that act as agent ideal points.
2. It constructs initial singleton coalitions mirroring these ideal sentences if the coalition formation process introduces noise (i.e., for simulations runs with $I = true$).
3. It proposes diverse options for aggregating two sentences, presenting methods to combine opinions from different coalitions represented as text. The process then determines the most suitable sentence by evaluating which has the embedding that is the closest to the weighted average of the two coalition embedded sentences.

Remark 10. *We assume that Euclidean distance in the embedding space reflects agent preferences. That is, texts closer to an agent's embedded ideal point are considered more preferable. This assumption connects the embedding to the distance-based agent model, though it may not hold uniformly across domains.*

4.2 Simulations-Based Analysis

We conducted simulations to assess the robustness and resilience of the model; done as follows:

- **Parameter Tests and Scale:** The simulations were conducted with different numbers of agents, specifically $n \in \{10, 20, 30, 40, 50, 100, 1000\}$. We varied the parameters $\sigma \in \{0, 1, 1.5\}$ and $\alpha \in \{-1, 0, 1\}$, while also setting the boolean variable C— enforcing coalition discipline or not. Each parameter combination was tested across 50 repetitions, with all ideal sentences generated sharing a predetermined topic of ways to address global warming.
- **Coalition Formation Process:** We employed an iterative pursuit in an embedding space using *squared cosine similarity* pseudo metric; the prompt given to GPT was: "Give me T different sentences that are well structured about how to deal with Y with at most of 15 words" (T being the number of agents, and Y being any topic – global warming in our case); to initialize the singleton coalition with introducing noise ($I = True$), the LLM was requested to provide a sentence resembling the ideal sentence of each agent, rather than introducing additional noise through a normal distribution as conducted in the euclidean case presented in the appendix. These function as the singleton coalition sentences to be embedded into the Euclidean space. The prompt given to the GPT was: "Give me a well-structured sentence with a maximum of 15 words, resembling this sentence: Z" (where Z represents an ideal sentence of an agent).

which corresponds to $\sqrt{2} \cdot \sin(\theta/2)$, where θ is the angle between vectors A and B. This is a metric when restricted to the unit sphere (or more generally, to rays through the origin), but only a pseudo metric over the entire embedding space: co-linear vectors have distance zero, even if they differ in magnitude. Since our mediator constructs compromise vectors via (non-normalized) weighted averages, we do not assume that vectors lie on a sphere.

- **Sentence Selection Process:** For each proposed sentence to the two coalitions, the LLM was tasked with generating 10 sentences that effectively combined both coalition sentences. We followed best practices for structured prompt design and multi-step reasoning [18], including these concepts:
  - *Structured Prompt Design:* Prompts should provide clear and concise instructions, ensuring that the LLM produced well-structured sentences.
  - *Encouraging Multi-Step Reasoning:* Prompts should be designed to guide the LLM through step-by-step reasoning, leveraging Zero-shot Chain of Thought (CoT) techniques to handle the task effectively.
  - *No Task-Specific Examples Needed:* Prompts should avoid the need for specific examples or task-specific training, enabling the model to generalize across different tasks.

  We used the following prompts and messages given to GPT 3.5 (5 options in total):
  - Mediator 1: **Prompt:** Generate 10 possible different well-structured sentences that aggregate the following two sentences. Make sure each sentence has at most 15 words. Number your answers (i.e., 1), 2), 3), 4), 5), and so on) for each sentence you propose. **Message:** You are a mediator trying to find agreed wording for how to deal with global warming based on existing sentences. Give a straightforward answer with no introduction to help people reach an agreed wording of a coherent sentence. (The proposed sentence was selected based on the minimal *squared cosine similarity* pseudo metric between its embedding and the weighted average embedding vector, considering the two embedding vectors of the coalitions and their sizes.)
  - Mediator 2: **Prompt:** Generate 10 concise and clear sentences that blend the following two sentences into one coherent idea: Ensure each sentence is no longer than 15 words. Number your answers (i.e., 1), 2), 3), 4), 5), and so on) for each sentence you propose. **Message:** As a mediator, you need to find a consensus on global warming solutions. Provide straightforward and numbered suggestions to help reach a clear and agreed-upon sentence.
  - Mediator 3: **Prompt:** Create 10 unique, well-structured sentences that combine these two sentences into one unified thought: Each sentence should be a maximum of 15 words. Number your answers (i.e., 1), 2), 3), 4), 5), and so on) for each sentence you propose. **Message:** You are acting as a mediator to achieve a common statement on global warming. Give direct and numbered suggestions to assist in forming a unified and coherent sentence.
  - Mediator 4: This baseline mediator involved soliciting several possibilities for sentence aggregation from the GPT and then selecting the sentence that minimized the distance from the average embedding vector of the two coalitions. Instead, we simply requested GPT to provide a single sentence. The prompt and message given to GPT were the same as those given for Option 1, but instead of 10 sentences, it was asked for 1 sentence only.
  - Mediator 5: This second baseline mediator denoted by **Option 5**, was to ask GPT for a completely random sentence.

We tested the number of iterations needed for coalitions to converge on a compromise, and the average distance between the compromise document and the ideal document of each agent within the coalition that halted the process.

5 Outlook and Discussion

Our findings closely align with the Euclidean case presented in the appendix:

Option	Mean Number of Iterations
Option 1	4.8000
Option 2	5.0750
Option 3	5.5250
Option 4	7.5000
Option 5	41.8125

Table 1: A comparison of different mediators-each corresponding to different prompts and LLM-usage strategy. The mean number of iterations until convergence is shown, validated with 95% confidence using ANOVA and Tukey HSD.

1) Processes with coalition discipline and deterministic agents always exhibit some cases of non-convergence (defined as exceeding 10,000 iterations), whereas all other combinations result in convergence; 2) A higher number of agents leads to more iterations; 3) Increasing α enlarges the mean average distance between the compromise sentence of the largest coalition and the ideal sentences of its members; 4) Coalition discipline reduces this mean average distance.

We also analyze the performance of different mediators, summarized in Table 1. An ANOVA test confirms statistically significant differences in the number of iterations across mediation approaches. A post-hoc Tukey HSD test (see appendix) further identifies significant pairwise differences, revealing that Option 1 achieves the fewest iterations on average.

Remark 11. *We conducted additional experiments using GPT-3 Davinci and GPT-4o Mini, both with a temperature of 0.75. As their results followed the same patterns and led to identical conclusions, we omit them here for brevity.*

5.1 Interpretation

Our simulations demonstrate the effectiveness of *AI-mediated coalition formation*, particularly when leveraging *Large Language Models (LLMs)*. AI-mediation significantly reduces the number of iterations required for coalitions to reach a compromise while minimizing the average distance between the final compromise and individual agents' ideal documents. Notably, LLMs combined with distance-based optimization consistently accelerate convergence compared to simpler mediator approaches.

The statistical tests reinforce that meaningful differences exist among mediation strategies, underscoring the adaptability of AI-mediation to different scenarios. The superior performance of Option 2 in minimizing iterations further suggests that careful tuning of the mediator's behavior can yield substantial efficiency gains.

5.2 Future Work

Several directions for future research are outlined below:

- **Theoretical Guarantees:** Analyze convergence under relaxed rationality and probabilistic behavior; study stability and fairness properties.
- **Strategic Behavior:** Extend the model to a game-theoretic setting that accounts for strategic agents who may misreport preferences and anticipate coalition dynamics.

- **Scalability:** Develop efficient mediator selection, distributed implementations, and hierarchical coalition structures for large-scale settings.
- **Bias and Interpretability:** Address AI-induced bias, enforce fairness constraints, and improve mediator transparency.
- **Application Domains:** Apply the model to other contexts, such as participatory budgeting, resource allocation, and collaborative drafting.
- **Empirical Evaluation:** Test the framework in real-world environments (e.g., DAOs, Wikipedia); assess adoption via human studies.
- **Adaptive Mediators:** Use reinforcement learning or game-theoretic tools to adapt mediator strategies over time.
- **Decentralized Use:** Support coalition-local mediator usage in decentralized systems with autonomous agent groups.
- **Proportionality:** Mitigate majority dominance using methods like Phragmén's rule [24] to ensure proportional influence in aggregation.
- **Forward-Looking Planning:** Extend mediators to evaluate merge sequences that optimize objectives (e.g., minimum distance or maximal support), under different assumptions about agent information.
- **Deliberation and Communication:** Extend the model to allow agent-to-agent communication, enabling persuasion or belief updates during the process.
- **Context-Sensitive Voting Behavior:** Our model compares proposals to a fixed status quo, following Elkind et al. [9]. A natural extension is to compare proposals to the current coalition point p_c, i.e., accept p if $d(p^v, p) < d(p^v, p_c)$. A more refined model would also consider coalition size and composition—agents may prefer large coalitions for influence, or avoid those with ideologically distant members.

Acknowledgment. The research was partially funded by the European Union, under the Horizon Europe project PERYCLES (Grant Agreement No. 101094862).

References

[1] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick et al. (2022): *Fine-tuning language models to find agreement among humans with diverse preferences*. Advances in Neural Information Processing Systems 35, pp. 38176–38189, doi:10.48550/arXiv.2211.15006.

[2] Illia Balush, Victoria Vysotska & Solomiia Albota (2021): *Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods.*

[3] Martin Beer, Mark d'Inverno, Michael Luck, Nick Jennings, Chris Preist & Michael Schroeder (1999): *Negotiation in multi-agent systems*. The Knowledge Engineering Review 14(3), pp. 285–289, doi:10.1017/S0269888999003021.

[4] Anna Bogomolnaia & Jean-François Laslier (2007): *Euclidean preferences*. Journal of Mathematical Economics 43(2), pp. 87–98, doi:10.1016/j.jmateco.2006.09.004.

[5] Laurent Bulteau, Gal Shahaf, Ehud Shapiro & Nimrod Talmon (2021): *Aggregation over metric spaces: Proposing and voting in elections, budgeting, and legislation*. Journal of Artificial Intelligence Research 70, pp. 1413–1439, doi:10.1613/jair.1.12388.

[6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar et al. (2018): *Universal sentence encoder*. arXiv preprint arXiv:1803.11175 N/A(N/A), p. N/A, doi:10.18653/v1/D18-2029.

[7] K. R. Chowdhary (2020): *Natural language processing for word sense disambiguation and information extraction*, doi:10.48550/arXiv.2004.02256.

[8] Edith Elkind, Piotr Faliszewski, Jean-François Laslier, Piotr Skowron, Arkadii Slinko & Nimrod Talmon (2017): *What do multiwinner voting rules do? An experiment over the two-dimensional euclidean domain*. In: Proceedings of the AAAI Conference on Artificial Intelligence, 31, AAAI Press, Palo Alto, CA, p. N/A, doi:10.1609/aaai.v31i1.10612.

[9] Edith Elkind, Davide Grossi, Ehud Shapiro & Nimrod Talmon (2021): *United for change: deliberative coalition formation to change the status quo*. In: Proceedings of the AAAI Conference on Artificial Intelligence, 35, AAAI Press, Palo Alto, CA, pp. 5339–5346, doi:10.1609/aaai.v35i6.16673.

[10] James M Enelow & Melvin J Hinich (1984): *The spatial theory of voting: An introduction*. Cambridge University Press, Cambridge, UK.

[11] Bernard Espinasse, Guy Picolet & Eugene Chouraqui (1997): *Negotiation support systems: A multi-criteria and multi-agent approach*. European Journal of Operational Research 103(2), pp. 389–409, doi:10.1016/S0377-2217(97)00127-6.

[12] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira & Manuel Wüthrich (2023): *Generative Social Choice*. arXiv preprint arXiv:2309.01291 N/A(N/A), p. N/A, doi:10.48550/arXiv.2309.01291.

[13] Samer Hassan & Primavera De Filippi (2021): *Decentralized autonomous organization*.

[14] Teresa W Haynes, Jason T Hedetniemi, Stephen T Hedetniemi, Alice A McRae & Raghuveer Mohan (2020): *Introduction to coalitions in graphs*. AKCE International Journal of Graphs and Combinatorics 17(2), pp. 653–659, doi:10.1080/09728600.2020.1832874.

[15] Jobst Heitzig, Forest W Simmons & Sara M Constantino (2024): *Fair group decisions via non-deterministic proportional consensus*. Social Choice and Welfare N/A(N/A), pp. 1–27, doi:10.1007/s00355-024-01524-3.

[16] Pavel Janovsky & Scott A DeLoach (2016): *Multi-agent simulation framework for large-scale coalition formation*. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, N/A, pp. 343–350, doi:10.1109/WI.2016.0055.

[17] Anna Maria Kerkmann (2022): *Stability, Fairness, and Altruism in Coalition Formation*. In Dorothea Baumeister & Jörg Rothe, editors: Multi-Agent Systems, Springer International Publishing, Cham, pp. 427–430, doi:10.1007/978-3-031-20614-6_25.

[18] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa (2022): *Large language models are zero-shot reasoners*. Advances in Neural Information Processing Systems 35, pp. 22199–22213, doi:10.48550/arXiv.2205.11916.

[19] Hideo Konishi & Debraj Ray (2003): *Coalition formation as a dynamic process*. Journal of Economic Theory 110(1), pp. 1–41, doi:10.1016/S0022-0531(03)00004-8.

[20] Sarit Kraus (1997): *Negotiation and cooperation in multi-agent environments*. Artificial Intelligence 94(1–2), pp. 79–97, doi:10.1016/S0004-3702(97)00025-8.

[21] Laurens Van der Maaten & Geoffrey Hinton (2008): *Visualizing data using t-SNE*.

[22] Francisco P Maturana & Douglas H Norrie (1996): *Multi-agent mediator architecture for distributed manufacturing*. Journal of Intelligent Manufacturing 7, pp. 257–270, doi:10.1007/BF00124828.

[23] Anna Mikhaylovskaya & Élise Rouméas (2024): *Building trust with digital democratic innovations*. Ethics and Information Technology 26(1), p. 1, doi:10.1007/s10676-023-09736-4.

[24] Dominik Peters (2024): *Proportional Representation for Artificial Intelligence*. In: *ECAI 2024*, IOS Press, pp. 27–31, doi:10.3233/FAIA240463.

[25] Talal Rahwan (2007): *Algorithms for coalition formation in multi-agent systems*. Ph.D. thesis, University of Southampton, N/A.

[26] Talal Rahwan, Tomasz P Michalak, Michael Wooldridge & Nicholas R Jennings (2015): *Coalition structure generation: A survey*. *Artificial Intelligence* 229, pp. 139–174, doi:10.1016/j.artint.2015.08.004.

[27] Ariel Rosenfeld & Sarit Kraus (2016): *Strategical argumentative agent for human persuasion*. In: *ECAI 2016*, IOS Press, pp. 320–328, doi:10.3233/978-1-61499-672-9-320.

[28] Samriddhi Sarkar, Mariana Curado Malta & Animesh Dutta (2022): *A survey on applications of coalition formation in multi-agent systems*. *Concurrency and Computation: Practice and Experience* 34(11), p. e6876, doi:10.1002/cpe.6876.

[29] Erich Schubert & Arthur Zimek (2019): *ELKI: A large open-source library for data analysis-ELKI Release 0.7.5 "Heidelberg"*. *arXiv preprint arXiv:1902.03616* N/A(N/A), p. N/A, doi:10.48550/arXiv.1902.03616.

[30] Ehud Shapiro & Nimrod Talmon (2022): *Foundations for grassroots democratic metaverse*. *arXiv preprint arXiv:2203.04090* N/A(N/A), p. N/A, doi:10.48550/arXiv.2203.04090.

[31] Onn M Shehory, Katia Sycara & Somesh Jha (1998): *Multi-agent coordination through coalition formation*. In: *Intelligent Agents IV Agent Theories, Architectures, and Languages: 4th International Workshop, ATAL'97 Providence, Rhode Island, USA, July 24–26, 1997 Proceedings 4*, Springer, N/A, pp. 143–154, doi:10.1007/BFb0026756.

[32] Shuohang Wang & Jing Jiang (2016): *A compare-aggregate model for matching text sequences*. *arXiv preprint arXiv:1611.01747* N/A(N/A), p. N/A, doi:10.48550/arXiv.1611.01747.

[33] Tom Wanyama & Behrouz H Far (2006): *Negotiation coalitions in group-choice multi-agent systems*. In: *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, N/A, N/A, pp. 408–410, doi:10.1145/1160633.1160704.

[34] Joshua C Yang, Marcin Korecki, Damian Dailisan, Carina I Hausladen & Dirk Helbing (2024): *Llm voting: Human choices and ai collective decision making*. *arXiv preprint arXiv:2402.01766*, doi:10.1609/aies.v7i1.31758.

[35] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong et al. (2023): *A survey of large language models*. *arXiv preprint arXiv:2303.18223*, doi:10.48550/arXiv.2303.18223.

[36] Gil Ben Zvi, Eyal Leizerovich & Nimrod Talmon (2021): *Iterative Deliberation via Metric Aggregation*. In: *International Conference on Algorithmic Decision Theory*, Springer, N/A, pp. 162–176, doi:10.1007/978-3-030-87756-9_11.

A Missing Text

A.1 A Concrete Example

Consider the following, toy example.

Example 12. *Consider a metric space X with a set of elements P and a given distance d. We have a status quo $r \in P$ and three agents A, B, and C, each with its ideal point, p^A, p^B, p^C. Furthermore:*

- *each agent is non-altruistic ($\sigma = 0$);*
- *there is no coalition discipline;*
- *the mediator's α is set to be 0;*
- *$\{p^A, p^B, p^C\}$ serve as the initial singleton coalition points.*

The distances between the different ideal points of the agents and the status quo within the metric space are as follows (note that it is indeed a metric): $\mathbf{d}(p^A,p^A)=\mathbf{d}(p^B,p^B)=\mathbf{d}(p^C,p^C)=\mathbf{d}(p^D,p^D)=0;\mathbf{d}(p^A,p^B)=3;\mathbf{d}(p^A,p^C)=5;\mathbf{d}(p^A,r)=9;\mathbf{d}(p^B,p^C)=2;\mathbf{d}(p^C,r)=6;\mathbf{d}(p^C,r)=8$. *Consider another element of the metric space,* d^{BC}*, with* $\mathbf{d}(p^B,p^{BC})=\mathbf{d}(p^C,p^{BC})=1$.

1. ***Initialization:*** *Each agent starts with its own coalition.*

$$D=\{(C_A,p^A),(C_B,p^B),(C_C,p^C)\}$$
$$C_A=\{A\},\quad C_B=\{B\},\quad C_C=\{C\}\ .$$

2. ***Iteration 1:*** *The mediator suggests the compromise point* p^{BC} *to the coalitions* (C_B,p^B) *and* (C_C,p^C)*. Both agents approve since* $1<6$ *and* $1<8$*. We arrive to the following coalition structure* D'*:*

$$D'=\{(C_A,p^A),(C_{BC},p^{CB})\}\ ,$$
$$C_{BC}=\{B,C\}\ .$$

3. ***Halting condition:*** *A coalition with an agent majority has been formed (as* $|C_{BC}|/|D|>0.5$*), thus the process halts.*

A.2 Mediators in a 2D Euclidean Space

In this section, we consider a rather simple setting where the metric space X contains points in a 2D Euclidean space and the distance is ℓ_2. This serves to illustrate the fundamental properties of our model and showcases the operation of our algorithms. As a usecase, consider a scenario in which an agent community collaborates to mutually select a location for a social event (e.g., a picnic).

A.2.1 Simulation-Based Analysis

We describe the design of the computer-based simulations we have conducted; and report and discuss the results.

We have generated instances of our model for the realization described above for a 2-dimensional Euclidean space. Next are details of the specific configuration used:

- Status Quo: Generated uniformly at random between $(0,200)\times(0,200)$.
- Ideal Points: Drawing inspiration from the literature [8], each agent was assigned an ideal point (x,y) with both coordinates sampled from either the uniform distribution between $(0,200)$ or from a 2-dimensional Gaussian Mixture Model (GMM). The GMM represents the overall probability distribution as a weighted sum of several Gaussian components with multiple peaks. In our simulations, we considered GMMs with g combined Gaussian distributions, for $g\in\{0,1,2,3,4\}$ with the mean of each Gaussian being distributed uniformly between 0 and 200 in each dimension, its deviation distributed uniformly between 0 and 50, and the weights signifying the importance of each Gaussian are distributed from the $Dirichlet(\alpha^g\in\mathbb{R}_0^{+g})$ distribution with α set to 1 (resulting in g numbers that sum to 1). This sampling of ideal points is demonstrated in the supplementary material. Note that we treat GMM with $g=0$ (i.e., 0 peaks) as the uniform distribution.
- The different number n of agents used in the simulations was $n\in\{10,20,30,40,50,100,250,1000\}$.
- Coalition Discipline: we evaluated and compared instances with coalition discipline and without (as described in the realization Subsection in the main text.

- For the mediator, we have used $\alpha = \{-1, 0, 1\}$ (as described in realization Subsection int he main text).
- We have used $\sigma_v \in \{0, 10, 20, 30\}$ as the degree of altruism, representing the smoothing of agents' approval functions (as described in the realization Subsection in the main text).
- For the initialization of the singleton coalitions we set a parameter $I \in \{True, False\}$: for $I = False$ the initial singleton coalition points were set to be the ideal points of each agent; while for $I = True$, the initial singleton coalition points were generated using a 2-dimensional Gaussian distribution with a mean being the ideal point p^v and with a covariance matrix with $\sigma_x \sim U(0, 10)$ and $\sigma_y \sim U(0, 10)$ on the diagonal (and zeros off-diagonal).

We conducted 100 independent repetitions for each configuration. Next we present our two evaluation metrics (the first measures the process speed, while the second measures the process quality):

- **Speed of convergence**: average number of iterations until the halting condition is met.
- **Quality of converged state**: average distance between the proposal of the coalition containing an agent majority to the ideal points of the agents within that coalition; formally, for the single coalition $d = (C, p)$ in the halting state, with $\frac{|C|}{n} \geq 0.5$ we compute $\frac{1}{|C|} \sum_{v=1}^{|C|} \mathbf{d}(p, p^v)$.

For efficiency, we halt our simulations whenever the number of iterations exceeds a threshold of $10,000$ (i.e., we treat an instance for which no convergence is reached within $10,000$ iterations as an instance that does not converge at all).

A.2.2 Results and Discussion

Next we discuss the main conclusions, drawn at a 5% significance level:

1. Processes with coalition discipline and non-altruist agents agents always result in some non-convergences (i.e., the number of iterations is greater than 10,000) while all other combinations result in convergence.
2. More agents result in more iterations (linearly), shorter mean distances, and higher log-odds of a converging process before 10,000 iterations.
3. Higher α leads to a larger mean average distance.
4. Coalition discipline shortens the mean average distance.
5. High interaction between n and α, n and coalition discipline, and σ and coalition discipline results in more iterations until the halting condition.
6. High interaction between n and number of peaks (GMM), and coalition discipline and number of peaks (GMM), leads to fewer iterations until the halting condition.

B Illustrating the Simulation Framework

Example 13. *To better illustrate the process, we present one of the simulations conducted with fixed parameters outlined in Figure 1. The simulation involves 10 ideal sentences of agents regarding dealing with global warming (of maximum 15 words) projected onto a 2D Euclidean space, showcasing the coalitions each agent belongs to by the time the process concludes (i.e., the halting condition is satisfied). The visualization method employed for multi-dimensional data is adapted from [21].*

Implementation, Code, and further Illustrations can be found here `https://github.com/EyalBriman/AI-Mediator`.

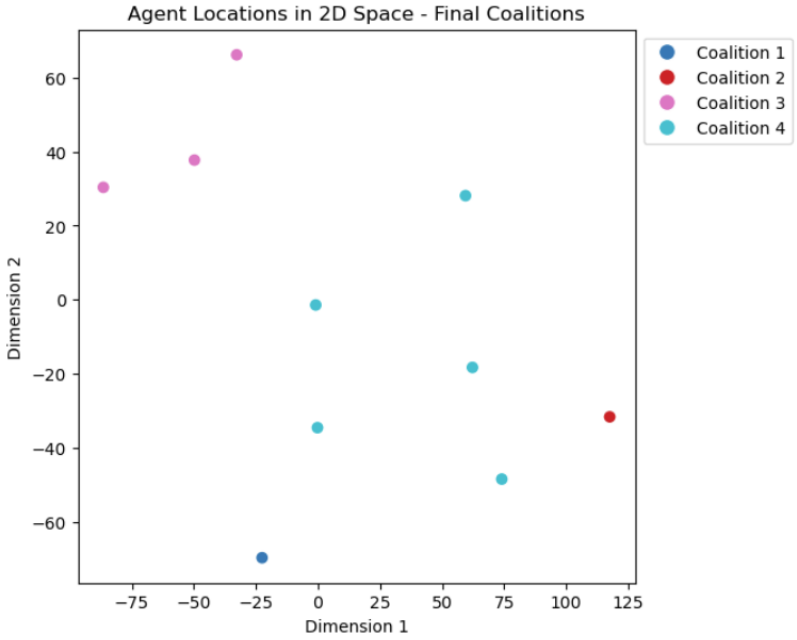


Figure 1: Coalition formation result- "Dealing with Global Warming", for $n = 10, C = \text{False}, \alpha = 0, \sigma = 0, I = \text{True}$.

Muddy Waters

Hans van Ditmarsch
University of Toulouse, CNRS, IRIT
hansvanditmarsch@gmail.com

In the 2013 Advent calender of the Berlin Mathematics Research Center MATH+, Gerhard Woeginger presents a novel hat problem with an uncommon initial announcement. Although the information given is insufficient for the hat bearers to learn their colour, they are informed that the colours have been chosen so that they can learn their colour. We formalize this announcement in public announcement logic and in an extension of public announcement logic with fixpoints.

1 A new hats riddle

1.1 History of the muddy children puzzle

The Muddy Children puzzle and similar hats riddles have delighted puzzle book enthusiasts since at least the 1930s [23, 25, 33, 19, 32] (with older roots going back to at least the early 19th century [16], see [13, 14] for further notes on its history) and has also fuelled the development of artificial intelligence and of dynamic epistemic logic [26, 2, 29, 17, 31, 10].

From an epistemic logical perspective such knowledge puzzles tend to follow a certain pattern. First, the problem solver has to represent the initial uncertainty of the agents in a multi-agent Kripke model. This initial model represents the background knowledge. Second, for the agents this may be insufficient information in order to solve their uncertainty by ignorance or knowledge statements (for example, when their ignorance is already common knowledge). Some extra information (an initial announcement) is then needed in order to make the agents' ignorance statements informative. Third, the agents then keep going announcing their ignorance until the problem is solved. In the case of muddy children: (i) the initial model encodes that the children can only see the mud on each other's foreheads (it is common knowledge that no child knows whether it is muddy); (ii) father then announces that at least one child is muddy; (iii) the children keep 'announcing' their ignorance of muddy in rounds until they know whether they are muddy. To synchronize the children's behaviour, a round is characterized by father repeating his request, or clapping in his hands, or ringing a bell. Stage (ii) to 'get the induction going' is not always there, for example not in Consecutive Numbers [25] and not in Sum and Product [18].

1.2 Mützen

The Berlin Mathematics Research Center MATH+ annually presents an Advent calender with mathematical riddles, of which the solutions are subsequently given after Christmas. In the 2013 calender problem number 10, by Gerhard Woeginger, is called Mützen ('Hats') [34]. It is a variation of the muddy children and hats puzzles with many hats and many announcements, in a Christmas setting where Santa informs his Little Helpers ('Wichtel', Gnomes) of the solution requirement. It is accompanied by a beautiful illustration implicitly challenging the reader to determine the famous mathematicians depicted there as gnomes with coloured hats (René Descartes, Leonhard Euler and Kurt Gödel are among them). Mützen is given in the Appendix in its original playful and beautiful German version and in English translation.

A. Bjorndahl (Ed.): Theoretical Aspects of
Rationality and Knowledge 2025 (TARK 2025)
EPTCS 437, 2025, pp. 433–447, doi:10.4204/EPTCS.437.33

© Hans van Ditmarsch
This work is licensed under the
Creative Commons Attribution License.

In the riddle's description we can easily see the three stages needed to solve it. First, the initial model encodes 126 gnomes only seeing the colours of others' hats. Nothing is known about these colours. Second, Santa makes a, in this case, curious initial announcement. We call this announcement **solvable**.

> *I chose the hat colours very carefully so that each of you can actually determine their colour through thinking during the game.* (**solvable**)

Third, the bell starts ringing, where at each ring those leave who know their colour, until finally all are having tea and cake.

What does Santa's initial announcement **solvable** say? In muddy children and hat problems it is always given what the set of colours is, or that children can be muddy or not, which can be seen as the set of colours consisting of black and white. In Mützen we only know that the hats are coloured. Just as in the muddy children problem and in other hat problems, without additional information that the problem solver provides ('at least one child is muddy') they never learn anything from the bell being rung and such.

The information content of **solvable** is much harder to grasp than that of father's announcement that at least one child is muddy. The latter can be easily verified, unlike the former. Atto gives an important hint when he says to Santa 'For example, if each of us had a different colour of hat, then no one would be able to figure out what colour it is'. Let us be precise about Atto's observed inability to solve the riddle.

If all gnomes wear a different unique colour, any gnome i wearing colour c also considers it possible that its hat has colour c' for some colour $c' \neq c$ that is also not worn by any of the other gnomes. Therefore no gnome knows the colour of its hat and therefore no one will leave the room when the bell rings. And also not in the next iteration. Not ever. The later distribution of hats reveals that Atto cannot have seen 125 all different colours. But this does not matter: as long as any gnome considers it possible that its own colour is unique, it considers it possible that this is any of two unseen colours and will therefore not leave the room when the bell rings. Therefore, it cannot have a unique colour, and therefore no gnome can have a unique colour, and therefore of each colour occurring there must be at least two hats. So that all the colours it sees, are all the colours there are. This demonstrates that the announcement **solvable** is therefore *at least as* informative as that of the announcement:

> *For each coloured hat worn by a gnome there must be another gnome wearing a hat with the same colour.* (**solvable**$'$)

We call that announcement **solvable**$'$. This is indeed the spark that gets the induction going, so that finally all can determine their colour. That demonstrates that **solvable** is *at most as* informative as **solvable**$'$. Therefore, **solvable** and **solvable**$'$ have the same information content.

Let us sketch how the iterations proceed that solve the problem:

At the first iteration all gnomes seeing only one gnome with a hat of a particular colour conclude that they must also have that colour, and leave the room. At the second iteration all (only) seeing two gnomes with a hat of a particular colour, leave. And so on. If in iteration n no one sees n gnomes of some colour, no one leaves the room; which is on condition, fulfilled in the riddle, that different colours can be seen. For example, let there be only 12 gnomes, with 2 white and 2 black and 8 rose hats. Then at ring 1 the gnomes with white and black hats leave, but already at ring 2 the gnomes with rose hats leave. *It does not take more rounds.* But if there had been 10 more pink and 10 more yellow hatted gnomes, there would have been no such shortening of rounds for the 8 rose hatted gnomes. Because then: at ring 1 white and black leave, at rings 2 to 6 no one leaves, at ring 7 the gnomes with rose hats leave, at ring 8 noone leaves, and at ring 9 the gnomes with pink and yellow hats leave. In the Mützen riddle shortening of rounds does not play a role, as there are two groups of the same maximum size that step forward in the last round. We do not know if this was by design, in order to avoid this extra modelling complication.

1.3 Solvable and unsolvable variations of muddy children

Santa's announcement **solvable** reduces the set of all colour distributions to the largest subset X such that all gnomes will leave the room after some finite iteration of ringing the bell, which informally means that this is a *fixpoint* with respect to solving the problem: any set of colour distributions of which X is a strict subset will not solve the problem, although there are sets of colour distributions that are strict subsets of X and that solve the problem. Furthermore, no actual colours are mentioned in **solvable**.

Before we delve into the logic and formally introduce fixpoints, let us investigate fixpoints in the simpler setting of the muddy children problem, with colours black (muddy) and white (clean), and for three children only. First, the standard solution: given that children only see the foreheads of others, father says that at least one child is muddy, after which at each round the children who know whether they are muddy step forward. This takes at most three rounds.

In Figure 1.i we visualize the execution of this branching protocol in a linear way, as iterated refinement of epistemic models, because all branching is a consequence of mutually exclusive public announcements. We also see father's initial announcement not as a model restriction but as a model refinement. The cube and its subsequent submodels represent the uncertainty of children a, b, c. The states (or worlds) are named with triples of digits indicating whether child a, b, c in this order are muddy (1) or clean (0). Agent labelled links between states indicate that the states are indistinguishable for those agents. The double arrows indicate model updates. Each arrow linking two models represents non-deterministic choice between mutually exclusive alternatives, where the first arrow is choice between two alternatives, often known as a *test* (whether), denoted with '?'. The final 'arrow' is a refinement that does not change the model and is therefore denoted $=$ instead of $\Rightarrow$.

Different ways to solve muddy children We will not explain this standard solution of the muddy children problem for the umptieth time. What we wish to point out is that apart from this standard initial refinement into a seven-state and a one-state model, in some other initial refinements all children also eventually learn whether they are muddy, whereas in yet other initial refinements this is not the case. We will not systematically review all these restrictions, but merely make some pregnant observations:

First, *any* initial restriction to seven states results in all children learning whether they are muddy, and by the exact same updates.

Second, *some* restrictions result in some children learning whether they are muddy, but not all, or not always. A typical one is the refinement where we separate an 'edge', such as 000—a—100 from its complement. In both parts a will never learn whether it is muddy. There is too much symmetry in the restriction.

Third, it is not so clear how to see any of the restrictions where all learn whether they are muddy as a fixpoint. In particular, the announcement **solvable**$'$ that there are at least two hats of every colour is trivialized because there are fewer colours (2) than agents (3). It is the restriction to $\{000, 111\}$. It is not a fixpoint, e.g. Figure 1.i and 1.ii both also contain 111. It is not even colour-blind: there are other restrictions that are also invariant for any permutation of colours over hats. For example, any other opposite states such as $\{011, 100\}$.

Fourth, did you, reader, ever realize that there are really two different ways to solve the muddy children problem? Take Figure 1.ii where in the three-state restriction $\{000, 100, 110\}$ after two rounds of the **bell** finally all children have stepped forward, whereas in the five-state complement restriction this takes three rounds instead. Now compare this to Figure 1.iii where the updates are a function of the announcement that nobody knows whether they are muddy (or its negation), later formalized as $\mathbf{bell}_\emptyset$. The updates are different! How come?

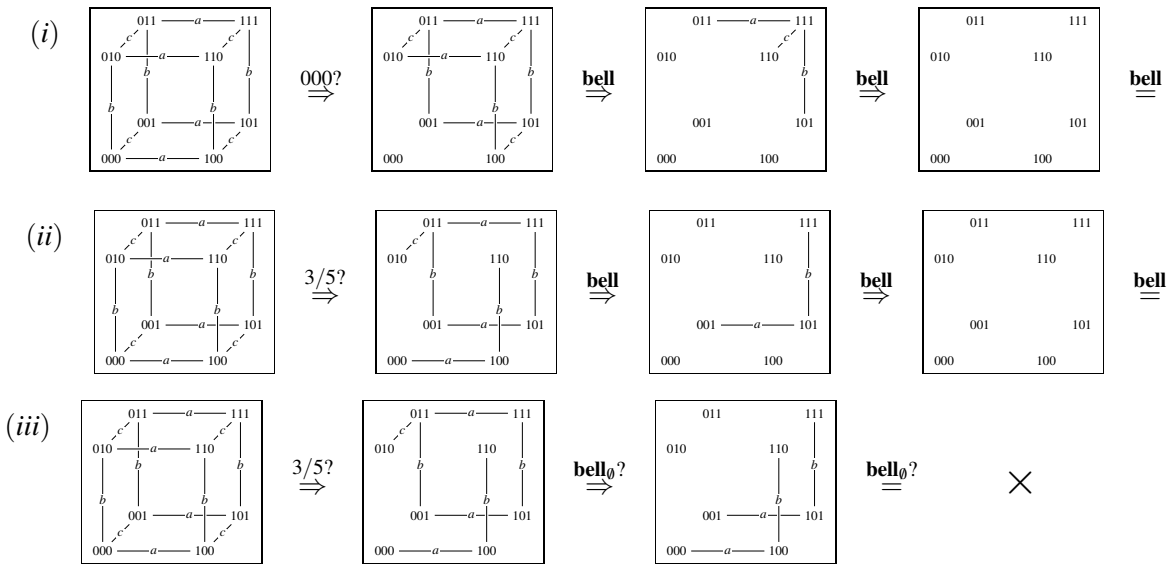


Figure 1: Different ways to solve three muddy children

In Figure 1.ii the updates represent that the children who know whether they are muddy step forward. The update **bell** is therefore non-deterministic choice between eight different announcements of which **bell**$_\emptyset$, nobody knows whether they are muddy, is only one. In Figure 1.iii the update **bell**$_\emptyset$? is non-deterministic choice between **bell**$_\emptyset$ and its negation: somebody knows whether they are muddy. These updates are different. For example, in the three-state restriction, in 000 only children b and c know whether they are muddy, in 100 only c knows that, and in 110 only a and c know that. Therefore at the first **bell** child a learns the difference between 000 and 100 (depending on whether b steps forward) and child b learns the difference between 100 and 110 (depending on whether child a steps forward). The links are cut. Similarly for the five-state restriction of the model at the second **bell**. Either way, eventually all children get to know whether they are muddy, and the problem is solved. But not with **bell**$_\emptyset$ updates. Some links are never cut.

Now do the same in Figure 1.i. There, it does not matter if the updates are with **bell**$_\emptyset$? or with **bell**.

After all these years teaching logic puzzles to audiences of school children and students, it came somewhat as a surprise to us that there are really different formalizations of muddy children, and that one is more faithful to the actions of individual children stepping forward than the other one.

Towards Mützen Let us now consider three children and four 'colours' $0,1,2,3$, so that even if all three children have a unique colour, a child is uncertain about its own colour. The uncertainty of the agents permits a similar cube model representation consisting of 64 states and with subdivided edges such as 000—a—100—a—200—a—300; we will not draw it. The restriction on this model due to **solvable** consists of states

$$\{000, 111, 222, 333\}$$

that are all singleton equivalence classes in which all three children will step forward immediately. Which is not a fixpoint. Consider the restriction consisting of

$$\{000, 111, 222, 333\} \cup \{100, 120\}$$

that contains a connected part 120—*b*—100—*a*—000. At the announcement of **bell**, in 111, 222, 333 all children know their colour and will step forward. In 000, *b* and *c* know that they are 0, in 100 only *c* knows it is 0, and in 120 *a* knows that it is 1 and *c* knows that it is 0. And in those cases all those would then step forward. Therefore, *a* learns the difference between 000 (wherein *b* knows its value) and 100 (wherein *b* does not know), and will step forward at the second **bell**. Similarly, *b* then learns the difference between 100 and 120. But, of course, the above restriction to $\{000,111,222,333\} \cup \{100,120\}$ is not invariant for permutations of colours; it is different from, for example, the restriction to

$$\{000,111,222,333\} \cup \{022,032\}$$

resulting from permutation $\iota(0) = 2$, $\iota(1) = 0$, $\iota(2) = 3$, $\iota(3) = 2$.

We are not there yet, because now consider another diagonal of the 64 state cube:

$$\{033,122,211,300\}$$

If we swap 1 and 3 we obtain

$$\{011,322,233,100\}$$

which is a different restriction. Although some permutations, like swapping 1 and 2, leave it unchanged. Restriction $\{000,111,222,333\}$ is the only one that is invariant for *any* permutation of colours, maximal, and such that all children always get to know their colour. In this case, in a boring way, because they immediately know. The complement of $\{000,111,222,333\}$ is also invariant for any permutation of colours and maximal. But the children now never get to know their colour (unlike in Figure 1.iii): any state *xyz* is indistinguishable for *a* from *vyz* where $v \neq x,y,z$. Therefore, $\{000,111,222,333\}$ is the only restriction satisfying **solvable** and **solvable**$'$.

All this, for three children and four colours, is just as in Mützen, only there it takes many rounds.

2 Mützen in epistemic logic

2.1 Introduction

We first formalize Mützen in standard public announcement logic where we have simplified the **solvable** announcement to the **solvable**$'$ announcement stating that no gnome wears a hat of a unique colour. We formalize it in an extended language including fixpoints. First, we survey the relevant literature.

Public announcement logic Public announcement logic [31] is a modal logic that is an extension of multi-agent epistemic logic (the logic of knowledge) [21] with a modality expressing the consequences of all agents synchronously observing (or hearing, or receiving) new information that is considered reliable. It is an example of what are known as dynamic epistemic logics [10, 6]. It contains modalities $K_a\varphi$ for 'agent *a* knows that φ' and modalities $[\psi]\varphi$ for 'after truthful public announcement of ψ, φ (is true)'. Works such as [10, 6] and many others are suitable background reading.

Fixpoints We can extend the language of public announcement logic with *fixpoints*. Fixpoint semantics have been investigated for epistemic modal logics in [20, 30, 4, 1, 24] (and in an epistemically motivated but not strictly epistemic setting in [5, 35]), where they are a more or less straightforward adaptation of those for the modal μ-calculus [22]. We add another inductive construct $\nu p.\varphi$ to the language. In fixpoint modal logic all occurrences of *p* in φ must be *positive*: bound by (zero or) an even number of negations. In fixpoint public announcement logic it is not that simple, as we will see, and our concern is only with existence of fixpoints.

Arbitrary iteration of announcements Arbitrary iteration of announcements is uncommon fare in dynamic epistemic logics, as the (satisfiability problem of the) resulting logic is undecidable and axiomatization are unknown [27]. Arbitrary iteration will be added as yet another inductive construct $[\varphi]^*\psi$ to the logical language.

Public assignments Assignments σ are functions that, for a finite subset of all atoms (for technical reasons), map each atom p in that set to a formula $\sigma(p)$ in the logical language. For $\sigma(p) = \varphi$ we also write $p := \varphi$. They come with yet another inductive construct $[\sigma]\varphi$. See e.g. [12, 9, 7]. It has been frequently used in works bridging the gap between epistemic logic and applications in artificial intelligence, such as epistemic planning [8, 3]. (Like public assignments, fixpoints also change the value of variables in a given model, namely that of the fixpoint variable p in $\nu p.\varphi$.)

2.2 Public announcement logic with fixpoints and assignments

We succinctly define the language, structures, and semantics.

Language Given disjoint countable sets of agents and atoms (the set of agents is often required to be finite), the language of public announcement logic is inductively defined as that of the *formulas*

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid K_a\varphi \mid [\psi]\varphi \mid \nu p.\varphi \mid [\psi]^*\varphi \mid [\sigma]\varphi$$

where p is an atomic proposition and a is an agent, where in clause $\nu p.\varphi$ atom p may have to satisfy a requirement (discussed later), and where σ is an assignment. Any other propositional connnectives are defined by notational abbreviation, as well as the dual modalities $\widehat{K}_a\varphi := \neg K_a\neg\varphi$, $\langle\varphi\rangle\psi := \neg[\varphi]\neg\psi$, $\langle\varphi\rangle^*\psi := \neg[\varphi]^*\neg\psi$, $\mu x.\varphi := \neg\nu x.\neg\varphi[x/\neg x]$ (where $\varphi[x/\neg x]$ denotes uniform substitution of all *free* —not bound by another νx occurring in φ— occurrences of x in φ by $\neg x$), and non-deterministic choice between announcements $[\varphi \cup \psi]\chi := [\varphi]\chi \wedge [\psi]\chi$ (and similarly for the diamond version, $\langle\varphi \cup \psi\rangle\chi := \langle\varphi\rangle\chi \vee \langle\psi\rangle\chi$, which more intuitively reflects choice). No dual is needed for assignments as they are self-dual. We omit parentheses (and) unless this creates ambiguity. We also write $\varphi?$ for $\varphi \cup \neg\varphi$.

Structures The language is interpreted in epistemic models $(W, \sim, V)$ where W is a set of *worlds* (or *states*), where $\sim$ consists of binary *indistinguishability relations* $\sim_a$ between worlds for each agent a, that are equivalence relations, and where *valuation* V maps each atom to a subset of the set of worlds W (these are the worlds where the atom is true).

Semantics We then define the semantics as follows, by way of a *satisfiability relation* $\models$ between pairs (M, w) and formulas φ, where $M = (W, \sim, V)$ is an epistemic model, world $w \in W$, and where $[\![\varphi]\!]_M$ abbreviates $\{w \in W \mid M, w \models \varphi\}$.

$$\begin{array}{lll}
M,w \models p & \text{iff} & w \in V(p) \\
M,w \models \neg\varphi & \text{iff} & M,w \not\models \varphi \\
M,w \models \varphi \wedge \psi & \text{iff} & M,w \models \varphi \text{ and } M,w \models \psi \\
M,w \models K_a\varphi & \text{iff} & M,v \models \varphi \text{ for all } v \text{ such that } w \sim_a v \\
M,w \models [\psi]\varphi & \text{iff} & M,w \models \psi \text{ implies } M|\psi, v \models \varphi \\
M,w \models [\psi]^*\varphi & \text{iff} & M,w \models [\psi]^n\varphi \text{ for all } n \in \mathbb{N} \\
M,w \models \nu p.\varphi & \text{iff} & M^{p\mapsto U}, w \models \varphi \text{ where } U = \bigcup\{X \subseteq W \mid X \subseteq [\![\varphi]\!]_{M^{p\mapsto X}}\} \\
M,w \models [\sigma]\varphi & \text{iff} & M^\sigma, w \models \varphi \text{ where } M^\sigma = (W, \sim, V^\sigma) \text{ with } V^\sigma(p) = [\![\psi]\!] \text{ for all } \sigma(p) = \psi
\end{array}$$

Here, $M|\psi$ is the *restriction* (or *update*) of the model M to the worlds in $[\![\psi]\!]_M \subseteq W$, defined as $M|\psi := \{W', \sim', V'\}$ where $W' = [\![\psi]\!]_M$, $\sim'_a := \sim_a \cap (W' \times W'$ for each agent a, and $V'(p) := V(p) \cap W'$ for each atom p. Then, for arbitrary $X \subseteq W$, $M^{p \mapsto X} := (W, \sim, V')$ where $V'(p) = X$ and $V'(q) = V(q)$ for $q \neq p$. Furthermore, $[\psi]^n \varphi$ represents iteration of the announcement ψ ($[\psi]^0 \varphi := \varphi$, $[\psi]^{n+1} \varphi := [\psi][\psi]^n \varphi$ and similarly for $\langle \psi \rangle^n \varphi$). A formula is *valid* if it is true in all worlds of all models.

We can also directly give the semantics of least fixpoint namely as:

$$M, w \models \mu p.\varphi \quad \text{iff} \quad M^{p \mapsto U}, w \models \varphi \text{ where } U = \bigcap \{X \subseteq W \mid [\![\varphi]\!]_{M^{p \mapsto X}} \subseteq X\}$$

The main thing about public announcement logic is that the meaning of $[\varphi]\psi$ (if φ is true and announced then ψ is true) is different from the meaning of $\varphi \to \psi$ (if φ is true then ψ is true), because in the former case ψ is interpreted in a different, updated, model. For example, $p \to K_a p$ is invalid (not all propositions are known!) whereas $[p]K_a p$ is valid (if we make p public, then a knows it). The puzzling feature of the logic is that a true formula may be false after its announcement, as in the muddy children puzzle (I know whether I am muddy / I don't know whether I am muddy), and the most typical example of that is the so-called Moore-sentence [28] $p \wedge \neg K_a p$ that is false after its announcement. It is easy to see that $[p \wedge \neg K_a p] \neg (p \wedge \neg K_a p)$ is even valid.

2.3 Positive and negative results about fixpoints

For the fixpoint semantics we did not specify the requirement for them to exist, such as occurring positively. The issues combining announcements and fixpoints are well-known [4, 1, 24]. We review some of the issues and have some minor novel results.

What is positive? How do we determine whether a fixpoint variable is 'positive' in a public announcement formula? Consider again a formula of shape $[\varphi]\psi$. When ψ is an atom p, $[\varphi]p$ is (always) equivalent to $\varphi \to p$, as atoms do not change their truth value in an updated model. Now consider $[\neg q]p$. This formula $[\neg q]p$ is equivalent to $\neg q \to p$, that is, $\neg\neg q \vee p$ (or $q \vee p$), wherein the occurrence of q is positive. We therefore should call the occurrence of q in $[\neg q]p$ positive because its occurrence in the formula $\neg q$ of the public announcement is *negative* (where negative is 'not positive'). So that, for example, the greatest fixpoint $\nu q.[\neg q]p$ exists.

But we cannot simply define that p is positive in $[\varphi]\psi$ iff it is negative in φ and positive in ψ (and that p is positive in $\neg\varphi$ iff it is negative in φ), from which we obtain that p is positive in $\langle \varphi \rangle \psi$ iff is is positive in φ and in ψ. Consider $\langle q \rangle \widehat{K}_a p$ and $\langle q \rangle K_a p$. Using the definitions by abbreviation and the reduction rules for public announcement [10], the former can be rewritten to: $\langle q \rangle \widehat{K}_a p$, iff $q \wedge \widehat{K}_a \langle q \rangle p$, iff $q \wedge \widehat{K}_a (q \wedge p)$. Both occurrences of q are positive. No problem. On the other hand, we rewrite the latter to: $\langle q \rangle K_a p$, iff $\neg[q]\neg K_a p$, iff $\neg(q \to \neg[q]K_a p)$, iff $\neg(q \to \neg(q \to K_a[q]p))$, iff $\neg(q \to \neg(q \to K_a(q \to p)))$, which with a bit of propositional work is equivalent to $q \wedge K_a(\neg q \vee p)$. Problem. The second occurrence of q is negative. This means that some smart strategy to expand the negation normal form $\varphi ::= p \mid \neg p \mid \varphi \vee \psi \mid \varphi \wedge \psi \mid \widehat{K}_a \varphi \mid K_a \varphi$ with clauses $\langle \psi \rangle \varphi$ and $[\psi]\varphi$ for announcement does not lead towards syntactic characterization of positivity of designated fixpoint variables in public announcement formulas. Note that the problematic $\langle q \rangle K_a p$ above is in this fragment.

Still, the following holds (where $\varphi(q)$ is a formula with possible occurrences of atom q).

Proposition 1 *If a public announcement formula $\varphi(q)$ is equivalent to a multi-agent epistemic formula wherein q occurs positively, then the fixpoint $\nu q.\varphi(q)$ exists.*

The proof is obvious as any public announcement formula is equivalent to one without announcements [31]. Note that we do not say '*any* equivalent epistemic formula'. For example, fixpoint $\nu q.[\neg q]p$ exists because given $[\neg q]p$, q is positive in the equivalent $q \vee p$. But it is not positive in the also equivalent $(q \vee p) \wedge (q \vee \neg q)$.

The fixpoint positive fragment That p is positive in a formula $[\varphi]\psi$ when it is negative in φ is not an unknown phenomenon, as it can be shown that the (also known as) positive fragment $\varphi ::= p \mid \neg p \mid (\varphi \vee \psi) \mid (\varphi \wedge \psi) \mid K_a\varphi \mid [\neg\psi]\varphi$ preserves truth after update [11, 15]. If we expand this fragment to the *fixpoint positive* fragment

$$\varphi ::= q \mid p \mid \neg p \mid (\varphi \vee \psi) \mid (\varphi \wedge \psi) \mid K_a\varphi \mid [\neg\psi]\varphi$$

where atoms q are designated fixpoint variables and atoms p are non-designated, then:

Proposition 2 *Fixpoints $\nu q.\varphi(q)$ exist for all formulas $\varphi(q)$ in the fixpoint positive fragment.*

Proof By induction on $\varphi(q)$ in the fixpoint positive fragment we extend the proof in [11] to show that there is an equivalent formula without public announcements wherein all occurrences of q are bound by (zero or) an even number of negations. We then use Proposition 1. □

We conjecture that fixpoints even exist for formulas in a further extended fixpoint positive fragment namely with an inductive clause $\nu q.\varphi$. Unfortunately, such results do not get us anywhere when modelling knowledge puzzles such as Muddy Children and Mützen, as the puzzling aspect is a consequence of flipping truth values of announced formulas, that are therefore not in the fixpoint positive fragment.

Existence of fixpoints We do not know of a syntactic characterization guaranteeing fixpoint existence in public announcement logic with fixpoints. When showing the existence of a fixpoint for Mützen we will therefore backtrack to directly showing their existence on the **muetzen** epistemic model, because, as we will see, not even monotonicity of our fixpoint formula is guaranteed for Mützen.

Assignments, announcements and fixpoints A useful property of the fixpoint semantics is that for atoms p and q, $\langle p\rangle\mu q.\varphi(q) \leftrightarrow p \wedge \mu q.\langle p\rangle\varphi(q)$ is valid [4, Fact 4]. However, it does not hold for announcements of (all) arbitrary formulas ψ instead of propositional variables p, and the interaction of assignments and announcements also cause additional complications [4, Fact 5]. Such complications are avoided in Mützen as our only assignments are permutations of colours over gnomes' hats.

Arbitrary iteration Whether arbitrary iteration of announcements is definable as a fixpoint is unclear, as in $[\varphi]^*\psi$ the meaning of φ can be different at each subsequent iteration. Given the flipping of positive into negative it is also unclear whether for $[\varphi]^*\psi$ one would rather have a least or a greatest fixpoint. We take the easy way out: to formalize Mützen we will not need arbitrary iteration, as there is an upper bound of the number of iterations of ignorance announcements in the inductive phase of solving the riddle. It is therefore sufficient to define $[\varphi]^*\psi := \bigwedge_{n=0}^{\max-1}[\varphi]^n\psi$, denoted $[\varphi]^{<\max}\psi$, where max is the number of colour distributions (the actual state will not be eliminated). Similarly, we define $\langle\varphi\rangle^{<\max}\psi$ as $\bigvee_{n=0}^{\max-1}\langle\varphi\rangle^n$. In that case we will see that this is, after all, a fixpoint $\mu q.\psi \vee \langle\varphi\rangle q$, where q does not occur in φ or ψ.

2.4 Formalizing Mützen in public announcement logic

We now can very efficiently formalize the Mützen riddle in public announcement logic for a set of agents that are 126 gnomes and a set of atomic propositions c_g for each gnome g and each colour c. Our encoding needs formulas wherein all atoms occur, so that we need to restrict ourselves to a finite set of colours. For our purposes it is then enough to have just one more colour than there are gnomes, ensuring that any gnome remains uncertain about its own colour, even it if knew that all gnomes wear different colours. So we have 127 colours. As worlds we take distributions $\delta : G \to C$ of colours over gnomes. We let $\Delta := C^G$ denote the set of colour distributions.

> *The epistemic model* **muetzen** *consists of the domain Δ of all distributions of colours over gnomes, where two worlds δ and δ' are indistinguishable for gnome g if it sees the same colours, so if for all $g' \neq g$, $\delta(g') = \delta'(g')$, and where $\delta \in V(c_g)$ if $\delta(g) = c$, that is, if g wears a hat of colour c.* (**muetzen**)

The announcement $\mathbf{solvable}'$ that no gnome wears a unique colour is the formula

$$\mathbf{solvable}' \quad := \quad \bigwedge_{g\in G}\bigwedge_{c\in C}(c_g \to \bigvee_{h\in G}^{h\neq g} c_h)$$

and the announcement **bell** of Santa ringing the Christmas bell is non-deterministic choice between mutually exclusive announcements $\mathbf{bell} := \bigcup_{L\subseteq G} \mathbf{bell}_L$ of formulas $\mathbf{bell}_L$ of L gnomes Leaving (**or having left**) the room and $G\backslash L$ gnomes staying in the room. As leaving means knowing and staying means being ignorant, we can define:

$$\mathbf{bell}_L \quad := \quad \bigwedge_{g\in L}\bigwedge_{c\in C}(K_g c_g \vee K_g \neg c_g) \wedge \bigwedge_{g\notin L} \neg \bigwedge_{c\in C}(K_g c_g \vee K_g \neg c_g)$$

The problem is solved when all know their colour, that is, when $\mathbf{bell}_G$ is true. To prevent Santa from having to ring the bell forever, it is preferable to distinguish rings where $L \neq G$ (there is remaining ignorance) from those where there is not, that we only want to feature once. We therefore define $\mathbf{bell}^{\mathbf{nlast}}$ as the non-deterministic announcement $\bigcup_{L\subset G} \mathbf{bell}_L$, and $\mathbf{bell}^{\mathbf{last}}$ as $\mathbf{bell}_G$. Given all this, the Mützen problem is to:

> *Determine $n \in \mathbb{N}^+$ such that* $\mathbf{muetzen}, \delta \models \langle \mathbf{solvable}' \rangle \langle \mathbf{bell}^{\mathbf{nlast}} \rangle^{n-1} \mathbf{bell}^{\mathbf{last}}$.

At the $(n-1)$th ring of the bell, all who did not yet know the colour of their hat now learn this and will therefore leave the room at the nth ring of the bell. Determining n requires the real colour distribution δ to become known to the problem solver. So, *which n* from 17—26 is the correct answer?

2.5 Formalizing Mützen with a fixpoint

We finally get to formalizing Mützen with a fixpoint, in the extended language. We keep the same set of propositional variables c_g for gnome g wearing a hat with colour c but we also need some designated fixpoint variables for which we write $p, q, \ldots$. First consider this (infelicitous) formalization of **solvable** (determine $n \in \mathbb{N}^+$ such that $\mathbf{muetzen}, \delta \models \langle \mathbf{solvable} \rangle \langle \mathbf{bell}^{\mathbf{nlast}} \rangle^{n-1} \mathbf{bell}^{\mathbf{last}}$):

$$\mathbf{solvable} \quad := \quad \nu p.\langle p \rangle \langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}} \qquad \times$$

As expected, we need a greatest and not a least fixpoint, as under the scope of an announcement positive become negative. However, the announcement $\langle p \rangle$ binds a plethora of knowledge and ignorance formulas, so that, even if we take $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^*$ to be $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^{<\max}$, there will not be an equivalent epistemic

formula obtained by eliminating announcements using the reduction axioms of public announcement logic wherein p is also positive (see **What is positive?** in the previous section). Could it be that maybe we do not need this to show existence? Indeed we don't, but let us postpone addressing this issue: this formalization of **solvable** is infelicitous, because the restriction of p to any singleton state in the model **muetzen**, including those where gnomes wear unique colours, already results in all gnomes knowing their colour. We resolve this by enforcing that only colour distributions can be considered that are invariant for permutations of colours. For this, we have the assignments σ and assignment modalities $[\sigma]$ at our disposal. Given a permutation $\iota : C \to C$ of the set of colours, we denote by $\boldsymbol{\iota}$ the assignment such that for all colours c and gnomes g, $\boldsymbol{\iota}(c_g) = \iota(c)_g$, and where $\boldsymbol{\iota}(p) = p$ for fixpoint variables p. Let I be the set of all permutations of colours. We also extend the use of colour permutations ι in the obvious way from colours to colour distributions and sets of colour distributions. Given colour distributions δ, δ', consider a relation $\delta \equiv_\iota \delta'$ if there is a permutation ι such that $\iota(\delta) = \delta'$. Relation $\equiv_\iota$ is an equivalence relation and induces a partition on the set of colour distributions Δ. An $\equiv_\iota$ equivalence class is characterized by a partition of all gnomes into *subsets of gnomes that have the same colour* (regardless of what that colour is).

Formalization of solvable This is our (felicitous) proposal for the formalization of **solvable**. Here, given a colour distribution $\delta : G \to C$ its *description* is defined as $\boldsymbol{\delta} := \bigwedge_{\delta(g)=c} c_g \wedge \bigwedge_{\delta(g)\neq c} \neg c_g$.

$$\textbf{solvable} \quad := \quad \nu p. \bigwedge_{\delta\in\Delta} ((\boldsymbol{\delta} \to p) \to \bigwedge_{\iota\in I} [\boldsymbol{\iota}](\boldsymbol{\delta} \to p)) \wedge \langle p \rangle \langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}}$$

Abbreviate the above formula **solvable** as $\nu p.\varphi(p)$. No matter how we wish to define positive, this formula is certainly not positive because of the part $\bigwedge_{\delta\in\Delta}((\boldsymbol{\delta} \to p) \to \bigwedge_{\iota\in I}[\boldsymbol{\iota}](\boldsymbol{\delta} \to p))$, where p is negative in the antecedent $\boldsymbol{\delta} \to p$ of the implication. So we have to determine the existence of a fixpoint directly.

There is one more obstacle on that path. The function $f_{\varphi(p)}$ mapping a subset $\Delta' \subseteq \Delta$ of colour distributions such that $[\![p]\!] = \Delta'$ to a subset $[\![\varphi(p)]\!] = \Delta'' \subseteq \Delta$ of colour distributions is **not** monotonic. However, it is peculiar. In the following, let a permutation-invariant Δ' be a Δ' such that $\iota(\Delta') = \Delta'$; this includes $\Delta' = \emptyset$.

- For any permutation-invariant Δ' that does not contain colour distributions with unique colours we have that $f_{\varphi(p)}(\Delta') = \Delta'$ so that $\Delta' \subseteq f_{\varphi(p)}(\Delta')$.
- For any permutation-invariant Δ' that contains colour distributions with unique colours we need to distinguish two cases. If Δ' only contains unique colour distributions we have that $f_{\varphi(p)}(\Delta') = \emptyset$. This is because $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}}$ then always fails: if $\delta' \in \Delta'$ and $\delta'(g) = c$ then also $\delta'' \in \Delta'$ that is as δ' except that $\delta''(g) = c'$ where c' is another unique colour. So, gnome g will never learn its colour. However, as also $f_{\varphi(p)}(\emptyset) = \emptyset$ this does not violate monotonicity. The other case is when Δ' also contains colour distributions without unique colours. In that case $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}}$ still fails on the part of Δ' with unique colours, but it is not guaranteed to succeed on the part of Δ' without unique colours (for example, if for four gnomes it contains 1100 as well as 2100 then a will now no longer find out whether it is muddy, but b, c, d will — whereas before, a also would in case 1100). Let Δ'' be the part without unique colours. We then will have for some such Δ'' that $\Delta'' \subseteq \Delta'$ whereas $f_{\varphi(p)}(\Delta'') \not\subseteq f_{\varphi(p)}(\Delta')$. However, with respect to Δ', Δ'' was then already maximal and all extensions of Δ'' will also fail to satisfy the fixpoint formula, and all distributions in the part $\Delta' \setminus \Delta''$ will fail. See the next case for how to treat that.

- For any Δ' that is not permutation invariant, we have that $f_{\varphi(p)}(\Delta') = \emptyset$ because the left conjunct $\bigwedge_{\delta\in\Delta}((\boldsymbol{\delta} \to p) \to \bigwedge_{\iota\in I}[\boldsymbol{\iota}](\boldsymbol{\delta} \to p))$ is now false. So that $\Delta' \not\subseteq f_{\varphi(p)}(\Delta')$. This breaks monotonicity and seems to spell trouble if such a Δ' extends a permutation invariant Δ'' for which $\Delta'' = f_{\varphi(p)}(\Delta'')$, because we now have that $\Delta'' \subseteq \Delta'$ whereas $f_{\varphi(p)}(\Delta'') \not\subseteq f_{\varphi(p)}(\Delta')$. But such a $f_{\varphi(p)}(\Delta'')$ does not contain members of $\Delta'\backslash\Delta''$, and in case $f_{\varphi(p)}(\Delta'')$ was not maximal there must be a permutation invariant Δ''' such that $f_{\varphi(p)}(\Delta'') \subseteq f_{\varphi(p)}(\Delta''')$, where it may even be that $\Delta'' \subset \Delta' \subset \Delta'''$. (As in Section 1.3 wherein $\{111\} \subset \{111,011\} \subset$ (complement of $\{000\}$). In 111 and in the seven-state restriction excluding 000 all finally learn whether they are muddy. But not in $\{111,011\}$.)

It is therefore still the case that the greatest fixpoint we want is the union of all permutation-invariant Δ' such that $\Delta' \subseteq f_{\varphi(p)}(\Delta')$. It exists! Work done.

Arbitrary iteration again Having come this far, why not consider another fixpoint, namely for the arbitrary iteration $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}}$, that so far we took to be the abbreviation of the maximum iteration $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^{<\max} \mathbf{bell}^{\mathbf{last}}$. This is actually straightforward. We can also define $\langle \mathbf{bell}^{\mathbf{nlast}} \rangle^* \mathbf{bell}^{\mathbf{last}}$ as the least fixpoint:

$$\mu q.\mathbf{bell}^{\mathbf{last}} \vee \langle \mathbf{bell}^{\mathbf{nlast}} \rangle q$$

First, it is easy to see that the single occurrence of q in $\mathbf{bell}^{\mathbf{last}} \vee \langle \mathbf{bell}^{\mathbf{nlast}} \rangle q$ remains positive in the multi-agent epistemic logic formula that we obtain by the rewriting axioms of public announcement logic. Second, the existence of an equivalent epistemic formula wherein q is positive in a standard modal μ-calculus therefore guarantees the existence of a least fixpoint by monotonicity. Third, in the **muetzen** model, that is a finite model, the iteration constructing the fixpoint will still be finite, which is an easy fix in any other sense of the word as well. But we can also easily constructively see the monotonicity because the fixpoint is the already computed extension of p of $\nu p.\varphi(p)$, and bottom-up we see that every singleton colour distribution satisfies $\mathbf{bell}^{\mathbf{last}}$.

So we can even formalize **solvable** as follows. Two fixpoints!

$$\mathbf{solvable} \quad := \quad \nu p. \bigwedge_{\delta\in\Delta}((\boldsymbol{\delta} \to p) \to \bigwedge_{\iota\in I}[\boldsymbol{\iota}](\boldsymbol{\delta} \to p)) \wedge \langle p \rangle (\mu q.\mathbf{bell}^{\mathbf{last}} \vee \langle \mathbf{bell}^{\mathbf{nlast}} \rangle q)$$

Again have that the solution of the Woeginger Mützen puzzle is formalized as:

Determine $n \in \mathbb{N}^+$ such that $\mathbf{muetzen}, \delta \models \langle \mathbf{solvable} \rangle \langle \mathbf{bell}^{\mathbf{nlast}} \rangle^{n-1} \mathbf{bell}^{\mathbf{last}}$.

So, merely asking once more: what is n? What a wonderful and creative hat problem Mützen is.

Acknowledgements I kindly acknowledge discussions with and encouragement or comments from Marta Bílková, Tim French, Malvin Gattinger, Barteld Kooi, Alexander Kurz, Roman Kuznets, Clara Lerouvillois, Yanjun Li, and Thomas Ågotnes. In particular, an early stage of this research involved much interactions and discussions with Barteld Kooi during my visit to Groningen in April 2024. I also very much thank the TARK reviewers for their comments. One reviewer found an error in one of the variations of Mützen presented: you are correct, pink and yellow only leave at ring 9, not at ring 8. I acknowledge support from the *Wolfgang Pauli Institute* (WPI) in Vienna where I was a Fellow during the completion of this research.

References

[1] A. Baltag, N. Bezhanishvili & D. Fernández-Duque (2022): *The Topology of Surprise*. In G. Kern-Isberner, G. Lakemeyer & T. Meyer, editors: *Proc. of 19th KR*, doi:10.24963/kr.2022/4. Available at https://proceedings.kr.org/2022/4/.

[2] J. Barwise (1981): *Scenes and other situations*. *Journal of Philosophy* 78(7), pp. 369–397, doi:10.2307/2026481.

[3] V. Belle, T. Bolander, A. Herzig & B. Nebel (2023): *Epistemic planning: Perspectives on the special issue*. *Artificial Intelligence* 316, p. 103842, doi:10.1016/j.artint.2022.103842.

[4] J. van Benthem & D. Ikegami (2008): *Modal Fixed-Point Logic and Changing Models*. In A. Avron, N. Dershowitz & A. Rabinovich, editors: *Pillars of Computer Science*, Springer, pp. 146–165, doi:10.1007/978-3-540-78127-1_9. LNCS 4800.

[5] L. Bozzelli, H. van Ditmarsch, T. French, J. Hales & S. Pinchinat (2014): *Refinement Modal Logic*. *Information and Computation* 239, pp. 303–339, doi:10.1016/j.ic.2014.07.013.

[6] H. van Ditmarsch, J.Y. Halpern, W. van der Hoek & B. Kooi, editors (2015): *Handbook of epistemic logic*. College Publications.

[7] H. van Ditmarsch, A. Herzig & T. de Lima (2012): *Public announcements, public assignments and the complexity of their logic*. *Journal of Applied Non-Classical Logics* 22(3), pp. 249–273, doi:10.1080/11663081.2012.705964.

[8] H. van Ditmarsch, A. Herzig & T. De Lima (2011): *From Situation Calculus to Dynamic Epistemic Logic*. *Journal of Logic and Computation* 21(2), pp. 179–204, doi:10.1093/logcom/exq024.

[9] H. van Ditmarsch, W. van der Hoek & B. Kooi (2005): *Dynamic Epistemic Logic with Assignment*. In: *Proc. of 4th AAMAS*, ACM, pp. 141–148, doi:10.1007/978-1-4020-5839-4_2.

[10] H. van Ditmarsch, W. van der Hoek & B. Kooi (2007): *Dynamic Epistemic Logic*. *Synthese Library* 337, Springer, doi:10.1007/978-1-4020-5839-4.

[11] H. van Ditmarsch & B. Kooi (2006): *The Secret of My Success*. *Synthese* 151, pp. 201–232, doi:10.1007/s11229-005-3384-9.

[12] H. van Ditmarsch & B. Kooi (2008): *Semantic Results for Ontic and Epistemic Change*. In: *Proc. of 7th LOFT*, Texts in Logic and Games 3, Amsterdam University Press, pp. 87–117.

[13] H. van Ditmarsch & B. Kooi (2015): *One Hundred Prisoners and a Light Bulb*. Copernicus, doi:10.1007/978-3-319-16694-0.

[14] H. van Ditmarsch, J. Ruan & R. Verbrugge (2007): *Sum and Product in Dynamic Epistemic Logic*. *Journal of Logic and Computation* 18(4), pp. 563–588, doi:10.1093/logcom/exm081.

[15] H. van Ditmarsch, W. van der Hoek & B. Kooi (2005): *Public Announcements and Belief Expansion*. In R. Schmidt, I. Pratt-Hartmann, M. Reynolds & H. Wansing, editors: *Proc. of 5th AiML*, King's College Publications, pp. 335–346.

[16] C. Esmangart et E. Johanneau (1823): *Oeuvres de Rabelais, ouvrage posthume*. 1, Dalibon, Paris.

[17] P. van Emde Boas, J. Groenendijk & M. Stokhof (1984): *The Conway Paradox: Its Solution in an Epistemic Framework*. In: *Truth, Interpretation and Information*, Foris Publications, Dordrecht, pp. 159–182, doi:10.1515/9783110867602.159.

[18] H. Freudenthal (1969): *Problem No. 223 (formulation of the Sum and Product problem)*. *Nieuw Archief voor Wiskunde* 3(17), p. 152.

[19] G. Gamow & M. Stern (1958): *Puzzle-Math*. Macmillan, London.

[20] J. Halpern & Y. Moses (1986): *Taken by surprise: The paradox of the surprise test revisited*. *J. Philos. Log.* 15(3), pp. 281–304, doi:10.1007/BF00248573.

[21] J. Hintikka (1962): *Knowledge and Belief*. Cornell University Press.

[22] D. Kozen (1982): *Results on the Propositional μ-Calculus*. In M. Nielsen & E. Meineche Schmidt, editors: *Automata, Languages and Programming, 9th Colloquium, Proceedings*, LNCS 140, Springer, pp. 348–359, doi:10.1007/BFB0012782.

[23] M. Kraitchik (1942): *Mathematical Recreations*. W. W. Norton, New York.

[24] Y. Li, J. Ren & T. Ågotnes (2025): *The Surprise Exam in Full Modal Fixed-Point Logic*. To appear in Proc. of 6th CLAR (International Conference on Logic and Argumentation).

[25] J.E. Littlewood (1953): *A Mathematician's Miscellany*. Methuen and Company.

[26] J. McCarthy (1990): *Formalization of two puzzles involving knowledge*. In V. Lifschitz, editor: *Formalizing Common Sense : Papers by John McCarthy*, Ablex Publishing Corporation, Norwood, N.J.

[27] J.S. Miller & L.S. Moss (2005): *The Undecidability of Iterated Modal Relativization*. *Studia Logica* 79(3), pp. 373–407, doi:10.1007/s11225-005-3612-9.

[28] G.E. Moore (1942): *A reply to my critics*. In P.A. Schilpp, editor: *The Philosophy of G.E. Moore*, Northwestern University, Evanston, pp. 535–677.

[29] Y.O. Moses, D. Dolev & J.Y. Halpern (1986): *Cheating husbands and other stories: a case study in knowledge, action, and communication*. *Distributed Computing* 1(3), pp. 167–176, doi:10.1007/BF01661170.

[30] R. Parikh (1992): *Finite and Infinite Dialogues*. In Y. Moschovakis, editor: *Logic from Computer Science*, Springer, pp. 481–497, doi:10.1007/978-1-4612-2822-6_18.

[31] J.A. Plaza (1989): *Logics of Public Communications*. In: *Proc. of the 4th ISMIS*, Oak Ridge National Laboratory, pp. 201–216.

[32] R.M. Smullyan (1982): *Lady or the Tiger? And Other Logic Puzzles Including a Mathematical Novel That Features Godel's Great Discovery*. Random House, New York.

[33] G. van Tilburg (1956): *Doe wel en zie niet om*. *Katholieke Illustratie* 90(32), p. 47. Breinbrouwsel 137.

[34] G. Woeginger (2013): *Mützen*. Available at http://www.mathekalender.de/wp/de/kalender/challenge-archive/. Problem no. 10 in the 2013 MATH+ Advent Calender.

[35] H. Xing (2024): *The extension of modal μ-calculus based on covariant-contravariant refinement*. Unpublished manuscript.

Appendix: Gerhard Woeginger's Hat Problem [34]

Der Weihnachtsmann hat 126 Intelligenzwichtel zu einem gemütlichen Nachmittag mit Kaffee und Kuchen eingeladen. Als die Wichtel den Saal betreten, bekommt jeder von ihnen hinterrücks eine neue Wichtelmütze auf den Kopf gesetzt. Das geht blitzschnell, sodass keiner von ihnen die Farbe der eigenen Mütze zu sehen kriegt. Der Weihnachtsmann eröffnet das Treffen mit einer kurzen Rede.

> *Meine lieben Intelligenzwichtel! Wir wollen diesen Nachmittag mit einem kleinen Denkspiel beginnen. Keiner von euch kennt die Farbe der eigenen Mütze, und jeder von euch kann die Mützen aller anderen 125 Wichtel sehen. Ziel dieses Spieles ist es, die Farbe der eigenen Mütze möglichst schnell und durch reines Nachdenken herauszufinden. Ich werde nun im Fünf-Minuten-Takt mit meiner grossen Weihnachtsglocke läuten. Wenn einer die eigene Mützenfarbe herausgefunden hat, so muss er beim nächsten Läuten sofort den Saal verlassen. Im Nebenzimmer bekommt er dann eine Tasse Kaffee und ein grosses Stück Sachertorte serviert.*

Der Weihnachtsmann will gerade zur Glocke gehen, als dem Wichtel Atto eine wichtige Frage einfällt:

> *Ja ist es denn wirklich für jeden von uns möglich, seine Mützenfarbe durch logisches Denken zu bestimmen? Wenn zum Beispiel jeder von uns eine andere Mützenfarbe hätte, dann könnte*

> *wohl niemand seine Farbe durch Denken herausfinden. Dann wäre das Spiel für uns doch nicht zu gewinnen!*

Der Weihnachtsmann entgegnet ihm ein wenig unwirsch:

> *Hätte, könnte, wäre!!! Natürlich kann jeder von euch dieses Spiel gewinnen! Ich habe die Mützenfarben sehr sorgfältig ausgewählt, sodass jeder von euch tatsächlich seine Farbe im Laufe des Spiels durch Denken herleiten kann.*

Und dann beginnen die Wichtel zu denken. Und der Weihnachtsmann beginnt zu läuten.

- Beim ersten Läuten verlassen Atto und neun andere Wichtel den Saal.
- Beim zweiten Läuten gehen alle Wichtel mit butterblumengelben, dotterblumengelben, schlüsselblumengelben und sonnenblumengelben Mützen aus dem Saal.
- Beim dritten Läuten gehen alle Wichtel mit karmesinroten Mützen, beim vierten Läuten alle mit kaktusgrünen Mützen, beim fünften Läuten alle mit aquamarinblauen Mützen, beim sechsten Läuten alle mit goldorangen Mützen, beim siebten Läuten alle mit bernsteinbraunen Mützen und beim achten Läuten gehen alle mit muschelgrauen Mützen.
- Beim neunten, zehnten, elften und zwölften Läuten verlässt niemand den Saal.
- Beim dreizehnten Läuten gehen alle Wichtel mit blütenweissen und alle Wichtel mit ebenholzschwarzen Mützen.

Und so geht es weiter. Beim N-ten Läuten des Weihnachtsmanns verlässt schließlich die letzte Wichtelgruppe den Saal. Ganze sieben Mal hat der Weihnachtsmann zwischendurch geläutet, ohne dass jemand aus dem Saal gegangen wäre (und bei diesen sieben Malen sind das neunte, zehnte, elfte und zwölfte Läuten bereits mitgezählt). Unsere Frage lautet nun: Wie groß ist N? Antwortmöglichkeiten:

1. $N = 17$	3. $N = 19$	5. $N = 21$	7. $N = 23$	9. $N = 25$
2. $N = 18$	4. $N = 20$	6. $N = 22$	8. $N = 24$	10. $N = 26$

Appendix: Gerhard Woeginger's Hat Problem [34] — English

Santa Claus invited 126 smartgnomes to a cozy afternoon with coffee and cake. When the gnomes enter the hall, each of them gets a new gnome hat placed on their head from behind. This happens at lightning speed, so that none of them can see the colour of their own hat. Santa opens the meeting with a short speech.

> My dear smartgnomes! We want to start this afternoon with a little brain teaser. None of you knows the colour of your own hat, and each of you can see the hats of all the other 125 gnomes. The aim of this game is to find out the colour of your own hat as quickly as possible and through pure thinking. I will now ring my big Christmas bell every five minutes. Once someone has found out their own hat colour, he must immediately leave the hall at the next ring of the bell. In the next room he is then served a cup of coffee and a large piece of Sachertorte.

Santa is just about to ring the bell when gnome Atto comes up with an important question:

> Is it really possible for each of us to determine the colour of our hat through logical thinking? For example, if each of us had a different colour of hat, then no one would be able to figure out what colour it is by mere deduction. Then we couldn't win the game!

Santa answers him a little gruffly:

> Would, could, were!!! Of course, each of you can win this game! I chose the hat colours very carefully so that each of you can actually determine their colour through thinking during the game.

And then the gnomes start thinking. And Santa starts ringing the bell.

- At the first ring of the bell, Atto and nine other gnomes leave the hall.
- At the second ring, all the gnomes walk out of the hall wearing buttercup yellow, kingcup yellow, primrose yellow and sunflower yellow hats.
- At the third ring all the gnomes leave with crimson hats, at the fourth ring all with cactus green hats, at the fifth ring all with aquamarine blue hats, at the sixth ring all with orange hats, at the seventh ring all with amber brown hats and at the eighth ring of the bell all leave with shell-gray hats.
- At the ninth, tenth, eleventh and twelfth rings no one leaves the hall.
- At the thirteenth ring, all the gnomes leave with blossom-white hats and all the gnomes with ebony-black hats.

And so it goes on. At the Nth ring of Santa, the last group of gnomes finally leaves the hall. Santa rang a total of seven times without anyone leaving the hall (and of these seven times, the ninth, tenth, eleventh and twelfth rings are already counted). Our question now is: What is N? Possible answers:

1. $N = 17$	3. $N = 19$	5. $N = 21$	7. $N = 23$	9. $N = 25$	
2. $N = 18$	4. $N = 20$	6. $N = 22$	8. $N = 24$	10. $N = 26$	[34]