Research Article

Selmer Bringsjord, Naveen Sundar Govindarajulu*, and Michael Giancola

Automated argument adjudication to solve ethical problems in multi-agent environments

https://doi.org/10.1515/pjbr-2021-0009
received May 7, 2020; accepted December 30, 2020

Abstract: Suppose an artificial agent a_{adj}, as time unfolds, (i) receives from multiple artificial agents (which may, in turn, themselves have received from yet other such agents…) propositional content, and (ii) must solve an ethical problem on the basis of what it has received. How should a_{adj} adjudicate what it has received in order to produce such a solution? We consider an environment infused with logicist artificial agents $a_1, a_2, \ldots, a_n$ that sense and report their findings to "adjudicator" agents who must solve ethical problems. (Many if not most of these agents may be robots.) In such an environment, inconsistency is a virtual guarantee: a_{adj} may, for instance, receive a report from a_1 that proposition ϕ holds, then from a_2 that $\neg\phi$ holds, and then from a_3 that neither ϕ nor $\neg\phi$ should be believed, but rather ψ instead, at some level of likelihood. We further assume that agents receiving such incompatible reports will nonetheless sometimes simply need, before long, to make decisions on the basis of these reports, in order to try to solve ethical problems. We provide a solution to such a quandary: AI capable of adjudicating competing reports from subsidiary agents through time, and delivering to humans a rational, ethically correct (relative to underlying ethical principles) recommendation based upon such adjudication. To illuminate our solution, we anchor it to a particular scenario.

Keywords: cognitive calculus, cognitive robotics, adjudication, adjudication for ethical principles, ethics

1 Introduction

Neurobiologically normal, mature human beings often enjoy the luxury of being able to make decisions in and unto themselves. A hot burner on a stove, if mistakenly touched, can lead to a rather quick decision to pull away; and while such a decision usually happens by reflex, the human in question can then inspect his/her finger and decide whether or not treatment is needed. But as we know, decision-making is not always this independent; sometimes what humans must decide must factor in what has been received from other humans. When this happens, the situation can be quite tricky. Perhaps this is especially true when the required decision is needed in order to try to resolve some ethical problem. Note that in the course of human affairs, profound ethical decisions have long needed to be made in these kinds of buzzing, dynamic, dialectical, multi-agent scenarios, where all the agents are humans. Deep and challenging legal cases provide a case in point,[1] as for that matter so do command-and-control challenges to humans in warfare, a domain that our case study given below relates to.[2] But our task herein is to formalize the AI correlate of this kind of tricky situation and to propose a way for a new kind of AI to solve the correlate.

This AI correlate, in broad strokes for the moment, has the following structure: An artificial agent a_{adj}, as

* **Corresponding author: Naveen Sundar Govindarajulu,** Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, United States of America, e-mail: naveen.sundar.g@gmail.com
Selmer Bringsjord: Department of Cognitive Science, Rensselaer AI & Reasoning (RAIR) Lab, Troy, NY 12180, United States of America; Department of Computer Science, Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, United States of America, e-mail: selmer.bringsjord@gmail.com
Michael Giancola: Department of Computer Science, Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, United States of America, e-mail: mike.j.giancola@gmail.com

1 It would, e.g., be quite interesting to see how an artificial agent of the type introduced in the present paper would, if suitably "armed" with starting information, adjudicate the Dreyfus case, covered brilliantly in literary fashion by Proust [68], and in hard-nosed journalistic fashion in ref. [69].

2 For a deeper, more complex case study in this domain, it would be interesting to see whether decision-making as to when to engage Pershing's new-world forces in WW I, which involved many a mind interacting with Pershing's, could be automated. For background, see ref. [70].

time unfolds, (i) receives from multiple artificial agents (which may, in turn, themselves have received from yet other such agents...) propositional content, and (ii) must solve an ethical problem on the basis of what it has received. How should a_{adj} adjudicate what it has received in order to produce such a solution? We consider an environment infused with logicist artificial agents a_1, a_2,..., a_n that sense, and report their findings to "adjudicator" agents who must solve ethical problems. (Many if not most of these agents may be robots.) In such an environment, inconsistency is a virtual guarantee: a_{adj} may, for instance, receive a report from a_1 that proposition ϕ holds, then from a_2 that $\neg\phi$ holds, and then from a_3 that neither ϕ nor $\neg\phi$ should be believed, but rather ψ instead, at some level of likelihood. We further assume that agents receiving such incompatible reports will nonetheless sometimes simply need, before long, to make decisions on the basis of these reports, in order to try to solve ethical problems. We provide a general solution to such a quandary: AI capable of adjudicating competing reports from subsidiary agents through time, and delivering to humans a rational, ethically correct (relative to underlying ethical principles) recommendation based upon such adjudication, with concomitant justification for the recommendation. To illuminate our solution, we anchor it to a particular scenario.

Note that the sort of quandary we seek automated solutions for have sometimes been called *ethical dilemmas*.[3] We do not wish to focus on ethical dilemmas in the present paper, but rather on a general form of an *ethical problem*, as defined below (Section 3.2). Yet we do offer here an observation regarding such dilemmas: namely, *they can't be resolved by logicist* (*intelligent*) *artificial agents that don't have the capacity to adjudicate competing, incompatible arguments in general*. Our emphasis in the present paper is to introduce formalisms and techniques for how such a capacity can be given to an artificial agent. Once that is accomplished, dividends will have been paid for use in the case of outright ethical dilemmas.

The remainder of the paper unfolds as follows. Next (Section 2), we introduce the methods by which we bring to bear AI agents which can adjudicate thorny ethical problems. We first (Section 2.1) explain the brand of AI that we pursue. We then summarize our approach to machine ethics, which is overall based on four general steps (Section 2.2). In Section 2.3, we quickly point out that, at a finer-grained level than our four general steps, lies a specific need to obtain AI able to handle reasoning that occurs as a dialectic through time. We then present our results in Section 3. First, we describe the particular formal logic (or, more accurately for reasons we explain, *cognitive calculus*) that is the basis for the kind of automated adjudication capability needed in multi-agent situations where the agents offer competing, incompatible recommendations in ethically charged situations (Section 3.1), define the concept of an ethical problem, in our general, abstract sense (Section 3.2), and define as well what it is for such a problem to be solved (Section 3.3). At this point, we are ready to show that our approach can handle a challenging scenario that demands adjudication of competing arguments through time, and do so in Section 3.4. We next anticipate and respond to a series of objections to our approach (Section 4). We conclude the paper in Section 4.4 by reflecting upon yet another objection to our approach, one that points to a need for future work of a certain kind.

2 Methods

2.1 Our overall logicist approach

AI has become a vast field, as chronicled and explained in ref. [1]. Accordingly, the pursuit of computing machines that qualify as intelligent, and indeed even the meaning of "intelligent" itself in some contemporary debates, are defined differently by different researchers and engineers, even though all of them work under the umbrella of "AI." Our approach is a logicist one, or – as it's sometimes said – a logic-based one. A full characterization of our approach to AI and robotics is of course beyond the reach of the present paper, but we must give at least information to orient the reader, and we do so now. We turn first to the generic concept of an *artificial intelligent agent*, or – since by context it's clear that we must have intelligence, in some sense, front and center – simply *artificial agents*.

2.1.1 Artificial agents/AI, generically speaking

For present purposes, we rely upon how dominant textbooks, for example ref. [2,3], characterize artificial agents. Their characterization is simply that such an agent com-

3 Perhaps the best starting point for learning about ethical dilemmas is the work of Kohlberg, and the scenarios that for him were given as classical ethical dilemmas. The ability to handle ethical dilemmas was for Kohlberg the hallmark of sophisticated ethical reasoning and decision-making; see ref. [71].

putes a function from what is perceived (*percepts*) to behavior (*actions*). All such agents are assumed to operate this way in a certain *environment*, but for present purposes, we can leave explicit consideration of this aspect of the AI landscape to the side; doing so causes no loss of generality or applicability for the work on machine ethics we relate herein. But what about the nature of the function from percepts to actions? As pointed out in the course of an attempt to show that the so-called Singularity[4] is mathematically impossible, ref. [4] notes the fact that in the dominant AI textbooks, these functions are firmly assumed to be recursive. In the present paper, we affirm this assumption, but the reader should keep in mind that despite this affirmation, our AI technology can still be based upon automated reasoning that is routinely applied to problems that are Turing-uncomputable *in the general case*. After all, all automated reasoners that are specifically automated theorem provers for first-order logic confront the *Entscheidungsproblem*, first shown unsolvable by Church (Church's Theorem).

2.1.2 The logicist approach to AI/robotics

We can now quickly state the heart of our logicist approach to AI and robotics, as follows. The artificial agents we specify and implement compute their functions (from, again, percepts to actions) via automated reasoning over a given formula ϕ in some formal language $\mathcal{L}$ for some formal logic $\mathscr{L}$. This means that what these agents perceive must ultimately be transduced into content expressed in such formulae; and it means that an action, before translated into lower-level information that can trigger/control an effector, must also be expressed as a formula. The reader will see this in action below, both abstractly (when we explain what an ethical problem in general is for us, and what a solution to such a problem consists in, generally speaking), and when we present the promised scenario. But how, specifically, are the functions computed in the case of such agents? The answer is straightforward: These functions are computed by automated reasoning. Of course, it has long been known that computation, while often understood in procedural terms (e.g., in terms of Turing machines), is fully reducible to, and usable as, reasoning.[5]

What about robotics, specifically? Well, first, the type of robotics we pursue is best called *cognitive* or – taking account of the terminological fact that sometimes the introduction of cognitive elements to a formalism makes that formalism *behavioral* in nature; see e.g. ref. [5] – *behavioral*. We specifically pursue cognitive robotics as defined in ref. [6],[6] with a slight formal tweak, and say simply that a cognitive robot is one whose macroscopic actions are a function of what the robot knows, believes, intends, and so on. As seen below, these verbs are at the heart of a *cognitive calculus*, the class of cognitively oriented logics we employ for machine ethics in general, and automated argument adjudication specifically. Cognitive calculi are explained in more detail in Appendix A1.

2.1.3 Ethics is rooted in logic

One final point about our approach, one the reader at this juncture will find thoroughly unsurprising: we view attempts to decide what an agent, whether artificial or human or – for that matter – extraterrestrial, ought (can, etc.) to do, to be, overall, a matter of what holds, declaratively speaking. This view on our part is simply derived from the observation that, from the standpoint of professional ethics as practiced and taught in the Academy, that which is obligatory (permissible, forbidden, uncivil, supererogatory, etc.) is determined by the standing of propositions, where those propositions are expressed in declarative statements. Given this, formal logic becomes a rather promising discipline for capturing ethics systematically; and in its computational guise, formal logic then becomes in turn a natural vehicle for computing what is obligatory (permissible, etc.). Beyond this, one might wonder when, specifically, *computational* logic arrives on the scene for – at least aspirationally speaking – imbuing a computing machine with a capacity to make ethical decisions unto itself. Presumably, this point in time may coincide with when people first rendered deontic logic in computation; an early example is ref. [7].

2.2 Our machine-ethics approach, brutally summarized

Our "Four Steps" to making morally correct machines, depicted pictorially from a high level in Figure 1, was first presented in ref. [8]. We now briefly explain the steps in question.

4 The point in future time at which, so the story goes, AIs reach human-level intelligence, and the immediately thereafter ascend to intellectual heights far, far above our own.

5 This is what allows proofs of the Halting Problem for Turing machines to be relied upon to prove the undecidability of the *Entscheidungsproblem*; see e.g. ref. [72].

6 As is pointed out in that paper, as far as most relevant thinkers know, it was actually Ray Reiter who coined and first defined the phrase "cognitive robotics."

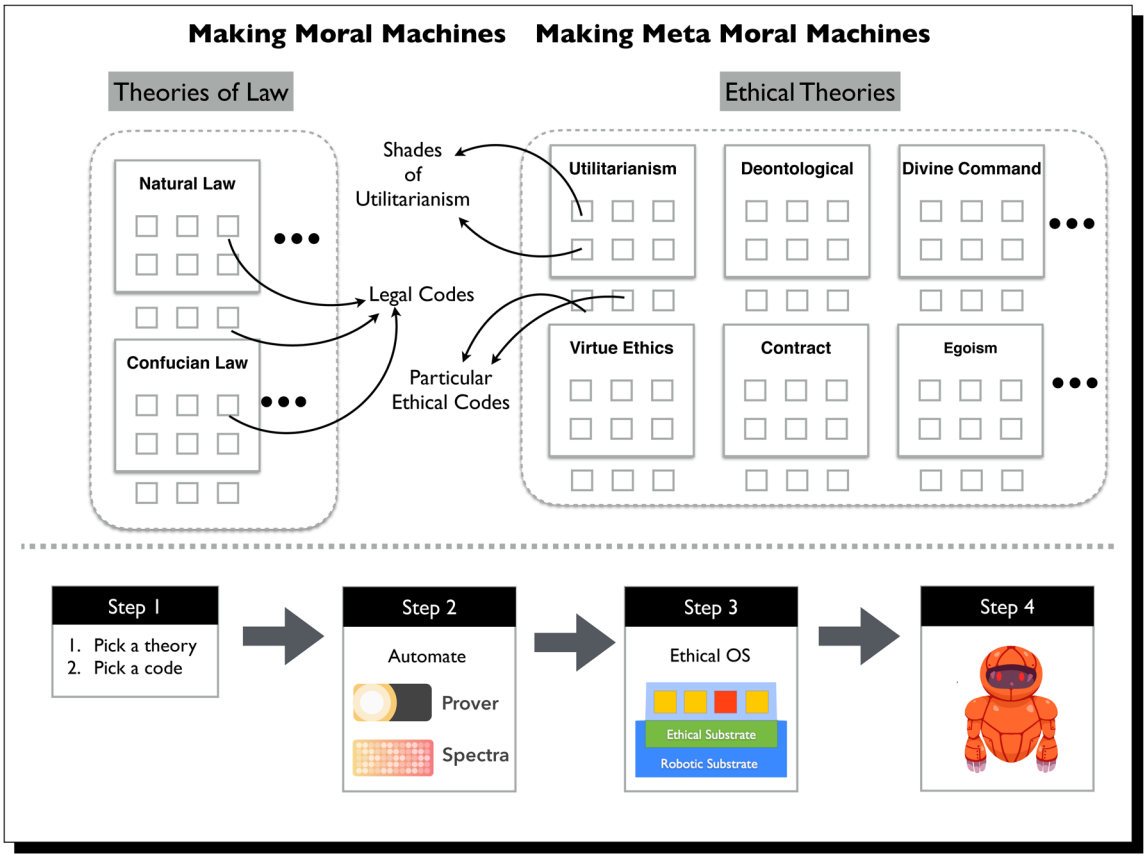

Figure 1: The four steps in making ethically correct machines.

The first step is the selection of an ethical theory (or theories), from a family of such theories.[7] The well-known families are shown in Figure 1. For instance, one family of ethical theories is *divine-command*; another is *utilitarianism*; a third is *virtue ethics*. An ethical theory that is a member of the second of these families would be standard act utilitarianism, according to which, put quickly and without nuance, one ought to always perform those actions from available ones that maximize happiness among the population that can be affected; for exposition of this ethical theory, see the old but venerable [9]. For the most part, in the past, we have, at one point or another, carried out work based on each family shown in Figure 1. For instance, for some prior work that reflects pulling from both *utilitarianism* and from *deontological* families, see ref. [10], which centers around the so-called *Doctrine of Double Effect*, a version of which we use below. Note that we do not advance a framework that requires one to commit to any particular one of these theories, or even to particular families of theories. Our framework is general enough that it can be applied to *any* ethical theory, or collection or family thereof. That said, there are a few high-level requirements for our general approach to machine ethics to be pursued, as follows. Unsurprisingly, these requirements are rooted in formal logic.

Suppose that we have a family $\mathscr{E}$ of ethical theories of interest. We require that any ethical theory $\mathcal{E} \in \mathscr{E}$ regiments how deontic operators that are invariants across all ethical theories (e.g., *obligatory*, *permissible*, *forbidden*, *supererogatory*, etc.) are to apply to either or both of states-of-affairs and actions performable by agents. In our approach, any ethical theory usable in The Four Steps must be formalized so as to capture these notions.

This formalization is made possible by a *cognitive calculus*. While details are provided in Appendix A1, such a calculus C is a pair $\langle \mathcal{L}, \mathcal{I} \rangle$ where $\mathcal{L}$ is a formal language (composed in turn, minimally, of a formal grammar,

7 This first step includes not only this selection, but the selection, immediately thereafter, of a particular domain-specific ethical code that falls under the selected theory, but this is left aside for economy here.

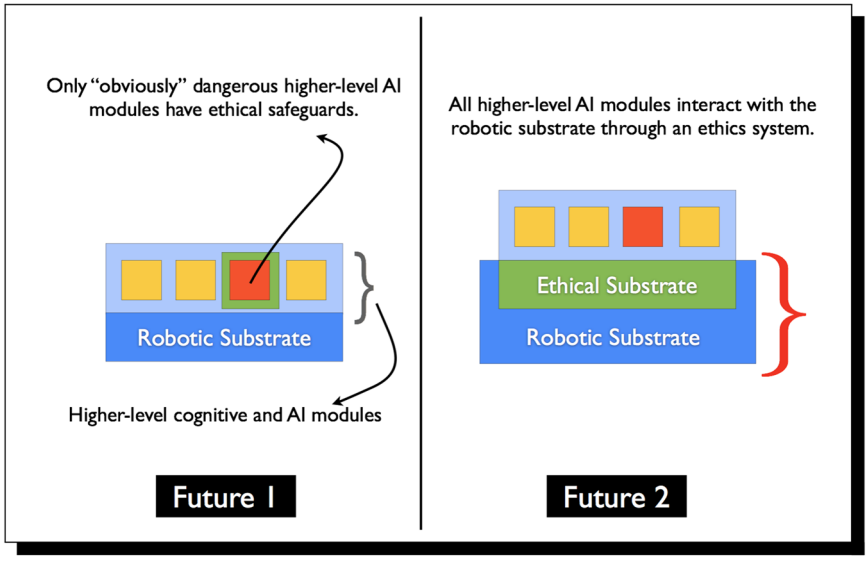

Figure 2: Two futures – with and without an ethical substrate. Higher-level modules are vulnerable to tampering. The ethical substrate protects the robotics substrate from rogue modules (figure from ref. [18]).

and an alphabet/symbol set), and $\mathcal{I}$ is a collection of inference schemata (sometimes called a *proof theory* or *argument theory*) $\mathcal{I}$. Within the present paper, as explained below, the cognitive calculus μC will be utilized.

The second of The Four Steps is to automate the generation of proofs of (un-)ethical behavior so that the reasoning can be utilized and acted upon by autonomous robots. As we explained above, logicist AI for us entails that the percepts-to-actions functions are handled by automated reasoning. We specifically use ShadowProver [11,12], an automated theorem prover for cognitive calculi.

Step 3 in The Four Steps is to integrate automated ethical reasoning into a cognitive robot's operating system (details available in ref. [8,13]). There are basically two possible approaches to this (see Figure 2). In the first, only "obviously" dangerous capabilities of an AI/robot are restricted with safeguards implemented above the OS level. In the second approach, all AI code must comply with an "Ethical Substrate" that is part of the OS. Unfortunately, while the first approach allows rapid engineering, unforeseen unethical behavior on the part of the AI/robot is entirely possible (see ref. [13]). Only by way of the second option is there any guarantee that the selected ethical theories and associated ethical codes will remain in force.

In the fourth and final step, we implement our ethical OS into a physical robot and arrive at a moral machine.[8] Specifically, a machine which can be formally verified to always act in accordance with an ethical theory.

2.3 The specific need to handle dynamic dialectic

So much for a high-altitude overview of our approach to machine/robot ethics, in the form of The Four Steps. We now draw the reader's attention to a specific capability we need in our AI for making the Four Steps concrete reality. In short, we need our automated reasoners to be able to handle, throughout time, ethical reasoning and

8 The fourth step in our Four Steps should not be taken lightly and is indeed a full-blown process unto itself. We have been told this in no uncertain terms: We have encountered roboticists of incontestable stature who maintain that while our formal work certainly comes with its own significant challenges for the computational logician, and while our automated-reasoning technology is robust and requires *software* engineering of high quality, the genuine implementation across the board of what eventuates from our Step 3 comes with its own deep, deep challenges – challenges that can only be met by engaging in a phase of concrete and practical *physical* engineering. We do not in any way regard this position with skepticism; in fact, we accept it. At the end of the day, even when our work is implemented on physical systems, that is, in physical robots (e.g. see ref. [73]), our work is in the realm of *cognitive* robotics [6], the hallmark of which is that substantive actions performed by the robots in question are a function of the cognitive attitudes of such robots (where such attitudes are, for us, represented in cognitive calculi). But physical robots, and for that matter "softbots" once implemented in physical environments, require an implementation phase that is long and challenging in its own right, even when this phase is supplied with the fruits of our Step 3.

counter-reasoning. We specifically need, therefore, automated reasoners with the capability to detect and resolve inconsistencies arising from competing arguments and positions on moral matters. But this is only one aspect of sevenfold desiderata for the kind of capability our AI must have. We dub this set of desiderata "$\mathcal{D}$," and lay down that an automated reasoner of the kind we seek must:

Desiderata ($\mathcal{D}$)

- d_1 be defeasible (and hence – to use the term frequently employed in AI – nonmonotonic) in nature through time;
- d_2 be able to resolve inconsistencies of various sorts, ranging, e.g., across ω-inconsistency to "cognitive inconsistency" (e.g., an agent α believing both ϕ and $\neg\phi$) to standard inconsistency in bivalent extensional logic when appropriate, and tolerate them when necessary in a manner that fully permits reasoning to continue;
- d_3 make use of values beyond standard bivalence and standard trivalence (e.g., beyond the Kleenean TRUE, FALSE, UNKNOWN trio), specifically probabilities *and* strength-factors (the latter case giving rise to multivalued inductive logic);
- d_4 be argument-based, where the arguments have internal inference-to-inference structure both in terms of declarative formulae (and possibly diagrams) and inference schemata, so that detailed verification, justification/explanation is available;
- d_5 have specified inference schemata (which sanction the inference-to-inference structure referred to in d_4), whether deductive or inductive, that are machine-checkable;
- d_6 be able to allow automated reasoning over the sociocognitive elements of knowledge, belief, desire, perception, communication, emotion, etc., of relevant artificial and human agents, where these elements are irreducibly intensional;
- d_7 be able to allow automated reasoning that can tackle Turing-unsolvable reasoning problems, e.g., queries about provability at and even above the *Entscheidungsproblem* (e.g., at and above $\sum_1^0$ and $\sum_1^1$ in the Arithmetical and Analytical Hierarchies, resp.)

2.3.1 Relevant prior work

2.3.1.1 OSCAR

One of the major modern contributors to research in argument-based reasoning through time and cognitive change on the part of the agents involved (sometimes called argument-based *defeasible* reasoning[9]) is John Pollock, a philosopher who made seminal contributions to AI. Pollock developed a robust theory of rationality which revolves around the ability to reason defeasibly. He also implemented this theory in an AI agent called "OSCAR." Among all those who have worked on defeasible reasoning in argument-centric fashion, there can be no denying that Pollock stands as hands down the most ambitious, since he repeatedly claimed and sought to defend the dual propositions that his line of work can be used to literally build artificial persons, and that the essence of this line is to formalize and computationally implement defeasible argumentation (note the titles in ref. [14,15]). For technical details regarding Pollockian work given in AIish terms, see ref. [16,17].

OSCAR employs first-order inference as well as Pollock's (coarse-grained) schemata for defeasible reasoning in order to solve problems. Input to OSCAR includes a list of formulae, termed *givens*, with corresponding rational-number strength values (not probabilities; Pollock adamantly rejected the probability calculus in all standard forms) and the ultimate epistemic interest of the artificial agent: the formula OSCAR will try to establish from the givens, once processing is launched. The strength of formulae is a rational number ranging from 0.0 to 1.0 (inclusive), where 1.0 means that the formula is known with absolute certainty to hold. Values less than 1.0 indicate levels of uncertainty in the truth of the statement and allow such statements to be defeated by arguments that rely solely on statements of higher strength.

OSCAR is impressive, but falls short of meeting the list $\mathcal{D}$ of desiderata prescribed above. First, while OSCAR includes a set of deductive inference schemata for first-order logic, it has no inference schemata whatsoever for its *inductive* arguments. As a result, its analysis of several arguments makes no directed use of the internal structure of individual inference steps. This generally corresponds to abstract treatments of arguments and the suppression of the specifics of individual inferences that are chained together to make an argument; a classic scheme in this tradition is presented in ref. [18]. OSCAR therefore doesn't satisfy d_4 and doesn't satisfy d_5. Also, as it is limited to

9 AI cognoscenti will know that "defeasible reasoning" and "defeasible logic" are phrases that are for the most part coextensive with – in AI – "nonmonotonic reasoning" and "nonmonotonic logic." We do not have the space to canvas, even just in part, work that falls under these phrases. Seminal work on nonmonotonic reasoning/logic was carried out nearly half a century ago, e.g., by McCarthy [74] and Reiter [75] – but this work was not, at least by our metrics, argument-based; and nor was subsequent effort in this vein.

first-order logic, OSCAR cannot satisfy d_6 without falling into unsoundness, as shown in ref. [19].

2.3.1.2 Other defeasible argumentation systems

What about work in defeasible argumentation systems, in general, with an eye to $\mathcal{D}$? A full answer to this question that takes account of all (or, for that matter, even most) prior systems is space-wise infeasible (a nice but now somewhat dated survey is provided in ref. [20]); we thus restrict ourselves to mentioning two pieces of impressive prior work, neither of which significantly overlaps our new approach, as we explain:

1. Ref. [21] presents a general framework for structured argumentation, and the framework is certainly computational in nature. This framework, ASPIC$^+$, is in fact Pollockian in nature, at least in significant part. More specifically, ASPIC$^+$ is based upon two fundamental principles, the second of which is that "arguments are built with two kinds of inference rules: strict, or deductive rules, whose premises guarantee their conclusion, and defeasible rules, whose premises only create a presumption in favor of their conclusion" (ref. [21], p. 31). This second principle is directly at odds with desideratum d_5. In our approach, all non-deductive inference schemata are mechanically checkable, in exactly the way that deductive inference schemata are. For instance, if some inference is analogical in nature, as long as the schema $\frac{\Phi}{C}$ (Φ for a collection of premises and C for the conclusion) for an analogical inference is correctly followed, the inference is watertight, no different than even *modus ponens*, where of course specifically we have $\frac{\phi \rightarrow \psi, \phi}{\psi}$.[10] Along this line, the reader will soon see that even the simplified cognitive calculus we use to obtain an implementation that meets an argument-adjudication challenge, the calculus μC is based on inference schemata purported applications of which can be mechanically certified as correct, or rejected.
2. Ref. [22] is an overview of implementations of formal-argumentation systems. However, the overview is highly constrained by two attributes. The first is that their emphasis is on *Turing-decidable* reasoning problems. As to the second attribute, the authors are careful to say that their work is constrained by the "basic requirement" that "conflicts" between arguments are "solved by selecting subsets of arguments," where "none of the selected arguments attack each other." Both of these attributes are rejected in our approach. In fact, with respect to the first, most of the interesting parts of automated-reasoning science and technology for us only *start* with problems at the level of the *Entscheidungsproblem*; see in this regard desideratum d_7. As to the second attribute, it also is not true of our approach, as will be seen below.

2.3.1.3 Work in judgment/argument aggregation

We now discuss, briefly, work in "aggregation" that is related to our research. From a scholarly point of view, this discussion must start with the fact that, as many readers will know, nobelist Arrow's [23] stunning "Impossibility Theorem" (AIT) roughly says that, without a "dictator" who holds sway, it's mathematically impossible for a group of agents in that group to have their individual preferences aggregated to yield preferences for the group as a whole.[11] An immediate and equally stunning corollary of AIT is that a "meta" agent cannot make a decision based on the input from an advisory group composed of agents, where that input is an aggregation of the preferences of the individuals in the group. Since we seek such an agent, how can we succeed? Inevitably, the constraints associated with AIT must be, at minimum, massaged. AIT is a negative result regarding the aggregation of *preferences*; moreover, AIT is a limitative theorem that only goes through under the constraint of certain axioms (that are out of scope in the present paper). Clearly, humans and machines (including on the machine side even simple sensors) routinely provide information to decision-makers that greatly exceeds a preference. For example, a commander might need to seek aggregation of a series of reports from the individual agents $\mathfrak{a}_i$ in a group that are each reporting information about the location of a bomb; such reports aren't preferences, but are rather propositions or claims or hypotheses (or perhaps even educated guesses). This broader problem, which is expressed in a family of theorems we denote via "AIT$^+$," has been the

10 For a discussion of this sort of explicit rigidity in the case of analogical inference, see ref. [76].

11 A very nice introduction to AIT is available in ref. [40]. An early chapter by A. Sen in this recent volume gives an economical and lucid proof of AIT.

genesis of the "judgment aggregation" field.[12] Yet our approach is specifically based upon not just the aggregation of judgments, but the aggregation, in particular, the *adjudication*, of *arguments*. Given this, what related work is there, and how does it compare and contrast with our approach and systems? We now answer this question, by way of two steps.

The first step is to emphasize that our approach is indeed best labeled *argument adjudication*. Argument adjudication is not to be confused with argument *aggregation* [24], which is based on the standard approach of treating arguments as abstract objects having none of the nuanced, internal structure analogous to what formal proofs have (this standard, abstract conception is given in ref. [18]). Since our formal theory of argumentation is fundamentally different than what is seen in argument aggregation, the axioms that constrain argument adjudication are different than those operative in argument aggregation. Along the same line, while we happen to often represent proofs and arguments in graphs (specifically hypergraphical digraphs), prior work on the so-called *graph aggregation* [25] is completely separate from our logico-mathematics and corresponding engineering, since in this other work, graphs are treated abstractly and any internal inferential structure they might have is ignored.

Nonetheless, and this is the second of our promised two steps, we find that some excellent work on argument aggregation (or as some would be inclined to say, on argument-centric judgment aggregation) is worth calling out here, in connection with our research on argument adjudication (in service, in the present paper, of machine ethics). Specifically, we cite and briefly comment on a few papers, as follows.

Ref. [26] introduces the concept of a *Value-Based Argumentation Framework* (VAF). VAFs "allow for attacks to succeed or fail, depending on the relative worth of the values promoted by the competing arguments." This is certainly impressive work – but it fails with respect to desideratum d_4, since it is built upon a definition of what counts as an argument deriving from the Dungian conception [18], according to which one can have a *bona fide* argument in the complete absence of particular inference schemata and particular content. A parallel diagnosis applies to ref. [27], for these authors have in no way considered the internal structure of deductive and inductive (e.g., analogical, abductive, enumerative inductive, etc.) arguments, and more importantly have not considered either arguments expressed in implemented systems for expressing and checking them. Our approach is radically different; for it gets off the ground only because we have particular inference schemata, and automated reasoning over them, and over detailed declarative content rendered in the formal languages of cognitive calculi.[13] The same divergence between our approach and [28] arises, because this work, certainly impressive as well, is explicitly devoted to the merging of arguments cast in Dungian style.

Of course, we do not contend to have taken account of literally all prior work that might meet $\mathcal{D}$. Of all that we have seen after considerable digging, the results are given immediately above, but it is also important to note that other computational logicist approaches to ethics exist, and we briefly review some of this work now.

2.3.1.4 Other computational logicist approaches to ethics

Noteworthy in this sphere is work by Ganascia [29], both because here he calls for the need to dynamically handle conflicting ethical positions and arguments, and because toward the end of this paper, Ganascia turns to consideration of the so-called "BDI" logics, which certainly move generally in the direction of the kind of cognitive expressivity that our cognitive calculi provide. Cointe et al. [30] present a very interesting approach to ethics-based cooperation that makes some use of BDI elements, and even mention in passing the Doctrine of Double Effect, which is central to our simulation below, and which, by virtue of ref. [10], we purport to have fully captured formally. Detailed formalization of ethical principles, and of argumentation by and among multiple artificial agents about such formalizations, is beyond the scope of this work, but – as will be seen below – at the heart of ours. Our work is thoroughgoingly proof-theoretic when it comes to both meaning and to process; in the latter case, we use automated reasoning over customized inference schemata the intensional operators in

12 Readers interested in learning more about the broader challenge of the apparent impossibility of judgment aggregation (i.e., in learning more about AIT$^+$) can consult [42]. Those readers who wish to start with a gradual pedagogical introduction to the broader-than-AIT problem in the form of negative results about judgment aggregation will be well served by reading [77].

13 In fact, the automated reasoner ShadowProver we use for our implementation (see Appendix A2) makes crucial use, during its processing, of *multiple* formal languages. This is true for technical reasons pertaining to the core algorithms of ShadowProver; for more details, see ref. [78].

which far exceed BDI (see Appendix A3). Finally, ref. [31] is a very interesting attempt to capture moral agency using modal logic. This work specifically employs a logic (DL-MA) that is a variant of STIT logic. There are major differences between this work and our paradigm. To mention but two differences from a long, long list: (i) we exploit at many a turn quantification in and over modal formulae, whereas ref. [31] is *propositional* modal logic; and (ii) we reject possible-worlds semantics (and have done so since the first cognitive calculus appeared on the scene in ref. [32]) in favor of semantics based exclusively upon proof-theory/argument-theory, whereas DL-MA is understood in terms of possible worlds.

As will be seen when we turn below to the particular problem scenario and our results, an AI with the capability described in $\mathcal{D}$ is indeed what will prove to be effective. But first we must be clear about the cognitive calculus we use to obtain our results, and about what we take, abstractly and generally, an ethical problem and a solution to it to be.

3 Results

3.1 Cognitive calculus used herein

The cognitive calculus we employ in the present paper is μC, a streamlined modal (specifically epistemic) first-order logic; this calculus is markedly simpler than $\mathcal{DCEC}$, which has been used (as said above) previously to fully model (among other things) robust ethical theories/codes/principles, and allows the capture and computational simulation of ethical reasoning and decision-making over these models. The reason we use here a simpler calculus is that we wish to facilitate and feature the exposition of the key aspects of intelligent argument adudication, unclouded by the (considerable) intricacies of robust cognitive calculi, which are among the most expressive formal logics we are aware of. Please see Appendix A1 for an account of a *cognitive calculus* in general, and Appendix A3 for specification of the cognitive calculus $\mathcal{DCEC}$ and its inductive correlate, $\mathcal{IDCEC}$; the latter, like μC, includes strength factors on epistemic attitudes (e.g., on belief), but in a much fuller way. For an introduction to the more robust calculus $\mathcal{DCEC}$, in a paper that also gives a full formalization of the Doctrine of Double Effect, see the Appendix in ref. [10]. The syntax for μC is given in the grammar shown in (1). The first line of the grammar sets out the conventional *terms* in μC, which are standard (we have variables, constants, and function symbols (denoted by "f") and can be iteratively applied to terms to yield richer terms. Next, well-formedness for formulae is specified. μC supplies standard Boolean formulae, but in addition, and crucially, $\mathbf{B}(a, t, \phi)$ denotes that agent a believes formula ϕ at time t; and $\mathbf{C}(t, \phi)$ denotes that it is common knowledge at time t that ϕ holds; in both formulae t is of course a term.

$$
\begin{aligned}
t &::= x : S \mid c : S \mid f(t_1,\ldots,t_n) \\
\phi &::= \begin{cases} t : \text{Formula} \mid \forall t : \phi \mid \exists t : \phi \\ \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \\ \mathbf{B}(a, t, \phi) \mid \mathbf{C}(t, \phi) \end{cases}
\end{aligned} \tag{1}
$$

The μC calculus lets us formalize, without slipping into inconsistency (a peril explained in ref. [19]), statements of the form "*John believes now that Mary believed that it was raining.*" One formalization of this specimen could be:

$$\exists t < now : \mathbf{B}(john, now, \mathbf{B}(mary, t, holds(raining, t)))$$

Given a set of formulae $\Gamma = \{\phi_1,\ldots,\phi_n\}$, the inference schemata $\mathcal{I}$ of the formal system determine whether certain formulae (plus sometimes meta-logical constraints placed on the schema in question) can be used to derive a given formula ψ; that is, whether or not (where $\vdash$ means provability all and only by inference schemata $I \in \mathcal{I}$):

$$\Gamma \vdash \psi$$

The inference schemata of μC includes the standard inference schemata for first-order logic, with quantification into modal formulae allowed. In addition, we have in μC *modal* inference schemata for inference with beliefs and common knowledge, shown in (2). Inference schemata are of the so-called *natural* kind (originated by Gentzen [33]). Each schema specifies what is required for an inference to be sanctioned (content above the horizontal line); if requirements are satisfied, moving to content below the line is permitted. While cognitive calculi such as $\mathcal{DCEC}$, and even the fragment thereof μC employed herein, make use of what may seem at first glance to be rather intricate inference schemata that are fully formal and hence suitable for use by our automated reasoners, in each case, the core inference is quite intuitively graspable. For instance, in the case of $I_{\mathbf{B}}$, shown immediately below, the schema can be interpreted to say that if an agent a believes at times prior to time t, a collection of m propositions that together can be used to proof ϕ, the agent a believes ϕ at t.

$$
\begin{array}{l}
\mathbf{B}(a, t_1, \phi_1),\ldots, \\
\mathbf{B}(a, t_m, \phi_m), \\
\dfrac{\{\phi_1,\ldots,\phi_m\} \vdash \phi, \quad \Gamma \vdash t_i < t}{\mathbf{B}(a, t, \phi)}[I_{\mathbf{B}}] \\
\dfrac{\mathbf{C}(t, \phi)}{\mathbf{B}(a_1, t,\ldots,\mathbf{B}(a_n, t, \phi))}[I_{\mathbf{C}}]
\end{array} \tag{2}
$$

3.2 Generalized ethical problems

Given some background knowledge Γ, at the core of our approach is an ethical principle ρ. The principle ρ tells us whether performance of the action α is ethically correct (usually, specifically, whether α is ethically *permissible* or *obligatory* or *forbidden*) for agent a at time t in a situation $\sum$. This can be written formally and schematically as shown in (3):

$$\Gamma \cup \Sigma \vdash \rho(a, \alpha, t)? \tag{3}$$

This approach can encapsulate different families of ethical theories, ranging from consequentialist/utilitarian to deontological to virtue ethics and beyond [10,34]. We reveal this in some detail below when we present and discuss the Doctrine of Double Effect, but to give the reader a sense at this point in the present paper as to how the rather abstract form of ρ can work, consider, for example, the standard biconditionals that have long been taken by formally inclined ethicists (see, e.g., the work of Feldman [9]) to capture key parts of ethical theories in the utilitarian family thereof. Specifically, consider the biconditional that for any agent a and any time t, α is obligatory for a if and only if α, among all other options at t for a, a's performing α maximizes happiness among all agents. This biconditional can clearly be expressed as a formula of the form of ρ. The reader will also see that if the biconditional is instead designed to express a "mental" form of utilitarian ethical theory, by, for instance, stipulating that the action is obligatory if and only the agent a here *believes* that α is a happiness maximizer, there will be no problem at all in having formula of the form of ρ do the job, since in accordance with μC we have at our disposal the belief operator $\mathbf{B}$.[14]

3.2.1 Clausification

The formal principle ρ is usually a logicized version of an informal version $\tilde{\rho}$ stated in a natural language. We assume that any such ethical principle ρ can be decomposed into *ethically relevant* clauses $\rho_1,\ldots,\rho_k$ such that the principle holds *iff* (if and only if) the clauses hold. Logically speaking, for any formula ϕ, there are an infinite number of ways to recast ϕ as clauses. We are mainly interested in breaking down ρ into clauses $\rho_1,\ldots,\rho_k$ that match the informal version $\tilde{\rho}$.

Informally:

$$\tilde{\rho} \ \textit{iff} \ \tilde{\rho}_1 \text{ and } \ldots \text{ and } \tilde{\rho}_k$$

Formally:

$$(\Gamma \cup \Sigma) \vdash \begin{pmatrix} \rho_1(a, \alpha, t) \wedge \\ \rho_2(a, \alpha, t) \wedge \\ \vdots \\ \rho_k(a, \alpha, t) \end{pmatrix} \leftrightarrow \rho(a, \alpha, t) \tag{4}$$

3.2.2 Agents

As part of the situation $\sum$, we have a set of agents $\{a_1,\ldots, a_n\}$ each having beliefs about which of the clauses hold. We can decompose $\sum$ into components as shown in equation (5). Our goal is then to ascertain what one particular agent a_{adj}, the adjudicator agent, *ought* to believe at some time t. *For example:*

$$\Gamma \cup \Sigma' \cup \left\{ \begin{array}{c} \mathbf{B}(a_1, t, \rho_2 \wedge \rho_4) \\ \mathbf{B}(a_2, t, \neg \rho_5) \\ \vdots \\ \mathbf{B}(a_n, t, \rho_1) \end{array} \right\} \vdash \mathbf{B}(a_{\text{adj}}, t, \rho)? \tag{5}$$

Each agent a believes a subset of $\beta_a \subseteq \{\rho_1,\ldots, \rho_k\} \cup \{\neg\rho_1,\ldots, \neg\rho_k\}$. Note that we allow agents to be inconsistent. This is useful for representing sensors or agents that are faulty. Our goal is now summarized as:

Goal

Given $\beta_{a_1},\ldots, \beta_{a_n}$ specify a procedure for computing $\mathbf{B}(a_{\text{adj}},t', \rho)$ where $t' > \textit{now}$.

3.3 Solution to a generalized ethical problem

The adjudication framework we present below is based upon an uncertainty system $\mathcal{S}$, so named in part because it uses *strength factors*; this system is detailed in ref. [35]. The prior strength-factor system $\mathcal{S}$, while computing the strength of propositions (i.e., uncertainties) for an agent a, took into account *only* a's beliefs. We hence present now a multi-agent version of $\mathcal{S}$ that takes into account a's beliefs about other agents' beliefs.

$\mathcal{S}$ assumes as a primitive the *reasonableness* operator $(\phi \succcurlyeq_t^{\alpha} \psi)$. The reasonableness operator tells us when an

14 While the calculus μC has no operator for what is obligatory, the aforementioned calculus $\mathcal{DCEC}$ does, and in it we can thus easily build a ρ-form formula such as: $\forall a, t, \alpha : [\mathbf{O}(a, \alpha, t) \leftrightarrow \mathit{Max}(\alpha, a, t)]$.

agent a at time t finds ϕ to be at least as *reasonable* as ψ. In ref. [35], the reasonableness operator was presented in terms of an individual agent's knowledge and information. We present a multi-agent version below. Briefly, everything else being equal, an agent a finds ϕ to be more reasonable than ψ if a believes that more agents believe ϕ than ψ. First, we define a new operator, the *withholding* operator $\mathbf{W}$ (this is "syntactic sugar," in AI parlance):

$$\mathbf{W}(a, t, \phi) \equiv \neg\mathbf{B}(a, t, \phi) \wedge \neg\mathbf{B}(a, t, \neg\phi)$$

We have two modal operators $\{\mathbf{B}, \mathbf{W}\}$. Let Θ and Ω be variables denoting one of the two modal operators $\{\mathbf{B}, \mathbf{W}\}$. Then:

Multi-agent reasonableness

$$\Theta(a, t, \phi) \succcurlyeq_t^a \Omega(a, t, \psi) \equiv \mathbf{B}\left(a, t, \forall a_i : \begin{pmatrix} \Omega(a_i, t, \psi) \\ \rightarrow \\ \exists a_j : \Theta(a_j, t, \phi) \end{pmatrix}\right)$$

The definition immediately above is written in μC and states that for every agent a_i that has an Ω formula in ψ, there is an agent a_j that has a Θ formula in ϕ. Using this operator, we can derive the four discrete uncertainty levels as shown immediately below.

Level 1 Agent a believes at least one other agent a_i believes that ϕ:

$$\mathbf{B}^1(a, t, \phi) \equiv \begin{pmatrix} \mathbf{B}(a, t, \mathbf{B}(a_i, t, \phi)) \\ \wedge \\ \mathbf{B}(a, t, a \neq a_i) \end{pmatrix}$$

Level 2 Agent a believes that it is more reasonable to believe ϕ than withhold ϕ:

$$\mathbf{B}^2(a, t, \phi) \equiv \begin{pmatrix} \mathbf{B}(a, t, \phi) \succcurlyeq_t^a \mathbf{W}(a, t, \phi) \\ \wedge \\ \mathbf{B}^1(a, t, \phi) \end{pmatrix}$$

Level 3 Agent a believes that it is more reasonable to believe ϕ than believe $\neg\phi$:

$$\mathbf{B}^3(a, t, \phi) \equiv \begin{pmatrix} \mathbf{B}(a, t, \phi) \succcurlyeq_t^a \mathbf{B}(a, t, \neg\phi) \\ \wedge \\ \mathbf{B}^2(a, t, \phi) \end{pmatrix}$$

Level 4 Agent a believes that every agent believes ϕ.

$$\mathbf{B}^4(a, t, \phi) \equiv \mathbf{B}(a, \mathbf{B}(a_i, t, \phi)); \text{ for every agent } a_i$$

3.4 Instantiation of the generalized problem: a scenario

We now, as promised, describe an ethically charged scenario, the solution of which will require AI capable of adjudicating inconsistent beliefs on the part of other artificial agents regarding propositional content crucial to a certain ethical principle. In short, the AI here faces an ethical problem in a multi-agent context.

The scenario is as follows. A NATO military squad acquires intel that an old hospital building is being used by terrorists to prepare for an attack on civilians. However, as it was originally a hospital, there is a possibility that there are still civilians inside. The squad wants to determine whether or not they should destroy the building.

The squad therefore utilizes several robotic systems, including high- and low-altitude drones and wall-penetrating radar[15] to look for evidence of people inside the building. The difficulty arises when the devices report inconsistent information regarding the presence of people inside the building.

The squad has an adjudicator agent a_{adj}. The agent a_{adj} relies on the Doctrine of Double Effect ($\mathcal{DDE}$), a well-known ethical principle that lies at the heart of the Occidental tradition of the so-called "just war." Anything here approaching full explication of $\mathcal{DDE}$ is infeasible, due to current space constraints; but we nonetheless very briefly reprise a robust treatment we have provided elsewhere ref. [10], and we direct readers wishing a very extensive essay on $\mathcal{DDE}$ to ref. [36]. $\mathcal{DDE}$ assumes that we have a utility or goodness function for states of the world, including states that are consequences of actions. For an agent a, an action α in a situation $\sum$ at time t is said to be $\mathcal{DDE}$-compliant *iff*:

- ρ_1 The action α by itself is not ethically forbidden (i.e., the action should be morally neutral or above neutral in an ethical hierarchy for deontic operators, such as the one given elsewhere by Bringsjord [37]);
- ρ_2 The net utility or goodness of the α in the situation is greater than some positive amount γ;
- ρ_3 The agent performing α intends only the good effects from this action;
- ρ_4 The agent does not intend any of the bad effects from α;

15 Such as that developed by Lumineye, LLC: https://www.lumineye.com. See: https://taskandpurpose.com/military-tech/army-technology-see-through-walls for more information.

ρ_5 The bad effects are not used as a means to obtain the good effects.

The action α in our scenario is the act *destroying the building*. The possible good effects are that an attack on civilians will be averted. The possible bad effects are that there will be loss of life and there might be civilians in the building who might be harmed.

Often in scenarios where $\mathcal{DDE}$ has to be employed, the clause that is most under scrutiny is ρ_2. This is the only clause that depends on our scenario. Clause ρ_1 is about the action of blowing up a structure. As a matter of empirical fact, this action is generally not forbidden by itself (unlike other actions, such as using biological weapons). $\mathcal{DDE}$ is dependent upon the state of the agent executing the action; clauses ρ_3, and ρ_4 reflect this. Finally, ρ_5 is about the cause-and-effect structure of the action: the bad effects of the action should not be used to cause the good effects; this can be decided by relying upon prior knowledge of the world, and we leave details regarding this aside. Hence, we are left with a focus upon only ρ_2; this clause has to be adjudicated based on possibly different sensory information by a diverse array of agents. Therefore, in the elaboration of the scenario given momentarily, the adjudicator a_{adj} only considers different reports regarding ρ_2. A much more detailed discussion of the clauses of $\mathcal{DDE}$, in connection not with a miliary situation but rather a railroad one, in which a version of the event calculus is employed, can be found in ref. [10].

Equation (6) shows the formalization of ρ_2 in $\mathcal{DCEC}$, which uses an adapted[16] version of the *event calculus* to represent time and change in the physical world. The event calculus has *actions/events* to represent change and *fluents* to represent physical states of the world [38,39]. Fluents are initiated or terminated through actions/events. Fluents that are initiated by action α carried out by agent a at time t are represented by $\alpha_I^{a,t}$, and terminated fluents are represented by $\alpha_T^{a,t}$. $\mu(f, y)$ represents the utility of a fluent f at time y. We are generally interested in modeling utility till some horizon $H > t$. Given these definitions, we can unpack state ρ_2 as given in equation (6) like this:

$$\rho_2 \equiv \sum_{y=t+1}^{H} \left(\sum_{f \in \alpha_I^{a,t}} \mu(f, y) - \sum_{f \in \alpha_T^{a,t}} \mu(f, y) \right) > \gamma \qquad (6)$$

16 For example, the axioms of the event calculus are taken as common knowledge in most work with $\mathcal{DCEC}$, which means that where ϕ is such an axiom, the common-knowledge operator **C** applies to ϕ.

Table 1: AI agents in the scenario

Agent	Description
hdrone	High-altitude drone
$ldrone_1$	Low-altitude drone (faulty)
$ldrone_2$	Low-altitude drone (fixed)
radar	Wall-penetrating radar

3.4.1 The challenge met

The artificial agents in the scenario are listed in Table 1. These agents report to the adjudicator agent their judgments regarding ρ_2. For reasons canvassed above, in the scenario, the adjudicator only needs to determine whether ρ_2 holds.

We now formalize the scenario using μC. To start, we formalize the query which the adjudicator knows will lead to deciding whether ρ_2 holds. That is:

> Are there people inside the building who are planning an attack *and* are there no civilians inside?

This can be expressed using the following formula:

$$\exists p(\mathit{Inside}(p, \mathit{building}) \wedge \mathit{PlanningAttack}(p)) \wedge \forall p(\mathit{Inside}(p, \mathit{building}) \rightarrow \neg \mathit{Civilian}(p))$$

However, what we would really like is a utility based on what subset of the query each agent believes is satisfied. To that end, Table 2 indicates the utility provided by the satisfaction of each formula.

That is, determining that there are terrorists and there are no civilians inside the building gives a utility of γ. Determining that there are no terrorists inside gives a utility of 0. Finally, if there are civilians inside (regardless of whether or not terrorists are inside), the utility is $-\gamma$.

Next, we walk through how this scenario could play out based on what each agent perceives and what beliefs they subsequently infer.

First, a high-altitude drone (*hdrone*) scans the building but cannot detect any humans inside.[17] Because this

17 Strength factors that modulate cognitive attitudes, specifically here the epistemic attitude of *belief*, are crucial for handling partial observability in logicist fashion – and it is partial observability that the low-level sensing agents such as *hdrone* must deal with. In, for example, the seminal work of Barwise and Etchemendy in connection with their Hyperproof system, observability of objects in the microworld they used can be partial (because objects can be occluded by other objects), but since no precise reasoning (by a human observer or by the execution of the system's own code to reason) is allowed over

Table 2: Utility (w.r.t. ρ_2) of the satisfaction of formulae

Utility	Formula
γ	$\exists p\,(Inside(p, building) \wedge PlanningAttack(p))$ $\wedge\forall p\,(Inside(p, building) \rightarrow \neg Civilian(p))$
0	$\neg\exists p\,(Inside(p, building) \wedge PlanningAttack(p))$
$-\gamma$	$\exists p\,(Inside(p, building) \wedge Civilian(p))$

drone has been preengineered for purposes of carrying out such scans, and is state-of-the-art in this regard (details beyond our scope), a fact it knows about itself, it therefore by deduction believes after its scan that there is no one inside the building.[18]

$$\mathbf{B}(hdrone, t_0, \neg\exists p\ \ Inside(p, building)) \tag{7}$$

Next, using background information, the adjudicator then derives the following:

$$\mathbf{B}(adj, t_1, \mathbf{B}(hdrone, t_0, \neg\rho_2))$$

To get a better look, a low-altitude drone ($ldrone_1$) is deployed to scan the building, but triggers a bug when scanning someone walking through a doorway, incorrectly detecting that there is a person who is inside and not inside the building simultaneously.

$$\mathbf{B}\left(ldrone_1, t_1, \exists p\ \begin{pmatrix} Inside(p, building) \wedge \\ \neg Inside(p, building) \end{pmatrix}\right) \tag{8}$$

Using background information, the adjudicator then derives the following:

$$\mathbf{B}(adj, t_2, \mathbf{W}(ldrone, t_1, \rho_2))$$

Finally, the squad activates a soldier equipped with wall-penetrating radar (*radar*) which is able to detect two people inside. It also notices that the occupants are standing near a desk, and seem to be assembling a weapon. This generates a belief that the people inside are planning an attack (and are therefore not civilians).

$$\begin{aligned}&\mathbf{B}(radar, t_2, \exists p\ \ Inside(p, building))\\ &\wedge \mathbf{B}(radar, t_2, \exists p\ \ (Inside(p, building)\\ &\wedge PlanningAttack(p))\\ &\wedge \forall p\ \ (Inside(p, building) \rightarrow \neg Civilian(p)))\end{aligned}$$

Once again, using background information, the adjudicator then derives the following:

$$\mathbf{B}(adj, t_3, \mathbf{B}(radar, t_2, \rho_2))$$

The squad then decides to apply a quick patch to the low-altitude drone ($ldrone_2$) and redeploy it. It is able to see inside a window, and determines that the men are actually civilians, and what appeared to be a weapon was actually a car engine.

$$\begin{aligned}&\mathbf{B}(ldrone_2, t_2, \exists p\ \ Inside(p, building))\\ &\wedge \mathbf{B}\left(ldrone_2, t_3, \exists p\ \begin{pmatrix} Inside(p, building) \wedge \\ Civilian(p) \end{pmatrix}\right)\end{aligned}$$

Finally, the adjudicator arrives at the following:

$$\mathbf{B}(adj, t_4, \mathbf{B}(ldrone, t_3, \neg\rho_2))$$

Figure 3 shows an overview of the situation. The different agents in the scenario, what they report to the adjudicator, the adjudicator's belief about the agents' beliefs (outer belief operator removed for clarity) and the adjudicator's belief is shown in that figure.

Table 3 summarizes how the adjudicator's belief uncertainty changes as the various agents report their beliefs. At the end of the scenario (time t_4), the adjudicator a_{adj} holds a belief at level 3 that ρ_2 does not hold. Therefore, a_{adj} believes that not all of the clauses of $\mathcal{DDE}$ are satisfied; hence, the detonation of the building is not $\mathcal{DDE}$-compliant and cannot be ethically sanctioned. Details about the implementation of this scenario are given in Appendix A2.

3.4.2 Reflecting on the "four steps"

To conclude our discussion of the case study, we briefly reflect upon our progress in implementing our "Four Steps," as we created a moral machine for this scenario.

The first step is fully complete: we selected the Doctrine of Double Effect/$\mathcal{DDE}$ as a principle in our ethical code (derived from both deontological and consequentialist families of ethical theories). The second step, given what we have done, is partially complete. We selected ShadowProver as the automated reasoner with which to implement the ethical reasoning in question. We confessedly lay claim to having implemented *part* of the reasoning in ShadowProver (see Appendix A2),

belief and knowledge that is affected by limited observability, machinery for belief and knowledge, including such machinery that represents graded belief, is entirely absent the Hyperproof system. In command-and-control challenge scenarios such as the one we consider and solve momentarily, we don't have the luxury of avoiding this machinery: it is needed for our solution, we have it, and we use it.

18 While as we say it's out of scope, fuller formalization would bring to bear our prior methodologies for enabling AIs and cognitive robots to reason about their own capabilities in cognitive calculi that include the "self-consciousness" operator $*$. See e.g., ref. [73].

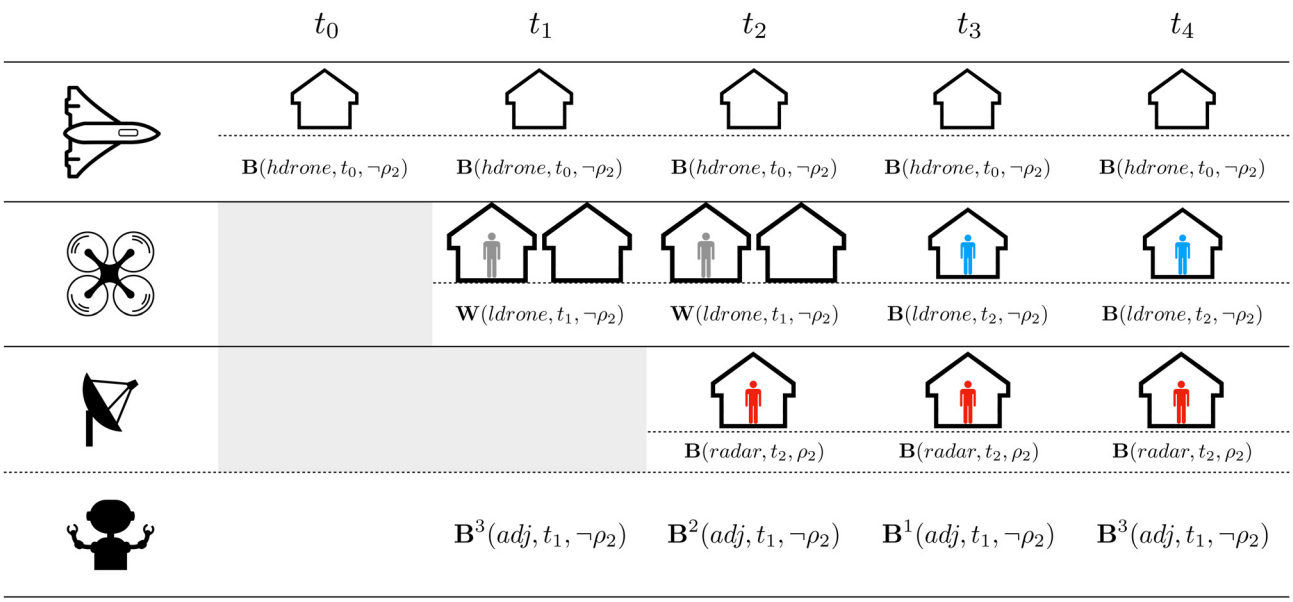

Figure 3: Overview of the scenario.

and leave more robust implementations to future work. The third step is also admittedly incomplete for this particular case study; however, full completion would not require any new research: one would simply follow the processes outlined in ref. [8,13]. Finally, Step 4 is also only partially finished, as its full completion is precluded by the merely partial implementation of Step 3. Overall, though, it should be clear that our Four Steps have been followed.

4 Discussion

We are under no such illusion as that our work will be embraced immediately by all. In general, we at this point anticipate two general classes of objections: one that contains technical worries, and a second aimed at alleged fundamental flaws in logicist AI, at least as such AI is pursued by us. In what now follows, in conformity with this two-part division, we first discuss a class of objections that relate to limitative theorems due, at least originally, to Arrow; and then, we proceed to present and rebut objections that claim our methodology is missing something crucial.

4.1 Dialectic arising from arrow's impossibility theorem and successors

In point of fact, there is no denying that Arrow's Impossibility Theorem (AIT) is directly relevant, logico-mathematically *and* implementation-wise, to our framework and technology for adjudication in multi-agent contexts. However, we cannot expect our readers in the present case to be familiar with AIT (very nicely presented and proved in ref. [40], and ably summarized without proof in ref. [41]). Hence, we must find a shortcut here; and we do, as follows. We can without loss of generality at the current juncture take AIT to be based upon the existence of n artificial agents $a_1,\ldots, a_n$ whose action repertoire consists solely in each of

Table 3: Overview of the beliefs. The adjudicator's beliefs about other agents' beliefs and its uncertainty level in $\neg\rho_2$

Time	hdrone	ldrone	Radar	Strength for ($adj, t, \neg\rho_2$)
t_1	$\mathbf{B}(hdrone, t_0, \neg\rho_2)$	Not considered	Not considered	$\mathbf{B}^3(\text{adj}, t_1, \neg\rho_2)$
t_2	$\mathbf{B}(hdrone, t_0, \neg\rho_2)$	$\mathbf{W}(ldrone, t_1, \neg\rho_2)$	Not considered	$\mathbf{B}^2(\text{adj}, t_2, \neg\rho_2)$
t_3	$\mathbf{B}(hdrone, t_0, \neg\rho_2)$	$\mathbf{W}(ldrone, t_1, \neg\rho_2)$	$\mathbf{B}(radar, t_2, \rho_2)$	$\mathbf{B}^1(\text{adj}, t_3, \neg\rho_2)$
t_4	$\mathbf{B}(hdrone, t_0, \neg\rho_2)$	$\mathbf{B}(ldrone, t_2, \neg\rho_2)$	$\mathbf{B}(radar, t_2, \rho_2)$	$\mathbf{B}^3(\text{adj}, t_4, \neg\rho_2)$

them reporting to some "overseeing" artificial agent a^* their respective preference $\mathfrak{p}_i$ at some fixed time. If we let $\mathcal{P}$ represent a set of attributes that are generally seen as desirable for the agents a_i, we can further stipulate that each agent has these properties (i.e., $\mathcal{P}(a_i)$). In addition, for AIT, we must say that a^* isn't allowed to be "dictatorial"; that is, approximately but not at all inaccurately, a^* isn't allowed to just lay down the law as to what is to be done, irrespective of the recommendations from the a_i. We can abbreviate this condition as "$\mathcal{D}(a^*)$." Given this, AIT can be taken to be this simple (material) conditional:

$$\text{AIT: } [\mathcal{P}(a_i) \wedge \neg\mathcal{D}(a^*)] \rightarrow \bot.$$

Put simply, this conditional says that if all the "recommender" AIs have the desirable properties in question, and they don't report to a dictator, but rather a reasonable aggregator, then outright contradiction ensues. Since a contradiction is an impossibility (at least in anything like a classical logico-mathematical venue), AIT can be said to point out an outright impossibility.

Given this exposition, we are confident that the reader can appreciate how skeptics might see direct relevance between AIT and the framework for multi-agent ethical decision-making we have presented. As a matter of empirical fact, we have had expressed directly to us the sentiment that our framework runs afoul of AIT. But does it? No, it does not, as we now briefly explain.

4.1.1 "But arrow's impossibility theorem (AIT) make the engineering of such an AI impossible!"

Here, now, is the first objection: "You have a framework, undeniably, in which n artificial agents are making recommendations to the AI you see as innovative, and efficacious. I see no harm is labeling your innovative AI as a^*. But your a^* is not allowed to be a dictator, for otherwise why would you even need the AI that this agent brings to bear in the first place? Hence, on the assumption that the agents involved in your scenario above, and in general in the analysis of information relevant to solving an ethical problem (as you have defined such a thing), have these desirable properties, your paradigm lapses into inconsistency."

It is actually easy for us to "surmount" AIT. The conditional that is AIT only applies when the subsidiary agents recommend preferences as their only available actions, and when the overseeing AI designed to adjudicate is not a dictator. But what we have presented, as a matter of technical facts, blocks the applicability of the conditional, for two general reasons. First, we insist that subsidiary agents do much, much more than express preferences (among other things, they communicate full-blown formulae, and proofs/arguments in support of those formulae). Second, we have been careful not to allow our overseeing AI to be anything like a dictator. We would prefer to call such an AI in the case of our framework a "benevolent dictator," or even better, a "philosopher king." To justify this, we summarize now the definition of the "dictator" in Arrow's framework, by drawing directly from ref. [40]:

> An agent a such that "if [a] likes candidate X best, then X is necessarily elected, regardless of how others feel about X." (ref. [40, p. 46])

In our framework, the adjudicator AI agent, at least essentially, has the property described in this quote. Hence, no violation of the conditional that AIT has occurred. However, our overseeing AI does not blindly make choices, disregarding the "voting" AIs. On the contrary, it processes their arguments within a formal system and makes its decision based on the strength of each agent's arguments.

Given that cognoscenti may well be among our readers, for good measure, we now consider another objection along the AIT line of thought.

4.1.2 "What about similar impossibility results in judgment aggregation?"

We anticipate this contrarian reasoning: "Okay, understood; points well-taken. Yet Dietrich and List [42] (among others) prove two similar impossibility results, ones involving not just simple preferences from subsidiary agents, but declarative information of the sort that is your bread and butter. D&L prove conditionals similar to the one you conveniently gave above: if a set of desirable properties is instantiated in a group of subsidiary agents who convey declarative content to a non-dictatorial but presiding agent charged with issuing judgement on the basis of this content, a contradiction deductively follows!"

Fortunately, our adjudication AI enables us to circumvent these results as well. How? Well, another feature of the work presented above comes to the rescue. Within our framework, the subsidiary agents a_i do not simply convey propositions with an assignment of either true or false. Instead, our a_i must provide some logically valid argument justifying their assertions (which may themselves be only likely or unlikely at a certain level); and our adjudicator must take into account this supplied reasoning when making its decision. For instance, specifically,

the adjudicator playing the role of $a^\star$ may disregard the declarative assertion of an a_i if another a_j voter presents a competing assertion with an argument of higher likelihood. This is in fact precisely the sort of situation seen in our scenario above (in Section 3.4).

4.2 Discussion of the "missingness" of formal logic relative to the realm of human ethics

4.2.1 "But emotions are central to morality and are beyond logic!"

We expect some skeptics to react to our approach to machine/robot ethics in general, and dynamic adjudication in particular, as follows:

> Since you yourselves point out that ethics in the human sphere, including in particular the practice of professional ethics, is a largely declarative affair, and hence suitable for what might be called 'logicization,' certainly you must agree as well that human -level ethics is veritably saturated with emotion. Ethicists, after all, aren't robots. And everyday morality is about real human beings – beings who, whether or not they are themselves ethicists paid to adopt a detached, clinical mindset, wisely base their own ethical decisions (at least in part) upon emotions. It should for instance be obvious to you that mistreating human beings is often judged immoral because of an understanding that victims too have emotions; hence empathy plays a crucial role. Even the simple facts that humans know they have desires, and know other humans do as well, and *also* know that quite often when desires are unfulfilled that causes other, unwanted emotions – these facts undergird morality in the human sphere.

Of course, we are humans ourselves, and readily accept that the phenomena here cited are real, and real important in the realm of ethics. At the same time, there simply is no good reason to maintain that artificial agents, including cognitive robots, actually have emotions. Yet, there is also no good reason to deny that the *abstract structure* of emotions can't be captured in formal logic; indeed, see, for instance, ref. [43]. Once logic captures some phenomena, which allows automated reasoning to then be deployed to support decision-making informed by that which is captured, engineering that benefits human beings is enabled. The mere fact that the world wants AI that can communicate in natural language with humans, despite the fact that, at least as the lead author of the present paper holds [44], AI can't really understand natural language in a human way, vindicates the approach we set out in the present paper.

4.2.2 "Well, but even if the emotions, structurally speaking, can be captured, consciousness, key to morality, can't!"

We fully expect the previous objection, despite our reply, to be sustained in the following form: "Thank you for your response. But it's one thing to say, as you are saying, that for instance a theory of emotions can be captured in formal logic, but surely it's quite another to say such a thing about *consciousness* in general. Yet how can we have a morally competent agent that isn't conscious? That doesn't seem to make any sense."

We are quite sympathetic. We agree that, for instance, so-called *phenomenal consciousness*, which includes what philosophers call "qualia," is not possessed by artificial agents, even when they are embodied as robots. But to echo our reply immediately above, that doesn't mean that the logico-mathematics of consciousness can't be captured in formal logic. A physicist like Einstein may in a real sense *feel* what simultaneity is, but that doesn't mean that relativity can't be captured in formal logic.[19] As a matter of fact, we expect that our own axiomatization of (cognitive) consciousness [45], combined with our theory, Λ, for the measurement of levels of cognitive consciousness [46], will in subsequent work be married to the formalisms and techniques we present in the present paper for machine-ethics in multi-agent environments.

4.2.3 "But even leaving consciousness aside, aren't you after genuine moral agents?"

The next anticipated objection flows from the preceding one and is expressed thus:

> Learning consciousness aside, I take your overarching, longer-term goal here to be the construction of genuine ethical/moral artificial agents (although it seems to me that the framework presented here would be perfectly functional outside of the domain of ethical reasoning). But then I struggle with the kinds of problems your framework is here aimed at – such as challenges in command-and-control, where there are multiple adjudicator agents responsible for interpreting, processing, and acting upon information pertinent to morally-charged situations. If your artificial adjudicators are to be the moral agents of folk psychology or philosophy of action, it seems they aren't actually agents at all: they are after all determined at the level of the operating system. In some sense, they are something like ethical "zombies." Now just maybe we humans are as well, but we certainly *believe* that we aren't, and there

19 In fact it can. See, e.g., ref. [79,80].

> hasn't been any profound scientific advance to suggest that we're wrong. So, given that the space of cognitive calculi is large, why not turn these attitudes about our genuine agency into full-blown artificial agents with beliefs about aspects of themselves and a folk psychology (of the sort that for instance Jerry Hobbs has pursued), so that you move closer to something like responsible agency, even if illusory?

This is a penetrating, profound concern/objection; however, it doesn't derail the research and development we present in the present paper; in fact, the concern catalyzes clarification of the nature of our work. When the interlocutor here refers to the work of J. Hobbs, no reference is given; but the recent [47] is in fact a polished compilation of longstanding effort on the part of Hobbs and collaborators to capture folk psychology (i.e., – to conveniently use the subtitle of the Hobbsian book we have cited – "how people think people think"). We applaud this work, but the nature of the efforts, formalisms, and implementation we report in the present paper differ fundamentally from what Hobbs et al. are doing, both methodologically and topically speaking. Methodologically speaking, what we are doing herein is formalizing not part of folk psychology, but part of precise, rational cognition, with – hopefully – none of the well-known deficiencies and biases of how "people think" in general. In addition, while Hobbs et al. at the end of the day base all their formal modeling on first-order logic, we find an exclusive focus on this simple logic to be crippling (which is reflected in d_6 in $\mathcal{D}$). Topically speaking, what we are doing herein is in line with an attempt to capture in formal logic what humans doing first-rate work in the formal sciences do when they discover and prove results in a manner that can be certified correct. We would thus start not with cognition seen "on the street," but with cognition seen in accomplished mathematicians – cognition that is now known in the discipline of reverse mathematics to exceed even second-order logic (see, e.g., ref. [48]). So, summing up, we do intend our artificial agents to be moral agents (and if the absence of subjective consciousness and qualia makes them "zombies," then so be it), but they must be, relative to the average human case, precise, rational, certifiably correct ones.

4.3 "But you leave aside the perception problem!"

Here is another objection we anticipate: "But your formalisms and algorithms and automated-reasoning capability is only brought to bear, at least as I understand them, *after* perception has happened. But perception is a huge challenge; arguably its severity rivals any of the challenges that you purport to have solved. Hence, your machine-ethics work, including specifically work – described above, of course – targeted at handling the need to adjudicate competing arguments dynamically in ethically charged scenarios, is all quite vulnerable. In fact I would respectfully go so far as to say that this work has, if you will, an Achilles heel."

We are painfully aware of the fact that human-level perception in AIs and, specifically, robots is nearly at the level of science fiction. This is in large measure why the threat of "killer robots," as pointed out in ref. [49],[20] is not yet to be taken too seriously. However, we took pains at the outset of the present paper to point out that (i) artificial agents in general (including, then, cognitive robots as we defined them) compute functions mapping percepts to actions, but that (ii) logicist artificial agents, the class our own fall into, cast this computation as automated reasoning *that only starts in earnest after the transduction of sense data into formulae in a cognitive calculus*. This basic conception of the overall pipeline is in line with how logicist AI has long been characterized by the lead author; see, for example, ref. [50]. This conception is also in line with our conception of cognitive robotics, as we defined this discipline above in our Methods section. In the scenario we gave above that featured our artificial adjudicator and its success, we confessedly took liberties in assuming that perceptual power was on hand, but we did so in order to present our contribution under logicist AI, a contribution that presents to the world formalisms, in precise and even implemented form, that we trust will be integrated with present and future cutting-edge research and engineering of AI/robot perception that is outside our forte.

We now pass to the final section of the paper proper, in which yet another objection is considered.

---

20 They tellingly write:

> For instance, how will the technology [="killer robots"] differentiate enemies from friends in asymmetric wars, where the soldiers don't wear uniforms? More generally, when humans are not able, on the basis of a given set of information, to discriminate cases that meet criteria from cases that don't, how will machines do better? … Will algorithms be able to recognize a particular individual from their facial features, a foe from their military uniform, a person carrying a gun, a member of a particular group, a citizen of a particular country whose passport will be read from a remote device? (ref. [49], p. 90).

4.4 Conclusion and future work

We end by considering what is in effect an additional objection; but because this objection is a direct catalyst of work on our part that is already underway, consideration of the objection makes for a suitable wrap-up, and a pointer to future work. What is this new objection? This: "I believe the three of you must concede that one unavoidable limiting factor afflicting any logic-based automated-reasoning system is the apparent necessity of hand-crafting for it the knowledge that such a system uses. In what you have presented, what your AI knows and believes, and reasons wonderfully from, appears, alas, as if by magic."

We make two points in rejoinder.

First, we plan that our reasoning systems will be integrated with ontologies and knowledge graphs to allow them to harness the epistemic content therein in order to refine and strengthen the arguments generated and managed by these systems. For example, in the military scenario we used to explain our approach, visual information from the low-altitude drone was exploited for an argument that the men were planning an attack. However, it is likely to be nigh impossible to prove from that data that the attack was being planned on the U.S. specifically. Such a narrow proposition, to some level of likelihood, would be the conclusion of an argument provided for instance by intelligence.[21] If such propositional knowledge was available to the low-altitude drone, it could've included it in its argument, which potentially would have strengthened its belief. We are actively working on the extension of our approach in this direction.

The second point we make in rebuttal, and with this we conclude, is that, actually, it is far from obvious that automated reasoning cannot, in and of itself, supplied with percepts, generate new knowledge and belief, which can then be further reasoned over in conjunction with new percepts, and so on as the life of the agent in question continues. In other words, perhaps automated reasoning can serve itself as the chief engine of coming to know, and to believe. This would be, if you will, a sort of "learning *ex nihilo*." We have in fact defined just such a type of learning [51], and future work for us is clearly to imbue adjudicating AIs discussed above with this form of learning.

Acknowledgments: Three anonymous reviewers provided excellent feedback, and we are deeply grateful to them. The authors are also grateful to ONR, both for its support of our current efforts to surmount Arrow's Impossibility Theorem and its descendants here special thanks to Paul Bello, Micah Clark, and Tom McKenna (with a special thanks here to the wise guidance and inspiration of Michael Qin), and for their past support of our R&D in robot/machine ethics (primarily under a MURI to advance the science and engineering of moral competence in robots; PI M. Scheutz, Co-PI B. Malle, Co-PI S. Bringsjord). Finally, we thank AFOSR for their support of our development of cutting-edge automated-reasoning systems for high levels of computational intelligence (special thanks to the long-term guidance and inspiration of James Lawton), for without the representational reach of these systems and their running in the real world on relevant formal content, the situation would be no better than a century back.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: All data generated or analysed during this study are included in this published article.

References

[1] S. Bringsjord and N. S. Govindarajulu, "Artificial intelligence," in *The Stanford Encyclopedia of Philosophy*, E. Zalta, Ed., 2018, Available: https://plato.stanford.edu/entries/artificial-intelligence/.

[2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, *3rd ed.*, Prentice Hall, Upper Saddle River, NJ, 2009.

[3] G. Luger, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, *6th ed.*, Pearson, London, UK, 2008.

[4] S. Bringsjord, "Belief in the singularity is logically brittle," *J. Conscious. Stud.*, vol. 19, no. 7–8, pp. 14–20, 2012.

[5] C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton, NJ, 2003.

[6] H. Levesque and G. Lakemeyer, "Chapter 23: Cognitive robotics," in *Handbook of Knowledge Representation, Foundations of Artificial Intelligence*, Elsevier, Amsterdam, The Netherlands, 2008, pp. 869–886, DOI: https://doi.org/10.1016/S1574-6526(07)03023-4.

[7] P. Bello, *"Toward a logical framework for cognitive effects-based operations: Some empirical and computational results,"* Ph.D. thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY, 2005.

[8] N. Govindarajulu, S. Bringsjord, A. Sen, J. Paquin, and K. O'Neill, "Ethical operating systems," in *Reflections on Programming Systems, Volume 133 of Philosophical Studies*, L. De Mol and G. Primiero, Eds., Springer, Cham, 2018, pp. 235–260, DOI: https://doi.org/10.1007/978-3-319-97226-8_8.

[9] F. Feldman, *Introductory Ethics*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

21 Professionally, not generically, speaking. It is customary to, e.g., speak of the "intelligence community."

[10] N. Govindarajulu and S. Bringsjord, "On automating the doctrine of double effect," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (*IJCAI-17*), C. Sierra, Ed., 2017, pp. 4722–4730, DOI: https://doi.org/10.24963/ijcai.2017/658.

[11] N. S. Govindarajulu, S. Bringsjord, and M. Peveler, "On quantified modal theorem proving for modeling ethics," in *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements, ARCADE@CADE 2019*, Natal, Brazil, August 26, 2019, vol. 311, EPTCS, M. Suda and S. Winkler, Eds., 2019, pp. 43–49, DOI: http://dx.doi.org/10.4204/EPTCS.311.7.

[12] N. S. Govindarajulu, *ShadowProver*, 2016, Available: https://naveensundarg.github.io/prover/.

[13] N. S. Govindarajulu and S. Bringsjord, "Ethical regulation of robots must be embedded in their operating systems," in *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, Cognitive Technologies, R. Trappl, Ed., Springer, Cham, 2015, pp. 85–99, DOI: https://doi.org/10.1007/978-3-319-21548-8_5.

[14] J. Pollock, *How to Build a Person: A Prolegomenon*, MIT Press, Cambridge, MA, 1989.

[15] J. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person*, MIT Press, Cambridge, MA, 1995.

[16] J. L. Pollock, "How to reason defeasibly," *Artif. Intell.*, vol. 57, no. 1, pp. 1–42, 1992, DOI: https://doi.org/10.1016/0004-3702(92)90103-5.

[17] J. Pollock, "Defasible reasoning with variable degrees of justification," *Artif. Intell.*, vol. 133, no. 1–2, pp. 233–282, 2001, DOI: https://doi.org/10.1016/S0004-3702(01)00145-X.

[18] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," *Artif. Intell.*, vol. 77, pp. 321–357, 1995, DOI: https://doi.org/10.1016/0004-3702(94)00041-X.

[19] S. Bringsjord and N. S. Govindarajulu, "Given the web, what is intelligence, really?" *Metaphilosophy*, vol. 43, no. 4, pp. 464–479, 2012, DOI: https://doi.org/10.1111/j.1467-9973.2012.01760.x.

[20] H. Prakken and G. Vreeswijk, "Logics for defeasible argumentation," in *Handbook of Philosophical Logic*, *vol. 4*, D. Gabbay and F. Guenthner, Eds., Springer, Dordrecht, The Netherlands, 2001, pp. 219–318, DOI: https://doi.org/10.1007/978-94-017-0456-4_3.

[21] S. Modgil and H. Prakken, "The *ASPIC*$^+$ framework for structured argumentation: A tutorial," *Arg. Comput.*, vol. 5, no. 1, pp. 31–62, 2014, DOI: https://doi.org/10.1080/19462166.2013.869766.

[22] F. Cerutti, S. A. Gaggl, M. Thimm, and J. Wallner, "Foundations of implementations for formal argumentation," in *The IfCoLog Journal of Logics and their Applications, Special Issue Formal Argumentation*, P. Baroni, D. Gabbay, M. Giacomin, and L. Van der Torre, Eds., College Publications, vol. 4, no. 8, pp. 2623–2705, 2017.

[23] K. Arrow, *Social Choice and Individual Values*, *2nd ed*., Cowles Foundation Monographs Series, Wiley, New York, NY, 1963.

[24] P. Dunne, P. Marquis, and M. Wooldridge, "Argument aggregation: Basic axioms and complexity results," in *Computational Models of Argument*, vol. 245, Series Frontiers in Artificial Intelligence and Applications, B. Verheij, S. Szeider, and S. Woltran, Eds., IOS Press, Amsterdam, The Netherlands, 2012, pp. 129-140, DOI: https://doi.org/10.3233/978-1-61499-111-3-129.

[25] U. Endriss and U. Grandi, "Graph aggregation," *Artif. Intell.*, vol. 245, pp. 86–114, 2017, DOI: https://doi.org/10.1016/j.artint.2017.01.001.

[26] T. Bench-Capon, "Persuasion in practical argument using value based argumentation frameworks," *J. Log. Comput.*, vol. 13, no. 3, pp. 429–448, 2003, DOI: https://doi.org/10.1093/logcom/13.3.4294.

[27] E. Awad, R. Booth, F. Tohmé, and I. Rahwan, "Judgment aggregation in multi-agent argumentation," *J. Log. Comput.*, vol. 27, no. 1, pp. 227–259, 2017, DOI: https://doi.org/10.1093/logcom/exv055.

[28] S. Coste-Marquis, C. Devred, S. Konieczny, M. C. Lagasquie-Schiex, and P. Marquis, "On the merging of Dung's argumentation systems," *Artif. Intell.*, vol. 171, no. 10–15, pp. 730–753, 2007, DOI: https://doi.org/10.1093/logcom/exv055.

[29] J.-G. Ganascia, "Non-monotonic resolution of conflicts for ethical reasoning," in *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, Cognitive Technologies, R. Trappl, Ed., Springer, Cham, Switzerland, 2015, pp. 101-118, DOI: https://doi.org/10.1007/978-3-319-21548-8_6.

[30] N. Cointe, G. Bonnet, and O. Boissier, "Ethics-based cooperation in multi-agent systems," in *Advances in Social Simulation, Springer Proceedings in Complexity*, H. Verhagen, M. Borit, G. Bravo, and N. Wijermans, Eds., Springer, Cham, 2020, pp. 101–116, DOI: https://doi.org/10.1007/978-3-030-34127-5_10.

[31] E. Lorini, "On the logical foundations of moral agency," in *Deontic Logic in Computer Science, DEON 2012, Lecture Notes in Computer Science, vol. 7393*, T. Ågotnes, J. Broersen, and D. Elgesem, Eds., Springer, Berlin, Heidelberg, 2012, pp. 108–122, DOI: https://doi.org/10.1007/978-3-642-31570-1_8.

[32] K. Arkoudas and S. Bringsjord, "Propositional attitudes and causation," *Int. J. Softw. Inform.*, vol. 3, pp. 47–65, 2009.

[33] G. Gentzen, "Untersuchungen über das logische Schlieben I," *Math. Zeitschrift*, vol. 39, pp. 176–210, 1935.

[34] N. S. Govindarajulu, S. Bringsjord, R. Ghosh, and V. Sarathy, "Toward the engineering of virtuous machines," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 29–35, DOI: https://doi.org/10.1145/3306618.3314256.

[35] N. S. Govindarajulu and S. Bringsjord, "Strength factors: An uncertainty system for quantified modal logic," in *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning"* (*LFU-2017*), *Melbourne, Australia*, V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade, and G. Qi, Eds., 2017, pp. 34–40, Available: https://arxiv.org/abs/1705.10726.

[36] A. McIntyre, "The doctrine of double effect," in *The Stanford Encyclopedia of Philosophy*, E. Zalta, Ed., 2004/2014, Available: https://plato.stanford.edu/entries/double-effect/.

[37] S. Bringsjord, "A 21st-century ethical hierarchy for humans and robots: EH," in *A World With Robots: International Conference on Robot Ethics* (*ICRE 2015*), I. Ferreira, J. Sequeira, M. Tokhi, E. Kadar, and G. Virk, Eds., Springer, Berlin, Germany, 2015, pp. 47–61.

[38] M. Shanahan, "The event calculus explained," in *Artificial Intelligence Today, Lecture Notes in Computer Science* (*Lecture Notes in Artificial Intelligence*), M. Wooldridge and M. Veloso, Eds., Springer, Berlin, Heidelberg, 1999, vol. 1600, pp. 409–430, DOI: https://doi.org/10.1007/3-540-48317-9_17.

[39] E. T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach*, *2nd ed*., Morgan Kaufmann, San Francisco, CA, 2014.

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

[41] M. Morreau, "Arrow's theorem," in *The Standford Encyclopedia of Philosophy*, E. Zalta, Ed., Winter 2016 ed., 2014, Available: http://plato.stanford.edu/entries/arrows-theorem/.

[42] F. Dietrich and C. List, "Arrow's impossibility theorem in judgment aggregation," *Soc. Choice Welf.*, vol. 29, no. 1, pp. 19–33, 2007, DOI: https://doi.org/10.1007/s00355-006-0196-x.

[43] B. Steunebrink, M. Dastani, and J.-J. Meyer, "A logic of emotions for intelligent agents," in *Proceedings of the 22nd National Conference on Artificial Intelligence*, AAAI Press, Palo Alto, CA, 2007, pp. 142–147.

[44] S. Bringsjord and R. Noel, "Real robots and the missing thought experiment in the Chinese room dialectic," in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, J. Preston and M. Bishop, Eds., Oxford University Press, Oxford, UK, 2003, pp. 144–166.

[45] S. Bringsjord, P. Bello, and N. Govindarajulu, "Toward axiomatizing consciousness," in *The Bloomsbury Companion to the Philosophy of Consciousness*, D. Jacquette, Ed., Bloomsbury Academic, London, UK, 2018, pp. 289–324, DOI: https://doi.org/10.5040/9781474229043.0025.

[46] S. Bringsjord and N. Govindarajulu, "The theory of cognitive consciousness, and Λ (lambda)," *J. Artif. Intell. Conscious.*, vol. 7, no. 2, pp. 155–181, 2020, DOI: https://doi.org/10.1142/S2705078520500095.

[47] A. Gordon and J. Hobbs, *A Formal Theory of Commonsense Psychology: How People Think People Think*, Cambridge University Press, Cambridge, UK, 2017, DOI: https://doi.org/10.1017/9781316584705.

[48] S. Simpson, *Subsystems of Second Order Arithmetic*, *2nd ed*., Cambridge University Press, Cambridge, UK, 2010.

[49] J.-G. Ganascia, C. Tessier, and T. Powers, "On the autonomy and threat of 'killer robots'," *Newslett. Philos. Comput.*, vol. 17, no. 2, pp. 87–93, 2018.

[50] S. Bringsjord, "The logicist manifesto: At long last let logic-based artificial intelligence become a field unto itself," *J. Appl. Log*, vol. 6, no. 4, pp. 502–525, 2008, DOI: https://doi.org/10.1016/j.jal.2008.09.001.

[51] S. Bringsjord, N. S. Govindarajulu, J. Licato, and M. Giancola, "Learning *ex nihilo*," in *GCAI 2020 – 6th Global Conference on Artificial Intelligence, Vol. 72, EPiC Series in Computing, International Conferences on Logic and Artificial Intelligence at Zhejiang University* (*ZJULogAI*), EasyChair Ltd, Manchester, UK, 2020, pp. 1–27, Available: https://easychair.org/publications/paper/NzWG.

[52] H. D. Ebbinghaus, J. Flum, and W. Thomas, *Mathematical Logic*, *2nd ed*., Springer-Verlag, New York, NY, 1994.

[53] M. Ashcraft, *Human Memory and Cognition*, HarperCollins, New York, NY, 1994.

[54] E. B. Goldstein, *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience*, *5th ed*., Cengage Learning, Boston, MA, 2018.

[55] G. Gentzen, "Investigations into logical deduction," in *The Collected Papers of Gerhard Gentzen*, M. E. Szabo, Ed., North-Holland, Amsterdam, The Netherlands, 1935, pp. 68–131.

[56] D. Prawitz, "The philosophical position of proof theory," in *Contemporary Philosophy in Scandinavia*, R. E. Olson and A. M. Paul, Eds., Johns Hopkins Press, Baltimore, MD, 1972, pp. 123–134.

[57] P. Schroeder-Heister, "Proof-theoretic semantics," in *The Stanford Encyclopedia of Philosophy*, E. Zalta, Ed., 2012/2018, Available: https://plato.stanford.edu/entries/proof-theoretic-semantics/index.html.

[58] K. Arkoudas and S. Bringsjord, "Toward formalizing common-sense psychology: An analysis of the false-belief task," in *PRICAI 2008: Trends in Artificial Intelligence, Lecture Notes in Computer Science*, T.-B. Ho and Z.-H. Zhou, Eds., Springer, Berlin, Heidelberg, 2008, vol. 5351, pp. 17–29, DOI: https://doi.org/10.1007/978-3-540-89197-0_6.

[59] K. Arkoudas and D. Musser, *Fundamental Proof Methods in Computer Science: A Computer-Based Approach*, *1st ed*., The MIT Press, Cambridge, MA, 2017.

[60] A. S. Rao and M. P. Georgeff, "Modeling rational agents within a BDI-architecture," in *KR'91: Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, R. Fikes and E. Sandewall, Eds., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991, pp. 473–484.

[61] C. Benzmüller and D. Miller, "Automation of higher-order logic," in *Handbook of the History of Logic, Vol. 9: Logic and Computation*, North Holland, Amsterdam, The Netherlands, 2014.

[62] C. Benzmüller and B. W. Paleo, "The inconsistency in Goödel's ontological argument: A success story for AI in metaphysics," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (*IJCAI'16*), S. Kambhampati, Ed., AAAI Press, New York, NY, USA, 2016, pp. 936–942.

[63] S. Shapiro, *Foundations Without Foundationalism: A Case for Second-Order Logic*. Oxford University Press, Oxford, UK, 1991.

[64] S. Bringsjord, J. Licato, and A. Bringsjord, "The contemporary craft of creating characters meets today's cognitive architectures: A case study in expressivity," in *Integrating Cognitive Architectures into Virtual Character Design*, J. Turner, M. Nixon, U. Bernardet, and S. DiPaola, Eds., IGI Global, Hershey, PA, 2016, pp. 151–180.

[65] B. F. Chellas, *Modal Logic: An Introduction*, Cambridge University Press, Cambridge, UK, 1980.

[66] J. Paris and A. Vencovská, *Pure Inductive Logic*, Cambridge University Press, Cambridge, UK, 2015.

[67] S. Bringsjord, J. Taylor, A. Shilliday, M. Clark, and K. Arkoudas, "Slate: An argument-centered intelligent assistant to human reasoners," in *Proceedings of the 8th International Workshop on Computational Models of Natural Argument* (*CMNA 8*), F. Grasso, N. Green, R. Kibble, and C. Reed, Eds., University of Patras, Patras, Greece, 2008, pp. 1–10.

[68] M. Proust, *In Search of Lost Time. The Complete Masterpiece* (Translated by: C. K. Scott Moncrieff, T. Kilmartin, A. Mayor; revised by D. J. Enright), Modern Library/Random House, New York, NY, 2003.

[69] J. -D. Bredin, *The Affair: The Case of Alfred Dreyfus*, Reprint ed. George Braziller Inc., New York, NY, 1986.
[70] G. Smith, *Until the Last Trumpet Sounds: The Life of General of the Armies John J. Pershing*, *1st ed*., Wiley, New York, NY, 1999.
[71] L. Kohlberg, "The claim to moral adequacy of a highest stage of moral judgment," *J. Philos.*, vol. 70, no. 18, pp. 630–646, 1973, DOI: https://doi.org/10.2307/2025030.
[72] G. S. Boolos, J. P. Burgess, and R. C. Jeffrey, *Computability and Logic*, *4th ed*., Cambridge University Press, Cambridge, UK, 2003.
[73] S. Bringsjord, J. Licato, N. Govindarajulu, R. Ghosh, and A. Sen, "Real robots that pass tests of self-consciousness," in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication* (*RO-MAN 2015*), IEEE, New York, NY, 2015, pp. 498–504, DOI: https://doi.org/10.1109/ROMAN.2015.7333698.
[74] J. McCarthy, "Circumscription – a form of non-monotonic reasoning," *Artif. Intell.*, vol. 13, no. 1–2, pp. 27–39, 1980, DOI: https://doi.org/10.1016/0004-3702(80)90011-9.
[75] R. Reiter, "A logic for default reasoning," *Artif. Intell.*, vol. 13, no. 1–2, pp. 81–132, 1980, DOI: https://doi.org/10.1016/0004-3702(80)90014-4.
[76] S. Bringsjord and J. Licato, "By disanalogy, cyberwarfare is utterly new," *Philos. Technol.*, vol. 28, pp. 339–358, 2015, DOI: https://doi.org/10.1007/s13347-015-0194-y.
[77] C. List, "The theory of judgment aggregation: An introductory review," *Synthese*, vol. 187, pp. 179–207, 2012, DOI: https://doi.org/10.1007/s11229-011-0025-3.
[78] N. Govindarajulu, S. Bringsjord, and M. Peveler, "On quantified modal theorem proving for modeling ethics," in *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements* (*ARCADE 2019*), *vol. 311, Electronic Proceedings in Theoretical Computer Science*, M. Suda and S. Winkler, Eds., Open Publishing Association, Waterloo, Australia, 2019, pp. 43–49.
[79] H. Andréka, J. X. Madarász, I. Németi, and G. Székely, "A logic road from special relativity to general relativity," *Synthese*, vol. 186, pp. 633–649, 2012, DOI: https://doi.org/10.1007/s11229-011-9914-8.
[80] N. S. Govindarajalulu, S. Bringsjord, and J. Taylor, "Proof verification and proof discovery for relativity," *Synthese*, vol. 192, pp. 2077–2094, 2015, DOI: https://doi.org/10.1007/s11229-014-0424-3.
[81] M. Nelson, "Propositional attitude reports," in *The Stanford Encyclopedia of Philosophy*, E. Zalta, Ed., 2015, Available at: https://plato.stanford.edu/entries/prop-attitude-reports/.
[82] N. Francez, *Proof-theoretic Semantics*, College Publications, London, UK, 2015.

Appendix A Supplemental data

A.1 What is a cognitive calculus? And why is it so named?

What is a cognitive calculus $\mathscr{C}$, and why is it denoted with the two words in question? (We use "$\mathscr{C}$" here as an arbitrary variable ranging over (the uncountably infinite space of) all cognitive calculi) (As the reader will recall, particular cognitive calculi that were called out and used above included μC, $\mathcal{DCEC}$, and $\mathcal{IDCEC}$. The latter two have their formal languages and inference schemata specified below.) In keeping with the mathematical-logic literature (e.g., ref. [52]),[22] we first take a *logical system* $\mathscr{L}$ to be a triple $\langle \mathcal{L}, \mathcal{I}, \mathcal{S} \rangle$ where $\mathcal{L}$ is a (often) sorted/typed formal language (based therefore on an alphabet and a formal grammar), $\mathcal{I}$ is a set of natural[23] inference schemata, and $\mathcal{S}$ is a formal semantics of some sort. For example, the familiar propositional calculus comprises a family of simple logical systems; the same holds for first-order logic; both families are of course at the heart of AI.[24] In the case of both of these families, a frequently included particular inference schema is *modus ponens*, that is:

$$\frac{\phi \rightarrow \psi, \phi}{\psi} I_{\text{MP}}$$

And in the case of the latter family, often *universal introduction* is included in a given I; a specification of this inference schema immediately follows.[25]

$$\frac{\phi(a)}{\forall x \phi\left(\frac{a}{x}\right)} I_{\text{UI}}$$

Note that both of the two inference schemata just shown are included in the particular cognitive calculi μC and $\mathcal{DCEC}$ we used in the present paper for modeling, and as a framework for automated reasoning. Note as well that both $\mathscr{L}_{\text{PC}}$ (approximately the propositional calculus) and $\mathscr{L}_1$ (= first-order logic) are *extensional*, which means essentially that the meaning of any formula ϕ in the relevant languages are given by compositional functions operating solely on the internal components of ϕ. If we, for example, know that ϕ is FALSE, then we know that the meaning of $\phi \rightarrow \psi$ is TRUE, for any ψ in the language, for both of these logical systems.

Moving from the concept of a logical system to that of a cognitive calculus is straightforward and can be viewed as taking but three steps, to wit:

S1 Expand the language of a logical system to include

(i) Modal operators that represent one or more mental verbs at the human level standardly covered in human-level cognitive psychology (e.g., see any standard, comprehensive textbook on human-level cognitive psychology, such as [53,54]), and regarded to be so-called "propositional attitudes" that give rise to propositional-attitude-reporting sentences, where these sentences are represented by operator-infused formulae in a cognitive calculus.[26] Such verbs include *knowing*, *believing*, *deciding*, *perceiving*, *communicating*,[27] *desiring*, and *feeling X* where "*X*" denotes some emotional state (e.g., possible *X* = *sad*, and so on). Note that such verbs break the bounds of extensionality, and hence make any logic that captures them an *intensional* logic.[28] Step S1(i) is the reason why we speak of a *cognitive* calculus.

(ii) Metalogical expressions (such as that from a set Φ of formulae a particular formula ϕ can be proved: $\Phi \vdash \phi$). Hence, cognitive calculi are not merely object-level elements of logics, but include metalogical elements as well. (This feature in the case of $\mathcal{DCEC}$ is required to fully capture the Doctrine of Double Effect/$\mathcal{DDE}$, central to what we did above, of course.) For example, a cognitive calculus can have a metaconditional saying that if some provability expression such as $\Phi \vdash \phi$ holds, then ϕ holds

22 Note in particular coverage in this excellent work of Lindström's Theorems, which pertain to the properties of certain logical systems (e.g., completeness).

23 Hence, when the schemata are deductive in nature, we specifically have natural deduction.

24 As can be confirmed by looking to the main textbooks of the field. For example, see ref. [2,3].

25 The standard provisos apply here to the constant a.

26 The attitudes are covered nicely in ref. [81]. Here's an informative quote from this work: Propositional attitude reporting sentences concern cognitive relations people bear to propositions. A paradigm example is the sentence 'Jill believes that Jack broke his crown.' Arguably, 'believes,' 'hopes,' and 'knows' are propositional-attitude verbs and, when followed by a clause that includes a full sentence expressing a proposition (a that-clause) form propositional attitude reporting sentences [81, p. 1].

27 Due to lack of space, we leave aside our approach to formal NLP on the basis of proof theory alone. For a truly excellent book on proof-theoretic semantics, including for natural language, we recommend [82].

28 This fact is discussed in some detail in ref. [19], and is replete with relevant proofs. As an example, note that the truth or falsity of "Jones believes that ϕ" is not determined by the truth or falsity of ϕ, since humans routinely believe that falsehoods hold.

(see, for example, the schema (2) in Section 3.1). Step S1(ii) is a necessary, preparatory step for S2.

S2 Delete $\mathcal{S}$; if desired, move selected elements of $\mathcal{S}$ into $\mathcal{I}$, which requires casting these elements as inference schemata that employ metalogical expressions secured by prior step S1(ii). S2 reflects the fact that cognitive calculi have purely *inferential* semantics, and hence are aligned with the tradition of *proof-theoretic semantics* [55–57]. (In particular, cognitive calculi thus do not employ possible-worlds semantics for modal operators. In possible-worlds approaches, e.g., *knows* doesn't get defined as justified true belief; but in general, in the cognitive-calculus approach, knowledge in a cognitive calculus holds iff the agent in question believes the known proposition on the strength of a proof or argument that constitutes justification; see ref. [51]. As the alert reader will have noted, the μC calculus does not employ an operator for *knows*, but only for *believes*, corresponding to the operator **B**. The meaning of **B** in μC is all and only expressed by how it can be used in inference (which is determined by inference schema (2) in Section 3.1). We might, for instance, wish to include an inference schema that regiments the idea that an agent knows that which is provable from what she knows. Step S2 is the reason why we speak of a cognitive *calculus* (instead of, e.g., a cognitive *logic*, or cognitive *logical system*).

S3 Expand $\mathcal{I}$ as needed to include inference schemata that involve the operators from S1(i). For instance, where **K** is the modal operator for "knows" and **B** for "believes," we might wish to have this inference schema in a given C:

$$\frac{\mathbf{K}\phi}{\mathbf{B}\phi} I_{\mathrm{KB}}$$

A.1.1 Automated reasoning in cognitive calculi

ShadowProver is a fast and consistent reasoner for cognitive calculi that is under continuous development and available under an open-source license [11].

A.1.1.1 Regarding related work

Much could be said about work/systems that are related to cognitive calculi, but sustained treatment of this issue is out of scope in this brief appendix, which is merely meant to supplement the paper coming before it. We will say only a few things, and hope they are at least somewhat enlightening; here goes. The first published, implemented cognitive calculus, a multi-operator modal logic (minus, by definition, and as explained earlier in the present appendix, any model-theoretic semantics) based on multi-sorted first-order logic, can be found in refs. [35,58]; the second of these publications is a refinement of the first. Implementation at that point was based upon Athena, a recent introduction to which, along with a study of proof methods in computer science, is provided in the excellent [59]. Related work as cited in this earlier work remains relevant over a decade later, and in particular, so-called "BDI logics" (e.g., ref. [60]) are related, and we applaud their advent – but such logics cover very few propositional attitudes present in adult and neurobiologically normal cognition (e.g., no communication operators, and no emotional states), and are not based on purely inferential semantics. Automated reasoning in the tradition of higher-order logic (HOL) as descended from Frege, and most prominently from Church, which is masterfully chronicled in ref. [61], is obviously related to cognitive calculi; this is especially true since HOL is now very much on the scene in twenty-first-century AI (e.g., ref. [62]). In contrast, cognitive calculi, and the automation thereof, are based on commitments guided by the study of human cognition; and as we see it, that cognition for matters formal and extensional is for the most part circumscribed by natural deduction in third-order logic in the complete absence of formal semantics (e.g., consider the raw material in the practice of mathematics that gives rise to the argument and analysis in ref. [63]) and in matters literary circumscribed by modal operators mixed with third-order logic (e.g., ref. [64]). Traditionally, in terms of the Frege-to-Church-to… history that HOL has, HOL is extensional; in contrast, cognitive calculi by definition cannot fail to have operators that cover human cognition. The final thing we mention here is that cognitive calculi are not in any way deductive and bivalent or trivalent; they can be infused with uncertainty, and have multiple values (e.g., ref. [35]). The paper that precedes the present appendix shows such values in action, as the reader has seen.

A.1.1.2 Regarding metatheory for cognitive calculi

Some readers may wonder what metatheoretical properties cognitive calculi in general have, or what properties of this sort a particular $\mathscr{C}$ has. The metatheory of cognitive calculi is rich and not uncomplicated. We thus say only a few words here.

Standard metatheorems for standard extensional logics such as $\mathscr{L}_1$ include the familiar ones students of formal logic learn early on: e.g., soundness, completeness, compactness, and decidability. Consider specifically completeness, which holds for $\mathscr{L}_1$, but does not hold for $\mathscr{L}_2$ (second-order logic). This dual fact can only be expressed (let alone proved) if the logic involved has both a model theory (according to which a given formula ϕ can be true on all *interpretations*) and a proof theory (according to which a given ϕ can be a theorem, i.e., provable from the null set); both $\mathscr{L}_1$ and $\mathscr{L}_2$ qualify, of course. However, no cognitive calculus qualifies in this way: no cognitive calculus can be complete. The reason should be obvious: a cognitive calculus only permits semantic meaning of any kind to be defined inferentially. It thus is nonsensical to say that some $\mathscr{C}$ is complete/incomplete; the same hold for soundness and – once we take care to isolate a set of relevant formulae to which the metaproperty is to be applied – compactness. Of course, there are well-known analogs for the metaproperties of soundness, completeness, and compactness in the case of standard intensional logics, such as **K**, **T**, **S4**, and **S5**. (For such logics, standard model theory is supplanted with a different account of TRUE and FALSE, one that often uses possible worlds (see, e.g., ref. [65] for an introduction); and precise deduction is augmented with inference schemata (that in fact often *do* turn up in the $\mathcal{I}$ for a $\mathscr{C}$). But for exactly parallel reasons, it is nonsensical to say, on the basis of these definitions for truth and falsity, that a given $\mathscr{C}$ is sound/unsound or complete/incomplete. After all, and again, all cognitive calculi are exclusively proof- and argument-theoretic in nature, traceable in this regard back to the dawn of proof-theoretic semantics.

However, decidability is quite another matter. When it is (correctly, of course) said that because of Church's Theory $\mathscr{L}_1$ is undecidable, this can mean either that theoremhood for this logic is Turing-undecidable, or that necessary truth (validityhood?) is. It is perfectly meaningful to ask, w.r.t. a given $\mathscr{C}$, whether theoremhood is Turing-decidable. The answer, for μC and $\mathcal{DCEC}$ follows Church's Theorem and is hence a negative; the proof is trivial, since these cognitive calculi include $\mathscr{L}_1$. What about the answer for $\mathcal{S}$ and $\mathcal{IDCEC}$? This question is out of scope. Metatheory for inductive logics based only on extensional formal languages (such as those for $\mathscr{L}_1$) is a supremely technical affair that has become the province of logicians and mathematician, with most in AI, behavioral/cognitive robotics, and computational cognitive science having no familiarity with this field; for an elegant (but not simple) introduction to it, required for mathematical understanding of $\mathcal{IDCEC}$, see ref. [66].

Two final words: First: Cognitive calculi do in fact include a formal semantics built upon hypergraphs that can be fairly viewed as replacing the role of interpretations in model theory (in the extensional case) and frames, etc., in possible worlds (in the intensional case). The first appearance of such hypergraphs in connection with an informal inductive cognitive calculi is in ref. [67]. Use of hypergraphs in a manner that allows theorems asserting that in fact, e.g., μC is sound and complete is out of scope here. And now the second word, perhaps not insignificant: Bringsjord hereby claims, in keeping with his announcement at University of Turin on November 14 2016, 200 years to the day after Leibniz's death, that cognitive calculi, as characterized above, accompanied by the used-herein machinery for inventing, specifying, and deploying such calculi in a given domain, constitutes none other than the arrival of the long-sought dream of Leibniz et al.: viz., "the art of infallibility" (to use Leibniz's French phrase), or his "universal characteristic." Further elaboration taking account of Leibniz's writings must wait for a subsequent occasion.

A.2 Implementation

The scenario we presented was simulated in ShadowProver [11]. ShadowProver is a quantified multi-modal prover capable of handling reasoning in cognitive calculi [12]. The inputs to ShadowProver are shown in Figure A1. Time taken to simulate the adjudicator's reasoning about other agents' beliefs is shown in Table A1.

Table A1: Time taken for reasoning

Agent	Time taken (s)
hdrone	3.52
$ldrone_1$	3.82
radar	2.88
$ldrone_2$	2.87

(a)

```
{:description "The adjudicator reasoning about hdrone. "
 :assumptions
 {;;; It is common knowledge at the start what is needed to satisfy clause 2.
  :common (Common! t0
                   (iff clause2
                        (and
                         (exists p
                                 (and (Inside p Building)
                                      (Planning p)))
                         (forall p (if (Inside p Building) (not (Civilian p)))))))

  ;;; Report from the high altitude drone.
  :report (Believes! adj t0 (Believes! hdrone t0 (not (exists p (Inside p Building))))))}

 :goal       (Believes! adj t1 (Believes! hdrone t0 (not clause2)))}
```

(b)

```
{:description "The adjudicator reasoning about the faulty ldrone. "
 :assumptions
 {;; It is common knowledge that if an agent reports inconsistent information,
 ;; then as protective measure, we assume that the agent has no belief either way.
  :inconsistent (Common! t0
                         (if (Believes! ldrone t1 False)
                           (and (not (Believes! ldrone t1 clause2))
                                (not (Believes! ldrone t1 (not clause2))))))
  ;;; Report from the high altitude drone.
  :report       (Believes! a t1
                           (Believes! ldrone t1
                                      (exists p
                                              (and (Inside p Building)
                                                   (not (Inside p Building)))))))}

 :goal        (Believes! a t2
                         (and (not (Believes! ldrone t1 clause2))
                              (not (Believes! ldrone t1 (not clause2))))))}
```

(c)

```
{:description "The adjudicator reasoning about information from the radar."
 :assumptions
 {;;; It is common knowledge at the start what is needed to satisfy clause 2.
  :common (Common! t0
                   (iff clause2
                        (and
                         (exists p
                                 (and (Inside p Building)
                                      (Planning p)))
                         (forall p (if (Inside p Building) (not (Civilian p)))))))

  ;;; Report from the radar.
  :report (Believes! adj t2
                     (Believes! radar t2
                                (and
                                 (exists p
                                         (and (Inside p Building)
                                              (Planning p)))
                                 (forall p (if (Inside p Building) (not (Civilian p))))))))}

 :goal       (Believes! adj t3 (Believes! radar t2 clause2))}
```

(d)

```
{:description "The adjudicator reasoning about information from the fixed ldrone."
 :assumptions
 {;;; It is common knowledge at the start what is needed to satisfy clause 2.
  :common (Common! t0
                   (iff clause2
                        (and
                         (exists p
                                 (and (Inside p Building)
                                      (Planning p)))
                         (forall p (if (Inside p Building) (not (Civilian p)))))))

  ;;; Report from the the fixed ldrone.
  :report (Believes! adj t3
                     (Believes! ldrone t3
                                (exists p (and (Inside p Building) (Civilian p))))))}

 :goal       (Believes! adj t4 (Believes! ldrone t3 (not clause2)))}
```

Figure A1: Inputs to ShadowProver to model the scenario.
(a) The adjudicator reasoning about hdrone.
(b) The adjudicator reasoning about the faulty ldrone.
(c) The adjudicator reasoning about information from the radar.
(d) The adjudicator reasoning about the fixed ldrone.

A.3 Definitions for $\mathcal{DCEC}$ and $\mathcal{IDCEC}$

Below is the signature of the standard $\mathcal{DCEC}$. It contains the sorts, function signatures, and grammar of this cognitive calculus.

$\mathcal{DCEC}$ Signature

$$S ::= \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Fluent}$$

$$f ::= \begin{cases} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Formula} \\ \mathit{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Formula} \end{cases}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} q : \text{Formula} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : \phi(x) \mid \exists x : \phi(x) \\ \mathbf{P}(a, t, \phi) \mid \mathbf{K}(a, t, \phi) \mid \mathbf{S}(a, b, t, \phi) \mid \mathbf{S}(a, t, \phi) \\ \mathbf{C}(t, \phi) \mid \mathbf{B}(a, t, \phi) \mid \mathbf{D}(a, t, \phi) \mid \mathbf{I}(a, t, \phi) \\ \mathbf{O}(a, t, \phi, (\neg)\mathit{happens}(\mathit{action}(a^*, \alpha), t')) \end{cases}$$

Perceives, **K**nows, **S**ays, **C**ommon-knowledge, **B**elieves, **D**esires, **I**ntends, **O**ught-to.

Next is the standard set of inference schemata for $\mathcal{DCEC}$.

$\mathcal{DCEC}$ Inference Schemata

$$\frac{\mathbf{K}(a, t_1, \Gamma), \quad \Gamma \vdash \phi, \quad t_1 \leq t_2}{\mathbf{K}(a, t_2, \phi)}[I_{\mathrm{K}}]$$

$$\frac{\mathbf{B}(a, t_1, \Gamma), \quad \Gamma \vdash \phi, \quad t_1 \leq t_2}{\mathbf{B}(a, t_2, \phi)}[I_{\mathrm{B}}]$$

$$\frac{}{\mathbf{C}(t, \mathbf{P}(a, t, \phi) \rightarrow \mathbf{K}(a, t, \phi))}[I_1]$$

$$\frac{}{\mathbf{C}(t, \mathbf{K}(a, t, \phi) \rightarrow \mathbf{B}(a, t, \phi))}[I_2]$$

$$\frac{\mathbf{C}(t, \phi), \quad t \leq t_1, \ldots, t \leq t_n}{\mathbf{K}(a_1, t_1, \ldots \mathbf{K}(a_n, t_n, \phi)\ldots)}[I_3] \qquad \frac{\mathbf{K}(a, t, \phi)}{\phi}[I_4]$$

$$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{K}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a, t_2, \phi_1) \rightarrow \mathbf{K}(a, t_3, \phi_2)}[I_5]$$

$$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{B}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a, t_2, \phi_1) \rightarrow \mathbf{B}(a, t_3, \phi_2)}[I_6]$$

$$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{C}(t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2, \phi_1) \rightarrow \mathbf{C}(t_3, \phi_2)}[I_7]$$

$$\frac{}{\mathbf{C}(t, \forall x. \quad \phi \rightarrow \phi[x \mapsto t])}[I_8]$$

$$\frac{}{\mathbf{C}(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg\, \phi_2 \rightarrow \neg\phi_1)}[I_9]$$

$$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \phi])}[I_{10}]$$

$$\frac{\mathbf{B}(a, t, \phi)\ \mathbf{B}(a, t, \phi \rightarrow \psi)}{\mathbf{B}(a, t, \psi)}[I_{11a}] \qquad \frac{\mathbf{B}(a, t, \phi) \quad \mathbf{B}(a, t, \psi)}{\mathbf{B}(a, t, \phi \wedge \psi)}[I_{11b}]$$

$$\frac{\mathbf{S}(s, h, t, \phi)}{\mathbf{B}(h, t, \mathbf{B}(s, t, \phi))}[I_{12}]$$

$$\frac{\mathbf{I}(a, t, happens(action(a^*, \alpha), t'))}{\mathbf{P}(a, t, happens(action(a^*, \alpha), t'))}[I_{13}]$$

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

Finally, the following two boxes specify the signature and inference schemata for $\mathcal{IDCEC}$, respectively. These specifications enable reasoning about uncertain belief and knowledge. (For additional information beyond what is provided in these specifications regarding strength factors, in connection with the uncertainty system that underlies the system $\mathcal{S}$ used above, see ref. [35].)

Additional Syntax for $\mathcal{IDCEC}$

$\phi ::= \{\mathbf{B}^{\sigma}(a, t, \phi) \mid \mathbf{K}^{\sigma}(a, t, \phi)$
where $\sigma \in [-6, -5, \ldots, 5, 6]$

Additional Inference Schemata for $\mathcal{IDCEC}$

$$\frac{\mathbf{P}(a, t_1, \phi_1), \quad \Gamma \vdash t_1 < t_2}{\mathbf{B}^{A}(a, t_2, \phi)}[I_{\mathrm{P}}^{S}]$$

$$\frac{\mathbf{B}^{\sigma_1}(a, t_1, \phi_1), \ldots, \mathbf{B}^{\sigma_m}(a, t_m, \phi_m), \{\phi_1, \ldots, \phi_m\} \vdash \phi, \{\phi_1, \ldots, \phi_m\} \nvdash \zeta, \quad \Gamma \vdash t_i < t}{\mathbf{B}^{min(\sigma_1, \ldots, \sigma_m)}(a, t, \phi)}[I_{\mathrm{B}}^{S}]$$

where $\sigma \in [0, 1, \ldots, 5, 6]$

$$\frac{}{\mathrm{C}(t, {}^{-\sigma}\ \mathbf{B}(a, t, \phi) \leftrightarrow \mathbf{B}^{\sigma}(a, t, \neg\phi))}[I_{\neg}^{S}]$$