On Quantified Modal Theorem Proving for Modeling Ethics

Naveen Sundar Govindarajulu
Rensselaer AI & Reasoning Lab
Troy, New York
naveensundarg@gmail.com

Selmer Bringsjord
Rensselaer AI & Reasoning Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York
selmer.bringsjord@gmail.com

Matthew Peveler
Rensselaer AI & Reasoning Lab
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York
matt.peveler@gmail.com

In the last decade, formal logics have been used to model a wide range of ethical theories and principles with the goal of using these models within autonomous systems. Logics for modeling ethical theories, and their automated reasoners, have requirements that are different from modal logics used for other purposes, e.g. for temporal reasoning. Meeting these requirements necessitates investigation of new approaches for proof automation. Particularly, a quantified modal logic, the **deontic cognitive event calculus** ($\mathcal{DCEC}$), has been used to model various versions of the doctrine of double effect, *akrasia*, and virtue ethics. Using a fragment of $\mathcal{DCEC}$, we outline these distinct characteristics and present a sketches of an algorithm that can help with some aspects proof automation for $\mathcal{DCEC}$.

1 Introduction

Modal logics have been used for decades to model and study a diverse set of subjects — e.g. temporal reasoning, multi-agent systems, linguistic content and phenomena, and game theory [1, Part 4]. While deontic modal logics have been used to study ethical principles, it is only recently that such logics have been considered in a rigorous manner [3] with the goal of either using them in a computational system or using such a logic to analyze computational systems.

For example, a quantified modal logic, the **deontic cognitive event calculus** ($\mathcal{DCEC}$), has been used recently to model various versions of the *Doctrine of Double Effect*, *akrasia*, and *virtue ethics* [14, 12, 15, 13, 7]. These ethical principles and theories have a unique set of characteristics when compared with other domains, e.g. with temporal reasoning, in which modal logics have been used.[1] This implies that logics for modeling ethical theories have requirements that are different than those for modal logics used for other purposes. These requirements dictate investigation of new approaches for proof automation. We present a central set of these requirements in this paper. Using a fragment $\mathscr{C}^1$ of $\mathcal{DCEC}$, we also present an algorithm that can help enable proof automation which partially satisfies these requirements.

[1] A note on the terms "ethical principles" and "ethical theories." An ethical theory is generally broader and more fundamental than an ethical principle. An ethical principle is ultimately a declarative statement usually cast under one or more ethical theories. E.g., the principle that one ought always to act with the intention to maximize utility for everyone would fall under the ethical theory known as *utilitarianism*. For a classic presentation of the main ethical theories and their key principles, see [10].

M. Suda, S. Winkler (Eds.): Second International Workshop
on Automated Reasoning: Challenges, Applications,
Directions, Exemplary Achievements (ARCADE 2019)
EPTCS 311, 2019, pp. 43–49, doi:10.4204/EPTCS.311.7

2 Requirements for Modeling Ethical Theories

To illustrate the unique characteristics required for modeling ethical theories and principle, we use the **Doctrine of Double Effect** ($\mathcal{DDE}$) augmented to handle self sacrifice. $\mathcal{DDE}$ is an ethical principle that can account for human judgment in **moral dilemmas**: situations in which all available actions have both significantly good *and* significantly bad consequences. According to $\mathcal{DDE}$, an action α in such a situation is permissible *iff*

> "(1) it is morally neutral; (2) the net good consequences outweigh the bad consequences by a large amount; and (3) some of the good consequences are intended, while none of the bad consequences are. [11]"

A formalization of $\mathcal{DDE}$ is presented in [11]. While $\mathcal{DDE}$ has some empirical support [9], it cannot account for instances of self-sacrifice. To handle self-sacrifice, an augmented version, $\mathcal{DDE}^*$, is presented and formalized in [13]. We now present an informal version of $\mathcal{DDE}^*$ to illustrate the requirements. We assume there is an ethical hierarchy of actions (e.g. *forbidden*, *morally neutral*, *obligatory*); see [2]. We also assume that we have a utility or goodness function for states of the world or effects. For an autonomous agent a, an action α in a situation σ at time t is said to be $\mathcal{DDE}^*$-compliant *iff*:

$\mathbf{C}_1$ At the time of the action, the agent a executing the action believes that the action is not forbidden (where, again, we assume an ethical hierarchy such as the one given by Bringsjord [2], and require that the action be morally neutral or above morally neutral in such a hierarchy);

$\mathbf{C}_2$ At the time of the action, the agent a believes that the net utility or goodness of the action is greater than some positive amount γ;

$\mathbf{C}_{3a}$ At the time of the action, the agent a performing the action intends only the good effects;

$\mathbf{C}_{3b}$ At the time of the action, the agent a does not intend any of the bad effects;

$\mathbf{C}_4$ the bad effects are not used by a as a means to obtain the good effects [unless a knows that the bad effects are confined to only a itself]; and

$\mathbf{C}_5$ if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action; that is, the action is unavoidable.

With $\mathcal{DDE}^*$ as the background, we outline the following requirements that are necessary in modeling not only $\mathcal{DDE}^*$ but also other ethical theories and principles, such as virtue ethics and *akrasia*. We split the requirements into two parts: requirements for the logic, and additional requirements for the reasoner.

Requirements for the Logic

$\mathbf{R}_1$ Multiple Modalities: Ethical principles have statements that take into account an agent's beliefs, intentions, obligations, etc. Any acceptable logic should be able to handle this.

$\mathbf{R}_2$ Time-Indexed Modalities: Intentions and beliefs at the time of an action matter rather than intentions and beliefs at other times.

$\mathbf{R}_3$ *De se* Agent Modalities Agent-indexed modalities are common in BDI (belief/desire/intention) logics [18], but $\mathbf{C}_4$ requires self beliefs known as *de se* beliefs. This requires modalities indexed by *de se* agents. This is needed to model statements such as *"a believes that a herself believes that ..."*. For more details on *de se* beliefs, please see [13] and [5].

$\mathbf{R}_4$ Quantifiers: Quantifiers are needed to handle comparisons between actions and for ordering actions by their consequences.

While the above core requriments are needed for $\mathcal{DDE}^*$, other features, such as the ability to represent uncertainty and counterfactuals, may be needed for some ethical theories. We omit these requirements from the core list above as there has not been as much discussion around these features. In addition to handling the above requirements, any reasoner for the logic should have the following capabilities:

Additional Requirements for the Reasoner

1. **Builtin Theories**: Handling of simple arithmetic and causation. This is required for efficiently computing consequences, and causes of actions.
2. **Justifications and Explanations**: Any reasoning system in an ethically charged scenario should be able to explain and present its reasoning in a verifiable and understandable manner.
3. **Answer Finding**: The reasoner should not only be used for proving that an action is ethical but should also be capable of finding the most ethical action in a given situation.

3 A Sparse Calculus $\mathscr{C}^1$

$\mathscr{C}^1$ is a straightforward modal extension of first-order logic that satisfies $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_4$. We have a modal operator **B** for belief, an operator **O** for obligation, and **G** to denote goals. The syntax and inference schemata of the system are shown below. Assume that we have a first-order alphabet augmented with a fixed finite set of symbols for agents $Ag = \{a_1, \ldots, a_n\}$ and a set of totally ordered symbols for time $T = \langle t_0, \ldots, t_n, \ldots \rangle$. Sometimes we use a for a_i and t for t_i. ϕ is a meta-variable for formulae, and A is any first-order atomic symbol. Given this, the grammar for wffs of $\mathscr{C}^1$ follows.

$$s_i ::= \text{standard first-order terms}$$

$$\phi ::= \begin{cases} A(s_1, \ldots s_n) \\ \neg\phi \mid \phi \vee \psi \mid \forall x.\phi \\ \mathbf{B}(a,t,\phi) \mid \mathbf{O}(a,t,\phi,\psi) \mid \mathbf{G}(a,t,\psi) \end{cases}$$

$\mathbf{B}(a,t,\phi)$ states that a believes at time t that ϕ holds. $\mathbf{O}(a,t,\phi,\psi)$ states that a ought to ψ at time t that if ϕ holds. $\mathbf{G}(a,t,\psi)$ states that a has as a goal ψ at time t.

Inference System We have three inference schemata: $\{I_R, I_B, I_O\}$, shown in Table 1. $\mathsf{unify}(a,b)$ denotes the most general first-order unifier of a and b. First-order reasoning is performed through I_R, which is just first-order resolution. Reasoning with beliefs is done with I_B. Beliefs propagate forward in time. Reasoning with obligations is handled with I_O: If an agent believes it has an obligation ψ when ϕ, and believes that ϕ, then it has a goal ψ.

Description	Inference Scheme
I_R	$\dfrac{\phi_1 \vee \ldots \chi \ldots \vee \phi_n \qquad \psi_1 \vee \ldots \neg\chi' \ldots \vee \psi_m \quad \text{where } \theta = \mathsf{unify}(\chi,\chi')}{(\phi_1 \vee \ldots \vee \ldots \phi_n \vee \psi_1 \vee \ldots \vee \psi_m)\theta}$ $[I_{\mathbf{Res}}]$
I_B	$\dfrac{\mathbf{B}(a,t_1,\phi_1 \vee \ldots \chi \ldots \vee \phi_n) \qquad \mathbf{B}(a,t_2,\psi_1 \vee \ldots \neg\chi' \ldots \vee \psi_m) \quad \text{where } t \geq t_1,t_2;\ \text{and } \theta = \mathsf{unify}(\chi,\chi')}{\mathbf{B}\big(a,t(\phi_1 \vee \ldots \vee \ldots \phi_n \vee \psi_1 \vee \ldots \vee \psi_m)\theta\big)}$ $[I_{\mathbf{B}}]$
I_O	$\dfrac{\mathbf{B}(a,\phi) \qquad \mathbf{B}(a,\mathbf{O}(a,\phi,\psi))}{\mathbf{G}(a,\psi)}$ $[I_{\mathbf{O}}]$

Proof from Γ to ϕ: A proof Π^{Γ}_{ϕ} from Γ to ϕ consists of a sequence of formulae $\phi_1, \phi_2, \ldots, \phi_n$ such that (i) $\phi_n \equiv \phi$; and (ii) for all $1 \leq i < n$, ϕ_i is derived from $\{\phi_j | j < i\}$ using $I_{\mathbf{R}}$, $I_{\mathbf{B}}$, or $I_{\mathbf{O}}$. $\Gamma \vdash \phi$ denotes that there is a proof Π^{Γ}_{ϕ} from Γ to ϕ.

4 An Algorithm Sketch

We now present an algorithm for handling the proof system for $\mathscr{C}^1$. Our goal is to leverage advances in first-order theorem proving to build the relevant reasoner. There are two straightforward but flawed ways this can be done. In the first approach, modal operators are simply represented by first-order predicates. This approach is the fastest but can quickly lead to well-known inconsistencies, as demonstrated in [4]. In the second approach, the entire proof theory is implemented intricately in first-order logic, and the reasoning is carried out within first-order logic. Here, the first-order theorem prover simply functions as a declarative programming system. This approach, while accurate, can be inefficient.

Our algorithm is based on a technique we term **shadowing**. At at high-level, we alternate between calling a first-order theorem prover and applying modal inference schemata. When we call the first-order prover, all modal atoms are converted into propositional atoms (i.e. the former are shadowed), to prevent substitution into modal contexts. This approach achieves speed without sacrificing consistency. The algorithm is briefly described below.

First we define the syntactic operation of **atomizing** a formula, denoted by A. Given any arbitrary formula ϕ, $\mathsf{A}_{[\phi]}$ is a unique atomic (propositional) symbol. Next, we define the **level** of a formula: $\mathsf{level} : \mathsf{Boolean} \rightarrow \mathbb{N}$.

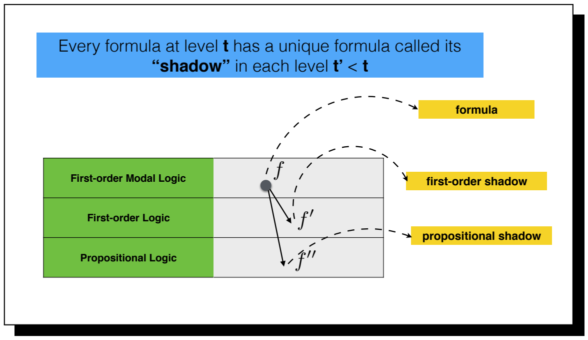
$$\mathsf{level}(\phi) = \begin{cases} 0; \phi \text{ is purely propositional formulae; e.g. } \mathit{Rainy} \\ 1; \phi \text{ has first-order predicates or quantifiers e.g. } \mathit{Sleepy}(\mathit{jack}) \\ 2; \phi \text{ has modal formulae e.g. } \mathbf{K}(a, t, \mathit{Sleepy}(\mathit{jack})) \end{cases}$$

Given the above definition, we can define the operation of **shadowing** a formula to a level. See Figures 1a and 1b.

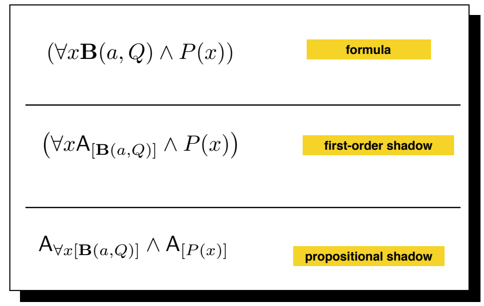
Shadowing

To shadow a formula χ to a level l, replace all sub-formulae χ' in χ such that $\mathsf{level}(\chi') > l$ with $\mathsf{A}_{[\chi']}$ simultaneously. We denote this by $\mathsf{S}[\phi, l]$.
For a set Γ, the operation of shadowing all members in the set is simply denoted by $\mathsf{S}[\Gamma, l]$.

Assume we have access to a first-order prover $\mathbf{P}$. For a set of pure first-order formulae Γ and a first-order ϕ, $\mathbf{P}(\Gamma, \phi)$ gives us a proof of $\Gamma \vdash \phi$ if such a first-order proof exists; otherwise **fail** is returned. See the algorithm sketch given below for a reasoner for $\mathscr{C}^1$:



(a) Shadowing a formulae to different levels (Overview)



(b) Shadowing a formulae to different levels (Examples)

Input: Input Formulae Γ, Goal Formula ϕ
Output: A proof of $\Gamma \vdash \phi$ if such a proof exists, otherwise **fail**
initialization;
while *goal not reached* **do**
  answer $= \mathbf{P}(\mathsf{S}[\Gamma, 1], \mathsf{S}[\phi, 1])$;
  if *answer* $\neq$ **fail** **then**
    return *answer* ;
  else
    $\Gamma' \longleftarrow$ **expand** Γ by using any applicable modal inference schemata;
    $\Gamma' \longleftarrow$ **expand** Γ' by recursively reasoning forward in all modal contexts;
    if $\Gamma' = \Gamma$ **then**
      `/* The input cannot be expanded further */`
      return **fail**
    else
      set $\Gamma \longleftarrow \Gamma'$
    end
  end
end

The algorithm alternates between applying $\{I_R\}$ and $\{I_B, I_O\}$. The algorithm is instantiated recursively to handle nested first-order reasoning within modal contexts as required in I_B.

5 Implementation

The reasoner is available as an open-source Java library [16]. A lightweight Python interface is available for quick prototyping and experimentation; see Figure 2. For the first-order prover, we use SNARK, due to its facilities for extension with procedural attachments and rewrite systems [17]. In addition, SNARK comes with theories for reasoning about simple arithmetic, lists, etc. For future work, we shall investigate and pursue integration with other theorem provers. Our prover is also integrated within the HyperSlate proof assistant, a modern extension of the Slate proof assistant [8]; see Figure 3 for an example.

6 Conclusion and Future Work

We have presented requirements that modal logics for modeling ethical theories should satisfy. A reasoning algorithm that can satisfy some of the requirements was presented. Future work involves extending

```python
from interface import prove

assumptions = [
    "(Believes! john happy)",
    "(Believes! john smiling)" ]

goal = "(Believes! john (and happy smiling))"

prove(assumptions, goal)
```

Figure 2: Use of the $\mathcal{DCEC}$ reasoner from Python.

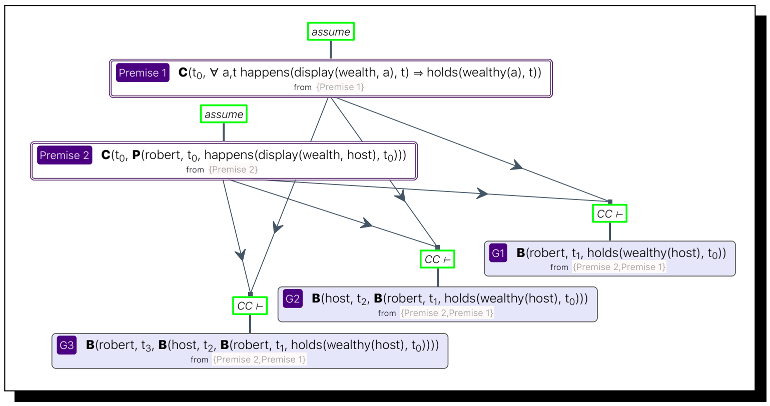


Figure 3: Use of the $\mathcal{DCEC}$ theorem prover within the HyperSlate workspace. Example from [6].

the reasoner to satisfy the other remaining requirements and proving that algorithm are sound and complete with respect to a core inference system. As there are no similar reasoning systems for $\mathcal{DCEC}$, direct comparison with other modal logic reasoners is not possible, but we plan to isolate fragments of $\mathcal{DCEC}$ that can enable benchmarks and comparisons with reasoners for other similar logics..

References

[1] Patrick Blackburn, Johan F. A. K. van Benthem & Frank Wolter (2006): *Handbook of Modal Logic*. Elsevier Science Inc., New York, NY, USA.

[2] Selmer Bringsjord (2017): *A 21st-Century Ethical Hierarchy for Robots and Persons: $\mathcal{EH}$*. In: *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, 84, Springer, Lisbon, Portugal, p. 47, doi:10.1007/978-3-642-31674-6_11.

[3] Selmer Bringsjord, Konstantine Arkoudas & Paul Bello (2006): *Toward a General Logicist Methodology for Engineering Ethically Correct Robots*. *IEEE Intelligent Systems* 21(4), pp. 38–44, doi:10.1109/MIS.2006.82. Available at `http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf`.

[4] Selmer Bringsjord & Naveen Sundar Govindarajulu (2012): *Given the Web, What is Intelligence, Really? Metaphilosophy* 43(4), pp. 361–532, doi:10.1111/j.1467-9973.2012.01760.x. Available at `http://kryten.mm.rpi.edu/SB\NSG\Real\Intelligence\040912.pdf`. This URL is to a preprint of the paper.

[5] Selmer Bringsjord & Naveen Sundar Govindarajulu (2013): *Toward a Modern Geography of Minds, Machines, and Math*. In Vincent C. Müller, editor: *Philosophy and Theory of Artificial Intelligence*, *Studies in Applied Philosophy, Epistemology and Rational Ethics* 5, Springer, New York, NY, pp.

151–165, doi:10.1007/978-3-642-31674-6_11. Available at http://www.springerlink.com/content/hg712w4l23523xw5.

[6] Selmer Bringsjord & Naveen Sundar Govindarajulu (2019): *Learning Ex Nihilo*. *CoRR* abs/1903.03515. Available at http://arxiv.org/abs/1903.03515.

[7] Selmer Bringsjord, Naveen Sundar Govindarajulu, Daniel Thero & Mei Si (2014): *Akratic Robots and the Computational Logic Thereof*. In: *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, Chicago, IL, pp. 22–29, doi:10.1109/ETHICS.2014.6893436. IEEE Catalog Number: CFP14ETI-POD.

[8] Selmer Bringsjord, Joshua Taylor, Andrew Shilliday, Micah Clark & Konstantine Arkoudas (2008): *Slate: An Argument-Centered Intelligent Assistant to Human Reasoners*. In Floriana Grasso, Nancy Green, Rodger Kibble & Chris Reed, editors: *Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)*, University of Patras, Patras, Greece, pp. 1–10. Available at http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf.

[9] Fiery Cushman, Liane Young & Marc Hauser (2006): *The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm*. *Psychological Science* 17(12), pp. 1082–1089, doi:10.1111/j.1467-9280.2006.01834.x.

[10] Fred Feldman (1978): *Introductory Ethics*. Prentice-Hall, Englewood Cliffs, NJ.

[11] Naveen Sundar Govindarajulu & Selmer Bringsjord (2017): *On Automating the Doctrine of Double Effect*. In Carles Sierra, editor: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, Melbourne, Australia, pp. 4722–4730, doi:10.24963/ijcai.2017/658. Preprint available at this url: https://arxiv.org/abs/1703.08922.

[12] Naveen Sundar Govindarajulu, Selmer Bringsjord & Rikhiya Ghosh (2018): *One Formalization of Virtue Ethics via Learning*. *CoRR* abs/1805.07797. Available at http://arxiv.org/abs/1805.07797. To be Presented at the 2018 International Association for Computing and Philosophy (IACAP) - Annual Meeting, June 21-23, 2018, Warsaw, Poland.

[13] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh & Matthew Peveler (2019): *Beyond the Doctrine of Double Effect: A Formal Model of True Self-sacrifice*, pp. 39–54. Springer International Publishing, Cham, doi:10.1007/978-3-030-12524-0_5.

[14] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh & Vasanth Sarathy (2019): *Toward the Engineering of Virtuous Machines*. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, ACM, New York, NY, USA, pp. 29–35, doi:10.1145/3306618.3314256.

[15] N.S. Govindarajulu & S. Bringsjord (2017): *On Automating the Doctrine of Double Effect*. In C. Sierra, editor: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, International Joint Conferences on Artificial Intelligence, pp. 4722–4730, doi:10.24963/ijcai.2017/658.

[16] Naveen Sundar Govindarjaulu (2017): *ShadowProver: A Fast and Exact Prover for Higher-order Modal Logic*, doi:10.5281/zenodo.1451808. Available at https://doi.org/10.5281/zenodo.1451808. URL: https://github.com/naveensundarg/prover.

[17] Mark Stickel, Richard Waldinger, Michael Lowry, Thomas Pressburger & Ian Underwood (1994): *Deductive Composition of Astronomical Software From Subroutine Libraries*. In: *Proceedings of the Twelfth International Conference on Automated Deduction (CADE–12)*, Nancy, France, pp. 341–355, doi:10.1007/3-540-58156-1_24. Available at http://www.ai.sri.com/~stickel/snark.html.

[18] M. Wooldridge (2002): *An Introduction to Multi Agent Systems*. MIT Press, Cambridge MA.