Toward Proof-Theoretic Semantics for the Deontic Cognitive Event Calculus[*]

(abstract)

Naveen Sundar Govindarajulu
naveensundarg@gmail.com

Selmer Bringsjord
selmer.bringsjord@gmail.com

Atriya Sen
atriya@atriyasen.com

Jan 7 2019 2230 NY

[*] We are grateful for support from the Office of Naval Research for our application of cognitive calculi to Arrow's Impossibility Theorem and judgement aggregation, and to the challenge of making morally correct machines.

Overview

We present some early work devoted to developing a robust, formal, proof-theoretic semantics for the **deontic cognitive event calculus** ($\mathcal{DCEC}$), a quantified multi-operator modal logic that has been used to model ethical theories, principles, and concepts, and enable their dynamic use. $\mathcal{DCEC}$ belongs to what we term the *cognitive calculi* family of logics. The first member of this family, the **cognitive event calculus** ($\mathcal{CEC}$), was introduced rather long ago by Arkoudas and Bringsjord [1] for their modeling of the *false-belief task*.[1] $\mathcal{DCEC}$ has been used to specifically formalize and automate richly intensional, ethical concept of *akrasia* (≈ an agent's succumbing to temptation to violate moral principles affirmed by that very agent) [4], the Doctrine of Double Effect [7], and elements of virtue ethics [10]. An extension of $\mathcal{DCEC}$ able to handle self-awareness (by modeling *de-se* statements) has been used to model ethical principles that involve self-sacrifice [11]. A smaller member of the family, $\mu\mathcal{C}$, **microcalculus**, has the ability to handle uncertainty [8]. During the workshop, we will present our ongoing construction of a proof-theoretic semantics for $\mathcal{DCEC}$. This builds upon our proof-theoretic semantics for $\mathcal{C}^0$, a fragment of $\mathcal{DCEC}$ [9]. One immediate application of our framework is to *social choice theory*, specifically the automation of a version of Arrow's Impossibility Theorem [12] applied to judgment aggregation, when judgments are about competing recommendations from multiple agents. A brief introduction to $\mathcal{DCEC}$ immediately follows.

Syntax (encapsulated)

$\mathcal{DCEC}$ is a sorted multi-operator quantified modal logic subsuming the first-order sorted *event calculus* [13]. Among the built-in sorts of $\mathcal{DCEC}$, the Agent, Action, and $\mathsf{ActionType}$ ones are not native to the event calculus. The modal operators present in the calculus include the familiar ones for knowledge $\mathbf{K}$, belief $\mathbf{B}$, desire $\mathbf{D}$, and intention $\mathbf{I}$. The general format of an intensional operator is $\Omega\,(a, t, \phi)$, which says that agent a Ω's at time t the proposition ϕ. Here ϕ can in turn be any arbitrary formula. Also, note the following modal operators: $\mathbf{P}$ for perceiving a state, $\mathbf{C}$ for common knowledge, $\mathbf{S}$ for agent-to-agent communication and public announcements, and finally and crucially, a dyadic deontic operator $\mathbf{O}$ that states when an action is obligatory or (using negation) forbidden for agents.

Syntax

$$
\begin{aligned}
S &::= \mathsf{Agent} \mid \mathsf{ActionType} \mid \mathsf{Action} \sqsubseteq \mathsf{Event} \mid \mathsf{Moment} \mid \mathsf{Fluent} \\
t &::= x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
\phi &::= \begin{cases} q : \mathsf{Formula} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \forall x : \phi(x) \mid \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\phi) \mid \mathbf{I}(a,t,\phi) \\ \mathbf{O}(a,t,\phi,(\neg)\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{cases}
\end{aligned}
$$

Inference Schemata (fragment)

The figure immediately below shows a fragment of the inference schemata for $\mathcal{DCEC}$. $R_{\mathbf{K}}$ and $R_{\mathbf{B}}$ are inference schemata that let us model idealized agents that have their knowledge and belief closed under the $\mathcal{DCEC}$ proof theory. While at least most humans are not deductively closed, and while some cognitive calculi avoid epistemic closure, this feature lets us model more closely how deliberate agents such as organizations and more strategic actors reason. R_{14} dictates how obligations get translated into intentions.

Some Inference Schemata

$$
\frac{\mathbf{K}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 < t_2}{\mathbf{K}(a,t_2,\phi)}\ [R_{\mathbf{K}}] \qquad \frac{\mathbf{B}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 < t_2}{\mathbf{B}(a,t_2,\phi)}\ [R_{\mathbf{B}}]
$$

$$
\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\mathbf{O}(a,t,\phi,\chi))\quad \mathbf{O}(a,t,\phi,\chi)}{\mathbf{K}(a,t,\mathbf{I}(a,t,\chi))}\ [R_{14}]
$$

[1] Cf. another proof-oriented modeling effort targeted at similar cognitive phenomena: [2].

Semantics (sketch only)

Meanings of our modal operators differ from what is available for so-called Belief-Desire-Intention (BDI) logics [14] in many important ways. We do not employ possible-worlds semantics or model-based reasoning in any form, instead opting for a *proof-theoretic* semantics and the associated type of reasoning commonly referred to as *natural deduction* [6, 5]. We of course applaud and appreciate possible-worlds/model-based work in general, but our main motivation is that our logic must be chiefly used to model computing agents whose intelligence for us consists in explicit inference in accordance with a given proof-/argument-theory, and we thus need artifacts to explain, process, confirm . . . such derivations and inferences. This is crucial when logics have to model cognition and computing machinery, as is the case for $\mathcal{DCEC}$. Briefly, in this approach, meanings of modal operators are defined via arbitrary computations over proofs.[2] From the cognition-is-reasoning point of view, truth-functional/model-functional semantics, by our lights, can be shown to be reducible to proof-theoretic semantics, and we give the kernel of our case for this position in the concluding section of the present abstract, below (§"Cognitive . . .").

As an instance of finding meaning in proofs, consider that in $\mathcal{C}^0$, we are given as primitives a set of **agents** a_i^s for each agent term a_i in a given signature. Each agent a_i^s has a set of formulae $\Gamma_{a_i^s}$ denoting its beliefs and a set of agents denoting its beliefs about other agents, given by a partial function mapping agent terms to agents. Semantics for different connectives are then given in terms of proofs based on these primitives. (See [9] for more details and a soundness theorem.) During the workshop, we will show how this model can be extended to include modal operators for knowledge, intention, and, importantly, for obligation.

Cognitive Reducibility Argument for Proof-Theoretic Semantics (enthymematic)

Here is a particularization of our reducibility argument for proof-theoretic semantics, which serves to convey the gist of this argument.

Consider a simple biconditional

$$\beta := (p_i \wedge p_j) \leftrightarrow (p_j \wedge p_i)$$

in the propositional calculus ($\mathcal{L}_{\text{PC}}$). Let ν be any customary truth-value assignment (t.v.a.) of TRUE or FALSE to every relevant propositional atom p_k. We say that any formula in the propositional calculus that's true on every t.v.a. is a **validity**. We can now ask whether β is true on a given t.v.a., and we can also ask if β is a validity. Take the second of these two queries. What is the answer? Of course, the correct answer is an affirmative one. But how does an agent know this? An agent, including us, knows this because it can use the relevant machinery of the formal truth-functional semantics of $\mathcal{L}_{\text{PC}}$ to prove

$$\nu \models \beta \tag{1}$$

and a key part of this machinery is this familiar clause:

$$\nu \models \phi \rightarrow \psi \text{ iff if } \nu \models \phi \text{ then } \nu \models \psi \tag{2}$$

Notice the occurrence in (2) of 'iff' and 'if' and 'then.' These terms are part of what we have called the relevant "machinery." *They are nowhere defined truth-functionally or model-theoretically at all*; they are meta-logical connective. Now, label this machinery '$\mathcal{M}$.' $\mathcal{M}$ includes (2), and also the background-logic proof theory $\Pi_{\mathcal{M}}$ for how 'iff' and 'if' and 'then' are to be reasoned with deductively (e.g. we have on hand *modus ponens*). So the reducibility in question has happened in front of our eyes. To make it even clearer, abbreviate the assertion that the combination of (1) and $\mathcal{M}$ can be used to prove (2) as

$$(\mathcal{M}\ + (2)) \vdash_{\Pi_{\mathcal{M}}} (1) \tag{3}$$

What we have just seen is that in $\mathcal{L}_{\text{PC}}$ getting to (1) reduces to (3). But this generalizes to every single truth-semantic target in $\mathcal{L}_{\text{PC}}$, for every cognizer coming to know that this target holds. In fact, since first-order logic $\mathcal{L}_1$ only augments $\mathcal{L}_{\text{PC}}$ with additional machinery for quantification in the same style, establishing model theoretic assertions of truth in $\mathcal{L}_1$, given what we have just seen, reduces to proof-theoretic semantics.[3]

[2] For a general argument that is the source of skepticism about possible worlds semantics, at least of the set-theoretic variety, see [3].

[3] In considering rich intensional logics (as opposed to $\mathcal{L}_{\text{PC}}$ and $\mathcal{L}_1$, both extensional) and the meaning of their formulae, we likewise look for the

References

[1] K. Arkoudas and S. Bringsjord. Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task. In T.-B. Ho and Z.-H. Zhou, editors, *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, number 5351 in Lecture Notes in Artificial Intelligence (LNAI), pages 17–29. Springer-Verlag, 2008.

[2] T. Braüner. Hybrid-Logical Reasoning in the Smarties and Sally-Anne Tasks. *Journal of Logic, Language and Information*, 23:415–439, 2014.

[3] S. Bringsjord. Are There Set-Theoretic Worlds? *Analysis*, 45(1):64, 1985.

[4] S. Bringsjord, N. S. Govindarajulu, D. Thero, and M. Si. Akratic Robots and the Computational Logic Thereof. In *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, pages 22–29, Chicago, IL, 2014. IEEE Catalog Number: CFP14ETI-POD.

[5] N. Francez and R. Dyckhoff. Proof-theoretic Semantics for a Natural Language Fragment. *Linguistics and Philosophy*, 33:447–477, 2010.

[6] G. Gentzen. Investigations into Logical Deduction. In M. E. Szabo, editor, *The Collected Papers of Gerhard Gentzen*, pages 68–131. North-Holland, Amsterdam, The Netherlands, 1935. This is an English version of the well-known 1935 German version.

[7] N. S. Govindarajulu and S. Bringsjord. On Automating the Doctrine of Double Effect. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4722–4730, Melbourne, Australia, 2017. Preprint available at this url: https://arxiv.org/abs/1703.08922.

[8] N. S. Govindarajulu and S. Bringsjord. Strength Factors: An Uncertainty System for a Quantified Modal Logic, 2017. Presented at Workshop on Logical Foundations for Uncertainty and Machine Learning at IJCAI 2017, Melbourne, Australia.

[9] N. S. Govindarajulu and S. Bringsjord. Toward Verification of Cognitive Calculi: An Initial Theorem. 2019. Available at http://www.naveensundarg.com/papers/UC_Semantics_Jan_6_2019.pdf.

[10] N. S. Govindarajulu, S. Bringsjord, and R. Ghosh. Toward the Engineering of Virtuous Machines. 2019. To appear in the proceeding of AAAI/ACM Conference on AI, Ethics and Society (AIES) 2019. arXiv preprint arXiv:1812.03868.

[11] N. S. Govindarajulu, S. Bringsjord, R. Ghosh, and M. Peveler. Beyond the doctrine of double effect: A formal model of true self-sacrifice. International Conference on Robot Ethics and Safety Standards, 2017.

[12] E. Maskin and A. Sen. *The Arrow Impossibility Theorem*. Columbia University Press, New York, NY, 2014.

[13] E. Mueller. *Commonsense Reasoning: An Event Calculus Based Approach*. Morgan Kaufmann, San Francisco, CA, 2006. This is the first edition of the book. The second edition was published in 2014.

[14] A. S. Rao and M. P. Georgeff. Modeling Rational Agents Within a BDI-architecture. In R. Fikes and E. Sandewall, editors, *Proceedings of Knowledge Representation and Reasoning (KR&R-91)*, pages 473–484, San Mateo, CA, 1991. Morgan Kaufmann.

conditions under which these formulae are provable.