EPiC Series in Computing

Volume 72, 2020, Pages 28–33

GCAI 2020. 6th Global Conference on Artificial Intelligence (GCAI 2020)

Adjudication of Symbolic & Connectionist Arguments in Autonomous-Driving AI

Michael Giancola[1], Selmer Bringsjord[1], Naveen Sundar Govindarajulu[1], and John Licato[2]

[1] Rensselaer AI & Reasoning (RAIR) Lab,
Rensselaer Polytechnic Institute (RPI), Troy NY 12180 USA
[2] Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida, Tampa FL 33620 USA
{ mike.j.giancola, selmer.bringsjord, naveen.sundar.g, john.licato } @gmail.com

Abstract

This paper discusses the tragic accident in which the first pedestrian was killed by an autonomous car: due to several grave errors in its design, it failed to recognize the pedestrian and stop in time to avoid a collision. We start by discussing the accident in some detail, enlightened by the recent publication of a report from the National Transportation Safety Board (NTSB) re. the accident. We then discuss the shortcomings of current autonomous-car technology, and advocate an approach in which several AI agents generate arguments in support of some action, and an adjudicator AI determines which course of action to take. Input to the agents can come from both symbolic reasoning and connectionist-style inference. Either way, underlying each argument and the adjudication process is a proof/argument in the language of a multi-operator modal calculus, which renders transparent both the mechanisms of the AI and accountability when accidents happen.

1 Introduction

On March 18, 2018, the first pedestrian was killed by an autonomous car. Of course, accidents do happen, and some are unavoidable. However, this one *was not*. Several key deficiencies in the design of Uber's autonomous car came into play in the moments immediately preceding the collision. We believe that, had there been a type of logicist automated reasoning in the car's automated systems, the fatality could have been prevented. In this paper, we present a quintet of desiderata describing such a logicist system, and outline an argument showing how things might have gone differently had the system been designed using a hybrid approach.

1.1 A Tragic Case Study

The accident occurred on a four-lane roadway, and began with the vehicle driving 44 mph in the rightmost lane, and the pedestrian walking a bicycle across the street starting from the (driver's) left side. The vehicle's radar first detected the pedestrian approximately 5.6 seconds before the

fatal collision [2]. Less than half a second later, the lidar detects the pedestrian but classifies her as "Other" and as an unmoving object. For the next 2.5 seconds, the lidar re-classifies her several times, alternating between "Vehicle" and "Other". The vehicle's automated-driving system (ADS) attempted to predict her direction of travel several times, but discarded any previous information about her trajectory every time it reclassified her. Finally, with 2.6 seconds until collision, the lidar classifies her as a bicycle, but as it was yet again changing her classification, discarded any past trajectory information, and hence determined that she was not moving. The upshot is that to this point the car had not taken any evasive or corrective action.

With 1.5 seconds left, the lidar re-classifies her yet again, this time as "Unknown". The system once again loses all of its tracking history. However, since at this point the pedestrian has entered the vehicle's lane, the ADS generates a plan to turn the car to the right to avoid her. Three hundred milliseconds later, the lidar re-classifies her as a bicycle, and determines that it would be impossible at this point to maneuver around her. With just 200 ms until collision, the ADS begins braking the vehicle, pitifully too late to stop in time.

2 A Hybrid Approach

Autonomous driving is — at least in the first author's opinion — an ideal application for a hybrid approach to AI, as the process of human driving at least seems to naturally call for both symbolic reasoning and statistical learning.

Consider the process of learning how hard to press the accelerator in order to move a car at various speeds. Within this task are many machine-learning sub-tasks, including computing a best-fit curve to an unknown function (e.g. a new driver learning the appropriate actuation of the pedal to gently accelerate without jolting the car forward) and transfer learning (e.g. an experienced driver adjusting the needed pressure on the pedal for a car they have never driven before). These tasks by our lights appear to be clearly best solved by machine learning.

Next, consider a scenario where you are trying to merge onto the highway in heavy traffic. You (d_1) see a space that you can merge into, and look to the driver (d_2) of the car behind that space. You make eye contact, and the other driver flashes his/her high beams at you. This has no formalized meaning, but you (rationally) interpret[1] it to mean that the driver **perceives** you and your vehicle, and hence **believes** that you **desire** to merge, and is **intending** to allow you to merge; therefore, you safely merge in front of them, and give a thank-you wave back. This scenario, unlike the last, isn't well-suited to being trained from numerical data. It is a reasoning process that one learns through one's ability to understand other humans at the "theory-of-mind" [16] (t-o-m) level.[2] This reasoning is best formalized in a modal logic which models the necessary t-o-m operators. In this paper we use the Inductive $\mathcal{DCEC}$ ($\mathcal{IDCEC}$), a nascent descendent of the $\mathcal{DCEC}$; both cognitive calculi are described in the following section. We can formalize the merging scenario above in standard $\mathcal{DCEC}$ as:

$$\mathbf{P}(d_2, now, \mathbf{D}(d_1, now, merge))$$
$$\mathbf{I}(d_2, now, Allow(d_1, merge))$$

With regard to the Uber accident, we believe that were there certain logicist elements present in the autonomous car, the accident could have been averted. Specifically, we propose the following list of desiderata (denoted by '$\mathcal{D}$') to characterize the type of automated reasoner we are looking to create:

[1] Though we do not discuss it here, such an act of interpretation may also be the kind of inferential leap that is better modeled by the hybrid approach we advocate [12, 17, 13].

[2] Each boldface verb in the prose immediately above is at the t-o-m level. For one of the first automated simulations of t-o-m reasoning, see [1].

Desiderata '$\mathcal{D}$'

We desire an automated reasoner that has the following properties:

- p_1 is defeasible (and hence nonmonotonic) in nature;
- p_2 is able to resolve or contain inconsistencies;
- p_3 makes use of values beyond bivalence (e.g. probabilities and/or strength-factors);
- p_4 is argument-based, where the arguments have internal inference-to-inference structure, so that justification (and hence explanation) is available; and
- p_5 is able to allow automated reasoning over the knowledge, belief, perception, etc. (t-o-m aspects) of humans involved in the scenario.

3 Steps Towards A Satisfactory Reasoner

We argue that a system which meets the above desiderata will be able to model and reason about not only the autonomous-car accident described above, but also a variety of scenarios in many domains. To the authors' knowledge, no such system currently exists. In order to create one, we will need to add several elements to our logicist armamentarium. The first addition is to turn to so-called *cognitive calculi*. We quickly summarize these calculi in the following subsections, then turn back to presenting our techniques with regard to the crash scenario.

3.1 Cognitive Calculi (Deductive)

Essentially, a deductive cognitive calculus is a quantified multi-operator modal logic such that its: proof/argument theory is specified in "natural deduction" form (traceable back to [7, 6]), operators cover all or most of human-level cognition (e.g., *believing*, *knowing*, *perceiving*, *communicating*, and also *obligations*, etc.), and semantics is exclusively proof-theoretic in nature.[3] Proof-theoretic semantics eschews model-theoretic and possible-worlds semantics in favor of the basic idea that meaning is provided to formulae and their constituents solely by virtue of the nature of proofs in which these things appear.[4]

In the present work, we specifically utilize elements of the Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) to model the perceptions (or lack thereof) and beliefs of the pilots, denoted **B** and **P** respectively. A dialect of $\mathcal{DCEC}$ is specified and used in [11]. $\mathcal{DCEC}$ can be thought of roughly as a quantified multi-operator modal logic with all of the introduction and elimination rules for first-order logic, plus a host of inference schemata to cover its many modal operators. Soundness proofs for cognitive calculi have been obtained but are out of scope. Also, an automated theorem prover for $\mathcal{DCEC}$ — ShadowProver [8] — has been created and is under active development.

3.2 Inductive Cognitive Calculi

$\mathcal{DCEC}$ is purely deductive and employs no uncertainty system, so it fails to satisfy p_1–p_4. Therefore, to meet desiderata p_1 and p_2, we use the *Inductive* $\mathcal{DCEC}$ ($\mathcal{IDCEC}$), which has been modified to handle inductive arguments.

To meet desideratum p_3, we employ "strength factors", in a nascent system for formalizing uncertainty in quantified modal logics, first presented in [9]. Strength factors can be viewed as a formalization of part of Chisholm's epistemology [3]. The current version, which the present work is based on, has a 13-value spectrum of strength, with zero being *counterbalanced* (no

[3] For specification of the formal language & proof theory of a cognitive calculus we direct readers to e.g. [11].
[4] For more on proof-theoretic semantics see [4, 5, 7, 15].

belief for or against some formula), increasing positive integers indicating stronger belief in favor of some formula, and decreasing negative integers indicating stronger belief against some formula. The relevant strength factors to this work (i.e. they are used in the arguments in the section below) are *likely* (level 2) and *more likely than not* (level 1).

Finally, to meet p_4, we need methods for adjudicating conflicting arguments with regard to which action to take in response to disagreement between multiple automated systems. Prior work in this area was presented in [10].

4 A Solution in $\mathcal{IDCEC}$

Our proposed solution is to install into an autonomous car a set of automated reasoners $\mathfrak{r}_1, \ldots, \mathfrak{r}_n$, and an AI adjudicator $\mathfrak{a}^\star$ that receives, analyzes, and weighs proofs/arguments generated by $\mathfrak{r}_i$.[5] Each automated reasoner would receive input from various sensors and determine what action it believes is appropriate at each timestep. They would each compute an argument as justification for their conclusion in the form of a proof in $\mathcal{IDCEC}$. When two (or more) automated reasoners disagree, the adjudicator $\mathfrak{a}^\star$ resolves the conflict.

For the purposes of the running example of this paper, denote two automated reasoners $\mathfrak{r}_1$ and $\mathfrak{r}_2$, and the adjudicator $\mathfrak{a}^\star$. $\mathfrak{r}_1$ receives input from the ADS and computes arguments that correspond with what the ADS did in the actual accident. $\mathfrak{r}_2$ takes a different approach; it retains trajectory information regardless of the classification of the object, and computes an argument in favor of braking. Due to space constraints, only a sketch of each proof is provided below. Denote the following time steps: t_0 the moment when the radar first perceives the pedestrian, t_1 the moment when the lidar first perceives the pedestrian, and t_2 a short moment after that. Finally, denote c the car and o^* the pedestrian. Also, while the necessary elements are not yet implemented in an automated reasoner to generate such proofs, we created simplified versions of the proofs to run in ShadowProver to show that this AI could have computed its argument fast enough to prevent the accident. The proof provided by $\mathfrak{r}_1$ was generated by ShadowProver in 0.35 seconds and in 0.32 seconds for $\mathfrak{r}_2$'s argument.

$\mathbf{B}^1(\mathfrak{r}_1, t_0, \mathit{Stationary}(o^*))$	$\mathbf{B}^1(\mathfrak{r}_2, t_0, \mathit{Stationary}(o^*))$
$\therefore \mathbf{B}^1(\mathfrak{r}_1, t_0, \neg \mathit{GoingToCollide}(c, o^*))$	$\therefore \mathbf{B}^1(\mathfrak{r}_2, t_0, \neg \mathit{GoingToCollide}(c, o^*))$
$\mathbf{B}^1(\mathfrak{r}_1, t_1, \mathit{Stationary}(o^*))$	$\mathbf{B}^2(\mathfrak{r}_2, t_1, \neg \mathit{Stationary}(o^*))$
$\ldots$	$\mathbf{B}^2(\mathfrak{r}_2, t_1, \neg \mathbf{P}(o^*, t_1, c))$
$\therefore \mathbf{B}^1(\mathfrak{r}_1, t_1, \neg \mathit{GoingToCollide}(c, o^*))$	$\therefore \mathbf{B}^2(\mathfrak{r}_2, t_1, \mathit{GoingToCollide}(c, o^*))$
$\therefore \mathbf{B}^1(\mathfrak{r}_1, t_2, \neg \mathit{NeedToBrake}(c))$	$\therefore \mathbf{B}^2(\mathfrak{r}_2, t_2, \mathit{NeedToBrake}(c))$
Argument of $\mathfrak{r}_1$	**Argument of $\mathfrak{r}_2$**

Clearly, $\mathfrak{r}_1$ and $\mathfrak{r}_2$ directly contradict each other. Therefore, $\mathfrak{a}^\star$ is called to adjudicate and make the final decision. In this case, the adjudication is easy: $\mathfrak{r}_2$'s belief that the car should brake is stronger than $\mathfrak{r}_1$'s belief that it doesn't need to.[6] Hence, $\mathfrak{r}_2$'s argument defeats $\mathfrak{r}_1$'s, and the vehicle begins braking 0.35 seconds[7] after the lidar first perceived the pedestrian. This would've given the car approximately 4.85 seconds to prevent the collision. With this amount of time, in the context of the conditions that framed the accident, it is reasonable to believe that the car, through some combination of adjusting its course and braking, would've been able to avoid hitting the pedestrian.

[5] The adjudicator's formal properties include but exceed those given to the aggregation function shown to be inadequate in Arrow's Impossibility Theorem (AIT; a nice proof and discussion is provided in [14]). Space constraints prevent coverage here of this dimension of our work, which is intended to surmount AIT.

[6] We justify assigning $\mathfrak{r}_2$'s belief a higher strength factor because its belief is based on *multiple* observations, whereas $\mathfrak{r}_1$'s are based on single observations.

[7] Plus a negligible amount of time for the — in this case, rather simple — adjudication process.

5 Conclusion

Combining symbolic and connectionist techniques can provide the best of both worlds. We gain the abilities enabled by deep learning and are able to generate proofs/arguments using symbolic techniques that afford transparency; in particular, when things go wrong, it will be clearer *why* they went wrong, and what specific component of the system is to blame. Such diagnoses should in turn make possible the engineering of safer AI systems, and save lives.

6 Acknowledgements

The authors are indebted to ONR for support of our r&d in the area of automated defeasible reasoning intended to surmount Arrow's Impossibiity Theorem, and to AFOSR for support of development of novel formalisms underlying this and other forms of automated reasoning.

References

[1] K. Arkoudas and S. Bringsjord. Propositional Attitudes and Causation. *International Journal of Software and Informatics*, 3(1):47–65, 2009.

[2] Ensar Becic. Vehicle Automation Report HWY18MH010. Technical report, National Transportation Safety Board, Tempe, AZ, 2019. Accessible as of November 22, 2019 here: https://dms.ntsb.gov/public/62500-62999/62978/629713.pdf.

[3] R. Chisholm. *Theory of Knowledge 3rd ed.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[4] M. Dummett. *Frege. Philosophy of Language (2nd ed).* Duckworth, London, UK, 1981.

[5] M. Dummett. *The Logical Basis of Metaphysics.* Duckworth, London, UK, 1991.

[6] F. Fitch. *Symbolic Logic: An Introduction.* Ronald Press, New York, NY, 1952.

[7] G. Gentzen. Untersuchungen über das logische Schlieben I. *Mathematische Zeitschrift*, 39:176–210, 1935.

[8] Naveen Sundar Govindarajulu. ShadowProver, 2016. https://naveensundarg.github.io/prover/.

[9] Naveen Sundar Govindarajulu and Selmer Bringsjord. Strength Factors: An Uncertainty System for Quantified Modal Logic. In V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade, and G. Qi, editors, *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning (LFU-2017)*, pages 34–40, Melbourne, Australia, August 2017.

[10] Naveen Sundar Govindarajulu and Selmer Bringsjord. Argument Aggregation in a Deontic Logic. In Mohammad O. Tokhi, Maria Isabel A. Ferreira, Naveen Sundar. Govindarajulu, Manuel Silva, Gurvinder S. Virk, Endre E. Kadar, and Sarah R. Fletcher, editors, *Intelligence, Robots and Ethics: Proceedings of the Fourth International Conference on Robot Ethics and Standards*, pages 17–18. CLAWAR Association Ltd, UK, 2019.

[11] 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. International Joint Conferences on Artificial Intelligence, 2017.

[12] John Licato and Zaid Marji. Probing Formal/Informal Misalignment with the Loophole Task. In *Proceedings of the 2018 International Conference on Robot Ethics and Standards (ICRES 2018)*, 2018.

[13] John Licato, Zaid Marji, and Sophia Abraham. Scenarios and Recommendations for Ethical Interpretive AI. In *Proceedings of the AAAI 2019 Fall Symposium on Human-Centered AI*, Arlington, VA, 2019.

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

[15] Dag Prawitz. The Philosophical Position of Proof Theory. In R. E. Olson and A. M. Paul, editors, *Contemporary Philosophy in Scandinavia*, pages 123–134. Johns Hopkins Press, Baltimore, MD, 1972.

[16] D. Premack and G. Woodruff. Does the Chimpanzee have a Theory of Mind? *Behavioral and Brain Sciences*, 4:515–526, 1978.

[17] Ryan Quandt and John Licato. Problems of Autonomous Agents Following Informal, Open-Textured Rules. In *Proceedings of the AAAI 2019 Spring Symposium on Shared Context*, 2019.