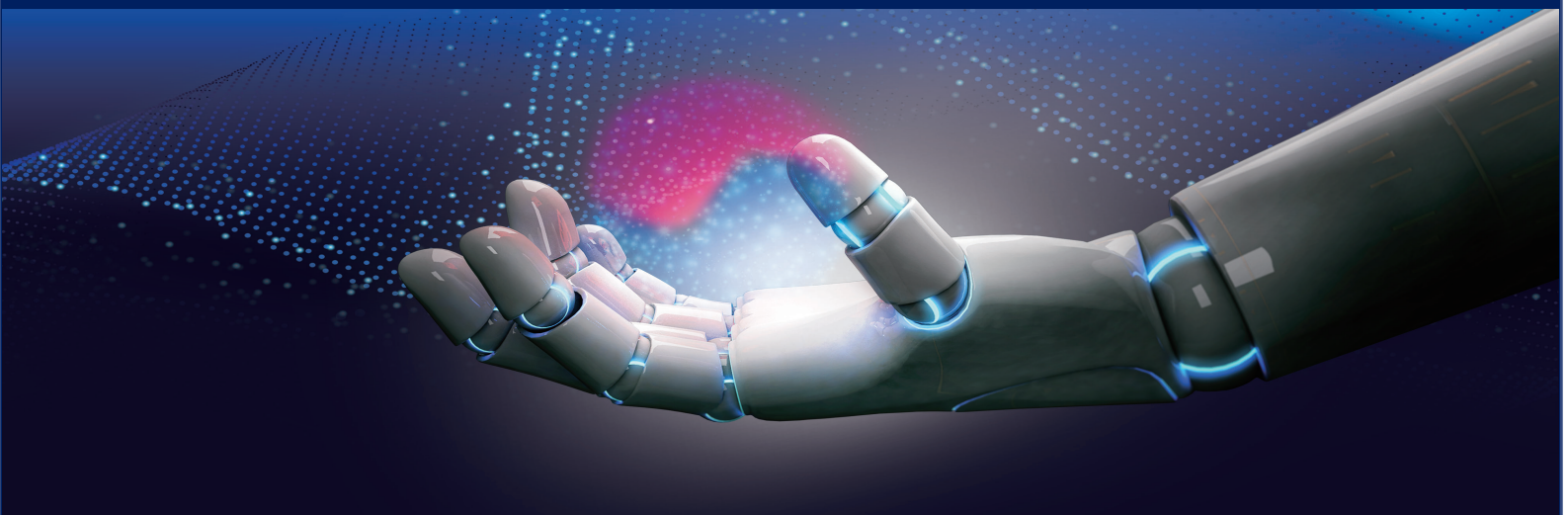
CLAWAR Association Series on
Robot Ethics and Standards



Value-Sharing between Humans and Robots

Editors

Sunyong Byun
Mohammad O. Tokhi
Maria Isabel A. Ferreira
Naveen S. Govindarajulu
Manuel F. Silva
Khaled M. Goher

VALUE-SHARING BETWEEN HUMANS AND ROBOTS

VALUE-SHARING BETWEEN HUMANS AND ROBOTS

ICRES 2022 Proceedings,
Seoul, South Korea, 18-19 July 2022

Editors

Sunyong Byun
Seoul National University of Education, & KSAIE, South Korea

Mohammad Osman Tokhi
London South Bank University, UK

Maria Isabel Aldinhas Ferreira
University of Lisbon, Portugal

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute, NY, USA

Manuel F. Silva
Porto Polytechnic, Portugal

Khaled M. Goher
University of Nottingham, UK

Published by

CLAWAR Association Ltd, UK (www.clawar.org)

Value-Sharing between Humans and Robots
Proceedings of the Seventh International Conference on Robot Ethics and Standards

ISBN 978-1-7396142-0-1

PREFACE

ICRES 2022 is the seventh edition of the International Conference series on Robot Ethics and Standards. The conference is organized by CLAWAR Association in collaboration with the Korean Society for Artificial Intelligence Ethics (KSAIE) in Seoul, South Korea during 18 – 19 July 2022.

ICRES 2022 brings new developments and new research findings in robot ethics and ethical issues of robotic and associated technologies. The topics covered include fundamentals and principles of robot ethics, social impact of robots, human factors, regulatory and safety issues.

The ICRES 2022 conference includes a total of 28 articles, and eight plenary lectures delivered by worldwide scholars. This number has been arrived at through rigorous peer review process of initial submissions, where each paper initially submitted has received on average three reviews. The conference additionally features special sessions on AI ethics education, trusting artificial intelligent systems, human robot interaction and Standardisation of Robot Systems and Evaluations. Furthermore, a discussion competition with elementary school pupils focusing on ethics of AI and technology is featured in the conference.

The editors would like to thank members of the International Scientific Committee and National Organising Committee for their efforts in reviewing the submitted articles, and the authors in addressing the comments and suggestions of the reviewers in their final submissions. It is believed that the ICRES 2022 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies.

S. Byun, M. O. Tokhi, M. I. A. Ferreira, N. S. Govindarajulu, M. F. Silva,
and K. M. Goher

CONFERENCE ORGANISERS

CLAWAR Association
www.clawar.org

Korean Society for Artificial Intelligence Ethics
https://ksaie.or.kr/

韓國倫理學會 **Korean Association of Ethics**

CONFERENCE SPONSORS AND SUPPORTERS

CONFERENCE COMMITTEES AND CHAIRS

Conference Chairs and Managers

Sunyong Byun (General Co-Chair)	– Seoul National University of Education, South Korea
Mohammad Osman Tokhi (General Co-Chair)	– London South Bank University, UK
Maria Isabel Aldinhas Ferreira (General Co-Chair)	– University of Lisbon, Portugal
Gurvinder S. Virk (Co-Chair IAC)	– CLAWAR Association, UK
Endre E. Kadar (Co-Chair IAC)	– University of Portsmouth, UK
Naveen S. Govindarajulu (Co-Chair ISC)	– Rensselaer Polytechnic Institute, USA
Shin Kim (Co-Chair ISC)	– Hankuk University of Foreign Studies, South Korea
Manuel F. Silva (Co-Chair ISC)	– ISEP & INESCTEC, Portugal
Jong-Wook Kim (Organising Committee Co-Chair)	– Dong-A University, South Korea
Tim Cheongho Lee (Organising Committee Co-Chair)	– Sangmyung University, South Korea

International Advisory Committee

Gurvinder S. Virk	– CLAWAR Association, UK
Endre E. Kadar	– University of Portsmouth, UK
Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Jen-Chieh Wang	– Industrial Technology Research Institute, Taiwan
Alan Winfield	– University of West England, UK

International Scientific Committee

Naveen S. Govindarajulu	– Rensselaer Polytechnic Institute, USA
Shin Kim	– Hankuk University of Foreign Studies, South Korea
Abdullah Almeshal	– College of Technological Studies, Kuwait
Sarah Fletcher	– Cranfield University, UK
Khaled M. Goher	– University of Nottingham, UK
Aman Kaur	– London South Bank University, UK
Philip Lance	– PA Consulting, UK
Manuel Silva	– ISEP-IPP and INESC TEC CRIIS, Portugal

National Organising Committee

Jong-Wook Kim	– Dong-A University, South Korea
Tim Cheongho Lee	– Sangmyung University, South Korea
Eunchan Bang	– Buyong Elementary School, South Korea
Jinwoo Jun	– KIRIA, South Korea
Joel Ryu	– KIRIA, South Korea

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.009

A Framework for Testimony-Infused Automated Adjudicative Dynamic Multi-Agent Reasoning in Ethically Charged Scenarios

Brandon Rozek and Michael Giancola and Selmer Bringsjord and Naveen Sundar Govindarajulu

Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI),
110 Eighth Street, Troy, NY 12180, USA
E-mail: {Rozek.Brandon, Mike.J.Giancola, Selmer.Bringsjord, Naveen.Sundar.G}@gmail.com

In "high stakes" multi-agent decision-making under uncertainty, testimonial evidence flows from "witness" agents to "adjudicator" agents, where the latter must rationally fix belief and knowledge, and act accordingly. The testimonies provided may be incomplete or even deceptive, and in many domains are offered in a context that includes other kinds of evidence, some of which may be incompatible with these testimonies. Therefore, before believing a testimony and on that basis moving forward, the adjudicator must systematically reason to suitable *strength* of belief, in a manner that takes account of said context, and globally judges the core issue at hand. To further complicate matters, since the relevant information perceived by the adjudicator changes over time, adjudication is a nonmonontonic/defeasible affair: adjudicators must dynamically strengthen, weaken, defeat, and reinstate belief and knowledge. Toward the engineering of artificial agents capable of handling these representation-and-reasoning demands arising from testimonial evidence in multi-agent decision-making, we explore herein extensions to one of our prior *cognitive calculi*: the *Inductive Cognitive Event Calculus* ($\mathcal{IDCEC}$). We ground these extensions in a recent, tragic drone-strike scenario that unfolded in Kabul, Afghanistan, in the hope that use by humans of our brand of logic-based AI in future such scenarios will save human lives.

Keywords: testimony-infused decision-making; multi-agent reasoning; argument adjudication

1. Introduction

Human agents often believe various propositions because they perceive part of their environment. Such an agent $\mathfrak{a}_h$ often believes for instance that there is a cup of coffee on the kitchen table because it sees the cup there courtesy of its own unaided sensors (eyes, e.g.); and in this case, all things being equal (e.g., the perceiver is not severely intoxicated), $\mathfrak{a}_h$ now as a result believes that there is a cup of coffee on the table. This basic picture stands at the heart of at least logic-based (= logicist) AI ([1], esp. Chap. 7 "Logical Agents;" and see as well dedicated treatments of logicist AI, e.g. [2]) and cognitive robotics of an overtly logicist sort [3], and is also a part of the very foundation of the empirical study of human cognition in information-processing terms (see e.g. [4]). However, agents, whether human or (present-day) artificial, are, we can all agree, not omnipresent; for this reason they often rely upon other agents to exceed the range of their own unaided sensors, by taking from these others *testimonies* (a term we use in its general sense, not in any narrow legal sense). A human agent located outside the kitchen may call to another agent inside it, "Is my coffee on the table in there?", and if hearing back an affirmative may rationally believe as a result that there is a cup of coffee on the kitchen table.

We shall in the present paper take a testimony to essentially have the basic shape of a triple $(\mathfrak{a}_w, \psi, \mathfrak{a}_{adj})$, where ψ is a declarative formula shared (via some form of communication, which make use of natural language expressing ψ) by a witness agent $\mathfrak{a}_w$ to an "adjudicator" agent $\mathfrak{a}_{adj}$. Adjudication is needed because whether it's rational for $\mathfrak{a}_{adj}$ to believe ψ at a given time frequently hinges on myriad factors, including competing, incompatible ones; and some of these competing factors can be testimonies themselves. The adjudicator in the case of the coffee example may receive in addition to an affirmative in response to a query, a negative

one — and now what should the adjudicator believe about the availability of desired caffeine? Realistically, we cannot assume that witness agents presenting testimonies are faultless. Such agents may have compromised perception or even ulterior motives. Therefore, when collecting and forming beliefs from testimonies, the adjudicator must reason over relevant, available information before fixing belief. And of course rational belief fixation engaged through time as the world changes and offers up new information, as has long been known in AI, is a temporally extended reasoning process that can't be exclusively deductive: this must pass into the realm of *inductive* logic, where inference is non-deductive, and uncertainty measures of some sort are used. In particular, new testimonies at time t may strengthen, weaken, or even defeat each other at that time, and may do the same to testimonies issued prior to t. Hence from the standpoint of logic-based (= logicist) AI and cognitive science, adept defeasible/nonmonotonic reasoning must be part of the adjudicator's cognitive arsenal.

To ground the rather admittedly abstract concepts and structures sketched in the previous paragraph, and the logico-mathematics behind them (which, in the form we prefer, we share soon: §3), we shall rely below upon an illuminating (and certainly sobering) case study of a recent drone strike in Kabul, Afghanistan. At the end of August 2021, as is widely known, US forces were evacuating Afghanistan. Three days before the incident we soon study, an ISIS K suicide attack killed 13 US troops and more than 60 afghan civilians [5]. The desire to prevent another attack was understandably high, as were tensions. In this emotionally and ethically charged context, authorization was given to employ kinetic counter-measures even under uncertainty, and as a matter of fact, such authorization was used — with tragic loss of innocent life. To use AI (or at least our brand of it) to prevent such tragedies in the future, automated reasoners must support, through time, ethical reasoning and counter-reasoning. We specifically need, as well, automated reasoners with the capability to detect and resolve inconsistencies arising from competing testimonies, arguments, and positions on profound moral matters. But this is only one desideratum (d_2, as will be shortly seen) among seven that constitute the requirements for the kind of capability our automated reasoning must have.

Enough introduction. The plan for the remainder is as follows: In the next section (§2), we enumerate, with brief exposition, the desiderata just alluded to. What follows next (§3) is a summary of the formal framework used in the work we report herein. Following that (§4), we return to the case study sketched in the previous paragraph, and establish (familiar, and broadly if not universally affirmed) conditions needed to permit a strike under the relevant type of contextual conditions by U.S. forces. We discuss related work and alternative approaches in §5, and compare them against our desiderata for automated adjudication of key information in the case study. In §6, we discuss the testimonies from outside intelligence sources and explain how, at least in our view, a rational defeasible system should handle them. Relevant and cognitively plausible inductive arguments are provided and treated in §7, along with a demonstration of our automated reasoner and how it can infer and adjudicate beliefs in a time-feasible manner that at least suggests the viability of AI-infused multi-agent decision-making in future situations analogous to the case study. We then close out the paper (§9) with with some final remarks, and recommended future actions.

2. Reasoning-System Requirements

Needless to say, any proposed set of requirements, or desiderata, for an automated reasoner (or for an ensemble of such systems) will directly reflect the general objectives and methodological orientation of the researchers and engineers involved in the pursuit at hand. We do not pretend that our overarching objectives are universally affirmed. For instance, for us,

any formal computational logic that fails to formalize and enable sophisticated *intensional* reasoning is unacceptable (relative to the applications we tend to emphasize), for reasons going back to the Frege, who while giving us the first fully rigorous and top-to-bottom presentation of first-order logic = $\mathscr{L}_1$, also presented us with the challenging observation that some rational agent $\mathfrak{a}$ can have beliefs about the morning star s_m, and radically different beliefs about the evening star s_e, and have no clue whatsoever that — expressed in the terms of extensional $\mathscr{L}_1$ — $s_m = s_e$.[a] This is specifically desideratum d_6 in the set of such, one we dub '$\mathcal{D}$', which is 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 $\mathfrak{a}$ 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* likelihood values or strength-factors (the latter case giving rise to multi-valued inductive logics corresponding to the cognitive calculus $\mathcal{IDCEC}$ used below);
- 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 (as opposed to purely abstract, meta-logical formalisms such as those of [7]), so that detailed step-by-step verification is possible, and over 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 socio-cognitive 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 (in, of course, particular instances), e.g. queries about provability at and even above the *Entscheidungsproblem* (e.g. at and above Σ^0_1 and Σ^1_1 in the Arithmetical and Analytical Hierarchies, resp.).

3. Formal Background

We make use herein of our previously erected formal framework for logicist AI and specifically automated reasoning, the chief component of which is a *cognitive calculus* $\mathcal{C}$ within an uncountably infinite family $\mathscr{C}$ of such. Full coverage of the family of cognitive calculi is out of scope.[b] We rely rather heavily in the present paper on the exemplar cognitive calculus

[a] A nice overview of intensional logic is given in [6], which in fact does discuss Frege and the example of Venus.

[b] Very briefly, the first building block of a cognitive calculus is simply a purely extensional and purely deductive *logical system* defined as in standard mathematical logic (e.g. in coverage of Linström's Theorems

presented momentarily, and used thereafter in the case study, to give readers (presumed to largely be cognoscenti) a good sense of what a cognitive calculus is; this cognitive calculus is the Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$), an inductive relative of the purely deductive $\mathcal{DCEC}$. For further information, the following resources among others are available: An efficient introduction to the family $\mathscr{C}$ is provided in [13]; use of a particular cognitive calculus for an ethically charged application handled by $\mathcal{DCEC}$ is provided in [14]; the first cognitive calculus that appeared in print is defined and used in [15]; for those who are more on the side of cognitive science than engineering-oriented AI, a sub-family of cognitive calculi are introduced and used in implemented form in [16], and sustained coverage in survey style of how formal logic can be used for cognitive modeling can be found in [17,18]. One final point before passing to specifics: A distinctive aspect of cognitive calculi is that they can be *heterogeneous*: their formal languages and inference schemata can allow diagrams and other pictorial elements, an approach given in the logic Vivid [19].

This works extends the $\mathcal{IDCEC}$ introduced in [20]. Briefly, a cognitive calculus consists of two parts: a signature and a set of inference schemata. The signature of the verison of $\mathcal{IDCEC}$ deployed herein is given in the box titled $\mathcal{IDCEC}$ **Signature**. It consists of three components: (1) a set of sorts (e.g. Agent, Action, etc. in order to capture states of the world and how it changes through time), (2) a set of types, and (3) a set of syntactic forms, including those of propositional and first-order logic as well as cognitive epistemic operators **B**elieves, **C**ommon-knowledge, **S**ays, and **P**erceives. The formulae are read in a fairly intuitive way. For example, $\mathbf{S}(a, b, t, \phi)$ is read as "Agent a *says* ϕ to agent b at time t."

$\mathcal{IDCEC}$ Signature

$$
\begin{aligned}
S &::= \mathsf{Agent} \mid \mathsf{ActionType} \mid \mathsf{Action} \mid \mathsf{Moment} \mid \mathsf{Fluent} \\
t &::= x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
\phi &::= \begin{cases} \psi : \mathsf{Formula} \mid \forall t : \phi \mid \exists t : \phi \\ \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \\ \mathbf{B}^{\sigma}(\mathfrak{a}, t, \phi) \mid \mathbf{C}(t, \phi) \mid \mathbf{S}(\mathfrak{a}, b, t, \phi) \mid \mathbf{P}(\mathfrak{a}, t, \phi) \end{cases}
\end{aligned}
$$

Cognitive uncertainty is captured through likelihood values shown in Table 1 and used

in [8]), which is then modified in three key ways: Move (i) is that all model-theoretic semantics are discarded but selectively recast and pulled back in purely proof-theoretic terms, by moving the central meta-logical expressions in such semantics (e.g. that some formula ϕ is satisfied by some interpretation $\mathcal{I}$, customarily written '$\mathcal{I} \models \phi$') into object-level syntax. And (ii), some set $\mathcal{I}$ of natural (hence when the schemata are deductive in nature, we specifically have natural-deduction schemata) inference schemata for a given calculus must be fixed, but are allowed to be inductive in nature (and hence draw from inductive logic, e.g. non-inferential *pure inductive logic* [9] or semi-formal inductive logic in the argument-centric tradition [10]) and make use of uncertainty measures (probabilities, likelihood values (which are used below), etc.), and done so in keeping with the third move (iii), which is the addition of modal operators that represent one or more cognitive 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 [11,12]), 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. Such verbs include: *knowing*, *believing*, *deciding*, *perceiving*, *communicating*, *desiring*, and *feeling* X where 'X' denotes some emotional state (e.g. possible $X = sad$, and so on. (Thus the reason we speak of a *cognitive* calculus should be plain.) "Off-the-shelf" modal logics are rejected because not only are they model-theoretic, possible-worlds-based, etc. instead of being purely inferential, but such semantics constrain what can be represented and automatically reasoned about, since e.g. perfectly meaningful English sentences are beyond the reach of any off-the-shelf logic (such as some version of quantified S5), e.g. "Selmer ought to bring it about that Brandon believes that at least three friends of Mike's know that Selmer just said 'None of Mike's friends save for one are angry'."

to ascribe a *quality* to the level of belief. This is represented formulaically as a superscript within the **B** operator and denoted by 'σ.' For example, $\mathbf{B}^4(a, t, \phi)$ can be read as "Agent a believes it is BEYOND REASONABLE DOUBT (cognitive likelihood value 4) that formula ϕ holds at time t."

As said above, we follow a proof-theoretic (or more accurately, an argument-theoretic[c]) as opposed to a model-theoretic (or, for the modal case, possible-worlds) approach. Proof-theoretic semantics for extensional logics, which avoid completely any Tarskian notion as a domain of discourse over which for example quantification ranges, was introduced by [21], extended in e.g. [22], and for a contemporary non-technical overview readers can consult [23]. The proof-theoretic semantics for cognitive calculi are beyond the scope of the present paper. An introduction to the core idea of extending proof-theoretic semantics for extensional logics to intensional ones, can be found in [24], which builds upon the natural-language-specific aspect of proof-theoretic semantics as set out in [25].

To summarize, the semantics of cognitive calculi are given exclusively via inference schemata, which dictate how new formulae can be derived and proofs can be constructed. The set of inference schemata for the version of $\mathcal{IDCEC}$ used herein is given in the box titled $\mathcal{IDCEC}$ **Inference Schemata**. Generally, an inference schema can be understood to say, "If the set of formulae above the line are true, then the formula below the line can be inferred." We will next describe how to interpret $[I_{WLP}]$, as it should then be clear how to interpret the others.

$[I_{WLP}]$ essentially implements the *Weakest Link Principle*, which dictates that an argument is only as strong as its weakest link. More formally, the schema says that if an agent a holds a set (of size m) of beliefs in formulae ϕ_1 to ϕ_m at likelihoods σ_1 to σ_m, and ϕ is provable from the set $\{\phi_1, \ldots, \phi_m\}$ *but* the set is not inconsistent (i.e. it cannot prove a contradiction), then at a later time t, a can infer a belief in ϕ, but only at the *minimum* likelihood of the beliefs used in the inference.

Given some background knowledge Γ, we desire our automated reasoner to make *ethical* decisions according to some ethical principle ρ. In $\mathcal{IDCEC}$, this will tell us whether the performance of an action α is ethically permissible, obligatory, or forbidden at time t. In section 4.1, we summarize the so-called Doctrine of Double Effect ($\mathcal{DDE}$), an ethical principle that draws from consequentialist, deontological, and divine-command ethical theories/traditions. As should be clear, nothing in the formalisms and technology that constitute the framework of our work is wed to $\mathcal{DDE}$: any credible and formalizable ethical principle can be used. (That said, $\mathcal{DDE}$ is in fact the basis for "just war" in the US and NATO.)

$\mathcal{IDCEC}$ Inference Schemata

$$\frac{\mathbf{B}^{\sigma}(\mathfrak{a}, t_1, \Gamma), \quad \Gamma \vdash \phi, \quad t_1 \leq t_2}{\mathbf{B}^{\sigma}(\mathfrak{a}, t_2, \phi)} \; [I_{\mathbf{B}}] \qquad \frac{\mathbf{P}(\mathfrak{a}, t_1, \phi), \quad \Gamma \vdash t_1 < t_2}{\mathbf{B}^{5}(\mathfrak{a}, t_2, \phi)} \; [I_{\mathbf{P}}]$$

$$\frac{\mathbf{B}^{\sigma_1}(\mathfrak{a}, t_1, \phi_1), \ldots, \mathbf{B}^{\sigma_m}(\mathfrak{a}, t_m, \phi_m), \{\phi_1, \ldots, \phi_m\} \vdash \phi, \{\phi_1, \ldots, \phi_m\} \nvdash \bot, \Gamma \vdash t_i < t}{\mathbf{B}^{min(\sigma_1, \ldots, \sigma_m)}(\mathfrak{a}, t, \phi)} \; [I_{WLP}]$$

[c]We note that argument-theoretic semantics are, in essence, the same as proof-theoretic semantics with one key distinction: it comes down to what differentiates a proof from a (formal) argument. Whereas a proof must be completely deductive, an argument can contain inductive/uncertain elements (e.g. uncertain beliefs, and/or inductive inference schemata), and hence an argument can conclude with a formula which contains a degree of uncertainty.

Table 1: The 13 Cognitive Likelihood Values

Numerical	Linguistic
6	*Certain*
5	*Evident*
4	*Beyond reasonable doubt*
3	*Very likely*
2	*Likely*
1	*More likely than not*
0	*Counterbalanced*
−1	*More unlikely than not*
−2	*Unlikely*
−3	*Very unlikely*
−4	*Beyond reasonable belief*
−5	*Evidently not*
−6	*Certainly not*

4. Case Study Part I

At the end of August 2021, the US was pulling its troops out of Afghanistan, primarily by way of the airport in Kabul. On August 26th 2021, two suicide bombers and gunmen attacked Kabul's airport, killing at least 13 US troops and 60 Afghans [5]. The desire to seek out and prevent future attacks on the airport was (naturally) acute. Six MQ-9 Reaper drones were deployed in order to search for potential ISIS K collaborators with the motive, means, and intentions[d] to bomb the airport. Once a suspect is located, a drone will track their activities to gather additional evidence, if possible. If a target is subsequently found, a sufficient level of evidence is sought, as are satisfaction of the ethical conditions. If conditions are met, then in this context, by U.S. policy, a hellfire missile may be fired to act as a counter-attack to prevent an attack on the airport.

4.1. *Conditions for Strike*

The DoD Law of War manual includes principles such as military necessity, proportionality, and distinction [28]. Military necessity justifies non-prohibited measures to end the war as quickly as possible. Proportionality limits the actions taken so that they are not unreasonable or excessive according to some utility function γ. Distinction ensures the protection of non-combatants and their objects. All of this is directly reflective of the longstanding Occidental tradition of "just war," going back to the Doctrine of Double Effect ($\mathcal{DDE}$), which originates with Augustine and was substantively refined by Aquinas. Drawing from consequentialism and deontology, $\mathcal{DDE}$ provides a set of conditions required to ethically permit an action α that may itself result in loss of life. (For an introduction to $\mathcal{DDE}$ in an AI context, see [14]; further analysis, formalization, and simulation with automated reasoning is provided in [29].) In the Kabul scenario, the ethical permissibility of firing a hellfire missile, requires a belief, at the level of *overwhelmingly likely* that ($\sigma_i \geq 4$):

$$C_1 \quad \mathbf{B}^{\sigma_1}(operator, t, \\ Capable(suspect, bomb(airport)))$$

[d]This is in general the correct approach, abstractly considered, in our opinion. We leave aside in this paper coverage of our reliance upon a formalization of the Wigmorean "MMOI" (motive, means, opportunity, and intent) pattern to persons of interest. For the background here, readers can consult [26] for a modern treatment, and [27] for original by Wigmore himself.

C_2 $\mathbf{B}^{\sigma_2}(operator, t, \mathbf{I}(suspect, t, bomb(airport)))$
$\mathcal{DDE}$ sanctions the strike.

When an operator encounters an imminent threat, a strike can be made under $\mathcal{DDE}$ as an act of *self-defense*. A large risk associated with this type of strike is the failure to detect and counter-attack within a fixed time period — which can often lead to catastrophic loss of life. In fact, an example of this is the attack on the airport three days prior, mentioned above. Therefore, if the former condition is not met, the following condition, in line with $\mathcal{DDE}$, acts as an alternative to permit a strike amid uncertainty.

C_1^* A positive belief of C_1 with $\sigma_1 > 0$.
C_2^* A positive belief of C_2 with $\sigma_2 > 0$.
C_3 $\mathbf{B}^{\sigma_3}(operator, t, \neg\exists t' > t :$
$Capable(operator, t', counterattack))$ with
$\sigma_3 \geq 4$.

Conditions for C_1 We define a suspect as capable of bombing the airport if they are near the airport and are in possession of an explosive item:

$$\frac{\begin{array}{l} B^{\sigma}(operator, t, \exists item \ : \\ Near(suspect, airport) \wedge \\ Explosive(item) \wedge \\ Has(suspect, item)) \end{array}}{B^{\sigma}(operator, t, Capable(suspect, t, bomb(airport)))}$$

Conditions for C_2 Assessing the intent of an individual is of course difficult. In the present case, our approach, as announced at the outset of the paper, is to rely upon testimonial evidence; our doing so is made plain §6.

Conditions for $\mathcal{DDE}$ Below are the informal conditions which express $\mathcal{DDE}$. A formalization of $\mathcal{DDE}$ in the cognitive calculus $\mathcal{DCEC}$ can be found in [14], and expressed in a formal but meta-logical manner in [30].

$\mathcal{DDE}_1$ The action α by itself is not ethically forbidden.
$\mathcal{DDE}_2$ The net utility of α in the situation is greater than some (non-trivial) positive amount γ.
$\mathcal{DDE}_3$ The agent performing α intends only the good effects from this action.
$\mathcal{DDE}_4$ The agent does not intend any of the bad effects from α.
$\mathcal{DDE}_5$ The bad effects are not used as a means to obtain the good effects.

Returning to the scenario, the drones were equipped with a camera which the operators used throughout the day to perceive and monitor multiple video feeds, all while incorporating outside intelligence into their process of belief fixation. The statements from this scenario are derived from [31] as well as [32]. Distinction between vehicles were communicated by their make, model, and color. The one we will see repeated multiple times is a white Toyota Corolla, represented as:

$$\begin{array}{l} \forall x : WTC(x) \iff \\ \quad White(x) \wedge Toyota(x) \wedge Corolla(x) \end{array}$$

We begin the scenario at time t_0 with an operator tracking a suspect driving a white Toyota

Table 2: Beliefs from Perception (at likelihood 5 from $I^{s}_{\mathbf{P}}$)

Label	Operator's Belief
B_1	$WTC(car)$
B_2	$driver(car) = suspect$
B_3	$Near(car, house)$
B_4	$Near(driver(car), house)$
B_5	$Near(house, safehouse)$
B_6	$Has(suspect, item)$
B_7	$Briefcase(item)$
B_8	$Near(suspect, airport)$

Corolla.

$$\mathbf{P}(operator, t_0, WTC(car) \wedge driver(car) = suspect)$$

At some time t_1, the suspect makes a stop at a house.

$$\begin{aligned}&\mathbf{P}(operator, t_1,\\ &Near(car, house) \wedge Near(driver(car), house))\end{aligned}$$

The house also appears to be near the suspected safehouse.

$$\mathbf{P}(operator, t_1, Near(house, safehouse))$$

Some time after the stop, while by the house, the suspect is seen carrying a laptop bag.

$$\begin{aligned}&\mathbf{P}(operator, t_2, Has(suspect, item) \wedge\\ &Briefcase(item) \wedge Near(item, house))\end{aligned}$$

Toward the end of the day, the drone sees the suspect park at a location that is three kilometers from the airport.

$$\mathbf{P}(operator, t_3, Near(suspect, airport))$$

These perceptions are then converted to beliefs as shown in Table 2.

5. Related Work

Much could be said about related work; we shall keep things brief in the present section, and touch upon a few distinctive apsects of our approach, with an eye to our desiderata $\mathcal{D}$.

To start with the general, as most readers well know, nonmontonic/defeasible reasoning and logics that support such reasoning can be traced back a number of decades. McCarthy [33] invented nonmonotonic logics based on circumscription, which is second-order and model-theoretic, the latter aspect at odds with our thoroughgoingly inference-by-inference-schemata approach (at odds with d_4 and d_5). In a firmly argument-centric orientation that specifically commits to the internal structure of arguments as crucial (satisfying d_4), Pollock later invented and implemented the defeasible-logic reasoner OSCAR [34–36] that

inspired us.[e] Early seminal work in nonmonotonic logic was carried out (temporally speaking) alongside McCarthy; for instance, specifically, we have the default logics of Reiter [37], in which epistemic possibilities hold in default of information to the contrary. However, none of the excellent work by these three pioneers was intensional in nature; no intensional operators, let alone intensional inference schemata, are to be found (failure of d_6). We leave aside in the interest of economy further assessment of the McCarthy-Reiter-Pollock trio w.r.t. to other desiderata. We do mention that more recently, Licato [38] modeled a complex case of deceptive reasoning and planning from the award-winning television series *Breaking Bad* using default logic. Their work did in fact use a cognitive calculus in the family $\mathscr{C}$ (the Cognitive Event Calculus, $\mathcal{CEC}$, which is devoid of deontic operators) to model the beliefs and intentions of various agents, but didn't have a formalism for assigning strengths to beliefs, and was in the realm of deductive logics, not inductive ones; therefore, while commendable on many fronts, their system does not satisfy d_3.

What about more work in defeasible argumentation systems, considering the promised eye to the desiderata we have laid down? We mention two pieces of impressive prior work, neither of which significantly overlaps our new approach, as we explain:

(1) [39] presents a general framework for structured argumentation, and the framework is certainly computational in nature. The framework, ASPIC$^+$, is in fact Pollockian in nature, at least in part. More specifically this framework 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" [p. 31, [39]]. This second principle is directly at odds with desideratum d_5. In our our intensional inductive calculi, including specifically $\mathcal{IDCEC}$, all non-deductive inference schemata are formally 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 even than even *modus ponens*, where of course specifically we have $\frac{\phi \rightarrow \psi, \phi}{\psi}$.[f]

(2) [41] 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 (at odds thus with d_7). 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 too is not true of our approach; in fact, adjudication for us is most needed when there is a complete absence of a state-of-affairs in which no arguments attack each other.

Work in testimonial evidence has a close tie to epistemology, and under that topic a tie

[e] Pollock has unfortunately passed away. Bringsjord's RAIR Lab currently maintains and offers OSCAR, courtesy of resurrection of the (Common Lisp) code by Kevin O'Neill. Today, there are strikingly few extant nattural-deduction automated deductive reasoners, and OSCAR is one at the level of at least $\mathscr{L}_1$.

[f] For a discussion of this sort of explicit rigidity in the case of analogical inference, see [40].

specifically to notions of trust, deontological justification of belief, judgement of character, and evidentialism. In our view, in this regard, helpful prior work is the account of occurrent trust in [42]; the account states that an agent $\mathfrak{a}_b$ trusts an agent $\mathfrak{a}_c$ to do an action α with respect to some goal ψ if and only if:

(1) Agent $\mathfrak{a}_b$ has the goal ψ.
(2) Agent $\mathfrak{a}_b$ believes that:

- Agent $\mathfrak{a}_c$ is capable of doing action α.
- By agent $\mathfrak{a}_c$ doing α, it will ensure ψ.
- Agent $\mathfrak{a}_c$ intends to perform action α.

As to deontological justification of belief, the idea here is that an agent $\mathfrak{a}_h$ is justified in believing ϕ if and only if $\mathfrak{a}_h$ is not obligated to refrain from believing ϕ. Since $\mathcal{IDCEC}^*$ is a deontic cognitive calculus, it can be easily expressed by us as:

$$\frac{S(\mathfrak{a}_s, \mathfrak{a}_h, t, \phi) \quad \neg O(\mathfrak{a}_h, t, \neg B^n(\mathfrak{a}_h, t, \phi), \chi)}{B^1(\mathfrak{a}_h, t, \phi)} \; I_B^D$$

As to evidentialism, [43] is certainly relevant, and when cast into our formal machinery states that an agent $\mathfrak{a}_h$ is justified in believing ϕ if and only if the belief of ϕ fits the evidence available; more precisely:

$$\frac{S(\mathfrak{a}_s, \mathfrak{a}_h, t, \phi) \quad \{\phi_1, \ldots, \phi_m\} \vdash \phi \quad \{\phi_1, \ldots, \phi_m\} \nvdash \bot}{B^1(\mathfrak{a}_h, t, \phi)} \; I_B^E$$

Finally, there is informal but remarkable work on testimonial evidence and character that has inspired us, and which will continue to do so; this is the work of Walton [44,45]. We return to this at the very end of the present paper.

6. Testimony-based Inferencing

Let us denote the agent or operator watching the video streams while gathering and adjudicating beliefs (following notation introduced earlier) as the adjudicator $\mathfrak{a}_{adj}$. Outside intelligence may come from satellite imagery, intercepted radio communications, sources at the site, etc. It is up to the adjudicator to determine whether it believes a *testimony* coming from an outside source.

6.1. *Inference Schemata for Testimonial Evidence*

In $\mathcal{IDCEC}$, testimonies are communicated using the **S**ays operator. (In this cognitive calculus, no NLU or NLG based on logically controlled natural language is used, so this operator is not associated with subsidiary computation for NLP in the present paper.) For example, a testimony from intelligence analyst $\mathfrak{a}_i$ for a claim ω can be represented as $\mathbf{S}(\mathfrak{a}_i, \mathfrak{a}_{adj}, t, \omega)$. That that a formula of this type conforms to the basic triadic structure of testimonies given earlier in the present paper.

We employ a confessedly naïve inference schema for handling testimonies, according to which the operator simply believes everything the intelligence analyst says, at the level of *highly likely* (we end the paper by pointing toward more sophisticated schemata):

$$\frac{\mathbf{S}(\mathfrak{a}_i, \mathit{operator}, t, \omega)}{\mathbf{B}^3(\mathit{operator}, t, \omega)} \; I_{naive}$$

7. Case Study Part II

Given the necessary formal background, we return to the scenario.

7.1. *Intelligence Reports*

Table 3: Beliefs from Testimonies (at likelihood 3 from I_{naive})

Label	Operator's Belief
B_9	$Collab(iperson_1, ISIS\ K)$
B_{10}	$\mathbf{I}(iperson_1, t', bomb(airport))$
B_{11}	$iperson_1 = driver(icar)$
B_{12}	$WTC(icar)$
B_{13}	$Near(iperson_2, safehouse)$
B_{14}	$Collab(iperson_2, ISIS\ K)$
B_{15}	$Has(iperson_2, iitem)$
B_{16}	$Explosive(iitem)$
B_{17}	$Near(driver(icar), safehouse)$
B_{18}	$Near(icar, safehouse)$

There are three pieces of intelligence that the operators received throughout the day. To add granularity, we assigned different intelligence analysts to each of them.

The first piece of intelligence, which comes from prior attacks is that a driver of a white Toyota Corolla is a collaborator of ISIS K and intends to bomb the airport.

$$\begin{aligned} &Collab(iperson_1, ISIS\ K) && \wedge \\ &\mathbf{I}(iperson_1, t', bomb(airport)) && \wedge \\ &driver(icar) = iperson_1 && \wedge \\ &WTC(icar) && (\psi_1) \end{aligned}$$

The next piece of outside information comes from intercepted communications and states that a meeting to hand off explosions to an ISIS K collaborator is happening at the safehouse.

$$\begin{aligned} &Near(iperson_2, safehouse) && \wedge \\ &Collab(iperson_2, ISIS\ K) && \wedge \\ &Has(iperson_2, iitem) && \wedge \\ &Explosive(iitem) && (\psi_2) \end{aligned}$$

The last piece of outside intelligence comes from a satellite analyst which states that a white Toyota Corolla left the safehouse.

$$\begin{aligned} &Near(driver(icar), safehouse) && \wedge \\ &Near(icar, safehouse) && (\psi_3) \end{aligned}$$

These pieces of outside intelligence are then converted to beliefs by the operator into Table 3.

Amidst uncertainty, the military need some way to associate objects which are potentially the same. In the next subsections, we will introduce the notion of cognitive transitive nearness as well as the Identity of Indiscernables under Uncertainty.

7.2. *Nearness Properties*

It is clear that the Predicate *Near* has the symmetric property:

$$\forall p_1, p_2 : Near(p_1, p_2) \iff Near(p_2, p_1)$$

The same cannot be said of transitivity. Otherwise, one can chain enough objects that are near each other to not satisfy Nearness. However, it is cognitively plausible that the transitive property of nearness holds, albeit with slightly less confidence with each chain.

$$\begin{aligned}\forall p_1, p_2, p_3 : \mathbf{B}^{\sigma_1}(\mathfrak{a}, t, Near(p_1, p_2)) \wedge \\ \mathbf{B}^{\sigma_2}(\mathfrak{a}, t, Near(p_2, p_3)) \implies \\ \mathbf{B}^{max(0,min(\sigma_1,\sigma_2)-1)}(\mathfrak{a}, t, Near(p_1, p_3))\end{aligned}$$

7.3. *Identity of Indiscernables under Uncertainty*

The standard Identity of Indiscernables state that two objects share all the same properties iff they are the same object as further described in [46] and [8].

$$\forall F : (Fx \iff Fy) \iff x = y$$

Now a version that allows for uncertainty is that it is believable that two objects are the same if you believe they're near each other and share two properties. The following statement is in third-order logic.

$$\begin{aligned}\exists F, G\ \forall x, y : F \neq G \wedge F \neq Near \wedge G \neq Near \wedge \\ \big(\mathbf{B}^{\sigma_1}(\mathfrak{a}, t, Near(x, y)) \wedge \\ \mathbf{B}^{\sigma_2}(\mathfrak{a}, t, Fx) \wedge \mathbf{B}^{\sigma_3}(\mathfrak{a}, t, Fy) \wedge \\ \mathbf{B}^{\sigma_4}(\mathfrak{a}, t, Gx) \wedge \mathbf{B}^{\sigma_5}(\mathfrak{a}, t, Gy)\big) \implies \\ \mathbf{B}^{max(0,min(\sigma_i)-1)}(\mathfrak{a}, t, x = y)\end{aligned}$$

7.4. *Associating Objects*

We first want to associate that the two people specified in the outside intelligence refers to the same person. We can do this by first using the symmetric and cognitive transitive nearness properties:

$$\begin{aligned}\big(\mathbf{B}^3(operator, t, Near(iperson_1, safehouse)) \wedge \\ \mathbf{B}^3(operator, t, Near(safehouse, iperson_2))\big) \implies \\ \mathbf{B}^2(operator, t, Near(iperson_1, iperson_2))\end{aligned}$$

Then we can use the Identity of Indiscernables under Uncertainty:

$$\begin{aligned}\big(\mathbf{B}^2(operator, t, Near(iperson_1, iperson_2)) \wedge \\ \mathbf{B}^3(operator, t, Collab(iperson_1, ISIS\ K) \wedge \\ \mathbf{B}^3(operator, t, Collab(iperson_2, ISIS\ K) \wedge \\ \mathbf{B}^3(operator, t, \mathbf{I}(iperson_1, bomb(airport)) \wedge \\ \mathbf{B}^3(operator, t, \mathbf{I}(iperson_2, bomb(airport)))\big) \implies \\ \mathbf{B}^1(operator, t, iperson_1 = iperson_2) \qquad (B_{19})\end{aligned}$$

Next we want to show a belief that the suspect driving the white Toyota Corolla is driving the same white Toyota Corolla from the pieces of outside intelligence. We first use

the nearness property to establish a belief that the suspect's car is near the safehouse.

$$\begin{aligned}&\big(\mathbf{B}^5(operator, t_2, Near(car, house)) \wedge \\ &\ \mathbf{B}^5(operator, t_2, Near(house, safehouse))\big) \implies \\ &\ \mathbf{B}^4(operator, t_2, Near(car, safehouse))\end{aligned} \tag{B_{20}}$$

We then use the cognitive transitive property of nearness again and symmetric property to establish that the suspect's car is near the car from the intelligence reports.

$$\begin{aligned}&\big(\mathbf{B}^4(operator, t_2, Near(car, safehouse)) \wedge \\ &\ \mathbf{B}^3(operator, t_2, Near(safehouse, icar))\big) \implies \\ &\ \mathbf{B}^2(operator, t_2, Near(car, icar))\end{aligned} \tag{B_{21}}$$

With the two cars near each other, we can then use the fact that they're both white Toyota Corolla's to infer a belief that they're the same object.

$$\begin{aligned}&\big(\mathbf{B}^2(operator, t, Near(car, icar)) \wedge \\ &\ \mathbf{B}^5(operator, t_2, White(car)) \wedge \\ &\ \mathbf{B}^3(operator, t_2, White(icar)) \wedge \\ &\ \mathbf{B}^5(operator, t_2, Corolla(car)) \wedge \\ &\ \mathbf{B}^3(operator, t_2, Corolla(icar))\big) \implies \\ &\ \mathbf{B}^1(operator, t_2, car = icar)\end{aligned} \tag{B_{22}}$$

We can then apply the function *driver* to induce:

$$\begin{aligned}&\mathbf{B}^1(operator, t_2, driver(car) = driver(icar)) \implies \\ &\mathbf{B}^1(operator, t_2, suspect = iperson_1)\end{aligned} \tag{B_{23}}$$

It is at this point that the military can infer (via substitution) albeit with a low degree of confidence that the suspect is capable and intends to bomb the airport.

$$\begin{aligned}&\big(\mathbf{B}^5(operator, t_3, Near(suspect, airport)) \wedge \\ &\ \mathbf{B}^1(operator, t_3, Has(suspect, iitem)) \wedge \\ &\ \mathbf{B}^1(operator, t_3, Explosive(iitem))\big) \implies \\ &\ \mathbf{B}^1(operator, t_3, Capable(suspect, t_3, bomb(airport))\end{aligned} \tag{B_{24}}$$

7.5. *Lack of Time Argument*

Since the level of belief in B_{24} is low, a strike will not be permitted unless a belief *beyond reasonable doubt* is held that there is no additional time to counterattack.

ϕ_1 From prior attacks, the operator has a *very likely* belief that the explosive item is either a suicide vest (SVest) or a rocket.

- $B^3(operator, t_3, SVest(iitem) \vee Rocket(iitem))$

ϕ_2 SVests are explosives that can fit into a briefcase.

- $\forall x : SVest(x) \implies Explosive(x) \wedge Briefcase(x)$

ϕ_3 Rockets would not be able to fit into a briefcase.

- $\forall x : Rocket(x) \implies \neg Briefcase(x)$

B_{25} We know from B_7, B_{16} that the suspect has an item that is explosive and an item that fits in a briefcase.

- $\mathbf{B}^3(operator, t_3,$
  $Explosive(iitem) \wedge Briefcase(item))$

B_{26} Through cognitive transitive nearness:

- $\mathbf{B}^2(operator, t_3, Near(item, iitem))$

B_{27} Using the Identity of Indiscernables under Uncertainty we can derive a belief that $iitem = item$ using the relations $Has(suspect, x), ObtainedNearSafehouse(x)$ with a belief level of 2.

- $\mathbf{B}^2(operator, t_3, iitem = item)$

B_{28} The explosive *item* fits inside a briefcase, therefore through disjunction syllogism, it must be an SVest.

- $\mathbf{B}^1(operator, t_3, Svest(item))$

ϕ_4 It is known that SVests are hard to counterattack in a populated area unless the suspect is enclosed.

- $\exists\ suspect, item:$
  $\big(Populated(suspect) \wedge SVest(item) \wedge$
  $Has(suspect, item) \wedge \neg Enclosed(suspect)\big) \implies$
  $\neg Capable(operator, t', counterattack)$

B_{29} The operator perceived the suspect park in a populated and enclosed location. (From perception and $I^{\mathbf{s}}_{\mathbf{P}}$)

- $\mathbf{B}^5(operator, t,$
  $Populated(suspect) \wedge Enclosed(suspect))$

B_{30} The operator held a belief through perception that the suspect would soon be not enclosed.

- $\mathbf{B}^5(operator, t', \neg Enclosed(suspect))$

B_{31} Once the suspect is not enclosed, opportunity would've been lost to counterattack. Therefore, there is no future time available to counterattack.

- $\mathbf{B}^1(operator, t_3, \neg\exists t' > t_3:$
  $Capable(operator, t', counterattack))$

7.6. *Simulations Achieved via Automated Reasoning*

ShadowProver [47] is an automated reasoner for the (purely deductive) $\mathcal{DCEC}$. In this work, we utilize a nascent automated reasoner — ShadowAdjudicator [48] — which contains a novel algorithm for reasoning about *inductive* cognitive calculi such as $\mathcal{IDCEC}$. It builds directly off of ShadowProver to enable reasoning about formulae with likelihood values.

ShadowAdjudicator is able to automatically find all of the arguments presented herein. The two main arguments, presented in §7.4 and §7.5, are each split into three sub-arguments.

The three sub-arguments of the first argument are: (1) the two people specified in the intelligence reports are the same person, therefore (2) the (perceived) suspect is the same person as the person identified in the intelligence reports, therefore (3) that the suspect is capable of bombing the airport. Those three sub-arguments are found by ShadowAdjudicator in 4.2s, 1.1s, and 0.3s respectively.

The three sub-arguments of §7.5 are: (1) the (perceived) item is the same item as the one mentioned in the intelligence reports, therefore (2) the item is a SVest, therefore (3) there is not enough time to wait to perform a counterattack. These three sub-arguments are found by ShadowAdjudicator in 0.27s, 0.34s, and 0.30s respectively.

8. Concerns; Replies

Before concluding the paper, we devote the present, short section to replies to a pair of concerns we anticipate arising in the minds of some thoughtful readers.

8.1. *What About Prior Logic-based Modeling?*

A reader might reasonably say: "You do have a section above on related work, which is appreciated, but there you focus on a general class of AI systems (into which yours falls), rather than on any specific modeling efforts in logic-based AI. Isn't there some modeling (and simulation) work in the past that is relevant to what you do here? After all, many folks have used formal logic to model various phenomena. For instance, Hayes [49] rather famously long ago modeled the behavior of fluids (in naïve-physics fashion), and much more recently Shanahan [50] modeled the cracking of an egg, which he rightly regards to be a 'benchmark' problem. How does your framework relate to this kind of impressive work?"

We have much respect for the work cited by this imagined reader. In fact we believe that Hayes inaugurated this line of work in nothing short of seminal fashion. However, this work is fundamentally different than our framework, and the work that has bee carried out to erect is. The differences are numerous; we have space here to mention but two; they are as follows. One, our research, our logico-mathematical formalisms (e.g. our cognitive calculi, including the one employed herein), and our automated-reasoning technology; all of this is *intrinsically intensional*. We are not interested the logicist modeling and simulation of objects and processes that are non-cognitive. Now of course it might be said that the cracking of an egg can be quite cognitive, but the fact is that in the aforementioned [50] there is no use of, and more importantly no need for, logics that have a singularly robust lineup of intensional/modal operators.[g] Our second point is this: Our framework is in no way given as first and foremost a contribution to modeling. No, our purpose is to engineer a framework in which AI can make decisions that save lifes, by adjudicating arguments regarding life-and-death questions.

8.2. *Is Your Framework Extensible?*

A second concern can be expressed thus: "Surely you agree that if your framework were a one-off affair, it would have precious little value. So, you must see that a skeptic would demand from you some assurance that your framework will carry over not only to other case studies, but to an entire class of multi-agent decision-making challenges of great consequence."

In reply, we cheerfully admit that our framework is intended to be extensible, and applicable well beyond the particular scenarios we have selected. (We also acknowledge that the work on cracking an egg cited above is very much intended to be applicable to a wide swathe of *physical* phenomena.) But does not one glance at the nature of for instance $\mathcal{IDCEC}$ reveal instantly the broad scope of our framework? After all, the inference schemata at the heart of all of our cognitive calculi are formal, abstract, and entirely domain-general. Indeed, it's very hard for us to fathom how, given the nature of these schemata, anyone rational could fail to see the extensibility and broad applicability of our framework. Consider perhaps the simplest inference schema known: *modus ponens*, which made a brief appearance above:

$$\frac{\phi \rightarrow \psi, \phi}{\psi}$$

[g] It is probably worth pointing out that in our cognitive calculi said to be "event calculi," the event calculus (which was invented by Shanahan himself; see e.g. [51]) is not used in its original form, but is rather cast in cognitive form. This has been the case going back to the very first cognitive calculus invented and used in modeling-and-simulation work: [15].

How can it fail to be that this inference schema is applicable to any number of spheres, and that it can be part of larger and larger collections of inference schemata? Likewise, the intensional inference schemata in $\mathcal{IDCEC}$ have the same wide scope, and can be parts of arbitrarily extended collections of inference schemata.

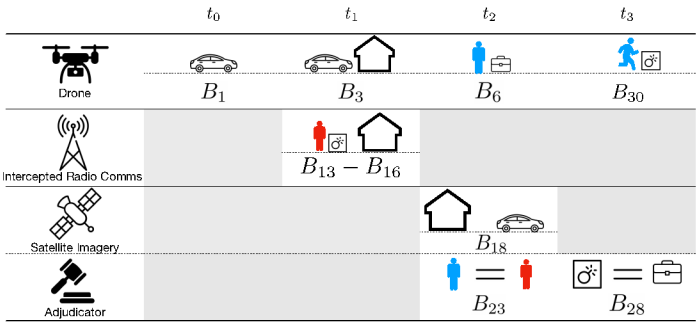


Fig. 1. Scenario

9. Conclusion & Next Steps

The scenario ends with a low degree of confidence that there was no additional opportunity to counterattack. This means that the two possible conditions to permit a strike were both unsatisfied: viz.,

- a belief *beyond reasonable doubt* that the suspect was capable and intending to bomb the airport; and
- a positive belief that the suspect was capable of and intending to bomb the airport, and a belief *beyond reasonable doubt* that there is no additional time to counterattack.

These conditions address the principle of distinction in the DoD law of war manual while also addressing the fog of war or uncertainty frequently present in warzones.

A cognitively plausible argument presenting a positive belief that the suspect was capable of and intent on bombing the airport, given the foregoing, is seen to be justified. This argument made use of inductive inference, which in turn more generally made use of higher-order logic in order to associate objects via both our "uncertainty-ized" Identity of Indiscernables, and transitivity of nearness.

Since an argument couldn't be made at the *beyond reasonable doubt* level that there was no additional time to counterattack, from a purely rational viewpoint we respectfully assert that more time should've been taken to track the suspect into the future, and gather newly arriving information. We believe that as logicist AI of the type we have featured herein progresses, the automated adjudication of arguments can extend beyond the technology that we exhibit herein, and take into account the blast radius of the missile, time between permitting a strike and it landing, velocity of the target, and so on, in order to maximize observation time and the accuracy of automated human/AI reasoning about agents of interest, and potential dangers arising therefrom.

To sum up, the main purpose of our work has been to provide to artificial agents an automated representation-and-reasoning capability sufficiently expressive and powerful to deal with ethically charged cases like the Kabul tragedy (which makes its expressive and automated-reasoning reach unprecedented, to our knowledge) — and this purpose has been achieved. Our chief overarching technical result is invention of the inductive logic we have

displayed, its use for a robust and relevant case study, and its automaticity and automated runs. Our AI automatically finds in this case study the relevant proofs and — since this is inductive logic, not deductive — arguments, , and we are the first group to achieve any such engineering, as far as we know.

What about next steps along the line of investigation described above? As alert and discriminating readers will doubtless have detected, the brute fact is that our inference schema for testimonial evidence is naïve. The next phase of our efforts will be to complete the specification of a robust and credible inference schema in a variant of $\mathcal{IDCEC}$, and achieve implementation via our automated reasoners.

Fortunately, there is some seminal prior work on such evidence under the umbrella of *in*formal logic, carried out by Walton [44,45]. Taking, we admit, considerable liberties in sharpening this work so as to make use of it in our formalisms and in the automated reasoners that enable such agents to compute, we can lay out at least a provisional formal version of one such inference schema that can be expressed in the inductive cognitive calculus $\mathcal{IDCEC}$ (see Figure 2), and with a presentation of it immediately below we end.

$$\frac{\begin{array}{c}\Vdash \mathbf{B}^{\sigma_1}(\mathfrak{a}_{adj}, \mathit{EpistemicPos}(\mathfrak{a}_w, \phi)) \\ \Vdash \mathbf{B}^{\sigma_2}(\mathfrak{a}_{adj}, \geq (\mathit{character}(\mathfrak{a}_w), k)) \\ \mathbf{K}(\mathfrak{a}_{adj}, \mathbf{S}(\mathfrak{a}_w, \phi, \mathfrak{a}_{adj}))\end{array}}{\mathbf{B}^{f(\sigma_1,\sigma_2,k)}(\mathfrak{a}_{adj}, \phi)}$$

Fig. 2. A Provisional Inference Schema for Testimonial Evidence. *The traditional single-turnstyle* $\vdash$ *for straight deductive provability is here replaced with a variant that indicates that what follows it is the conclusion of inference that may be either deductive or inductive, expressed, respectively, by a proof or formally valid argument.*

Acknowledgments

The work reported herein exploits key automated-reasoning theory & technology developed via multi-year support from both AFOSR (Award # FA9550-17-1-0191) and ONR of investigators Bringsjord & Govindarajulu, and of doctoral researchers (including Giancola) under their supervision in the Rensselaer AI & Reasoning (RAIR) Laboratory. B&G are especially grateful for ONR support (Award # (N000141912558) intended to surmount Arrow's Impossibility Theorem (AIT) via the brand of artificial adjudication described above — and this pair hereby asserts that such surmounting (courtesy in no small part of their requiring that "voting" must be supported by an argument) has come to pass. They are also grateful for ONR support under Award # N00014-17-1-2115 for the development of logic-based machine learning, which in part makes the formal schemes and reasoning therein shown herein possible.

References

1. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, New York, NY, 2020). Fourth edition.
2. S. Bringsjord, The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself, *Journal of Applied Logic* **6**, 502 (2008) `http://kryten.mm.rpi.edu/SB\_LAI\_Manifesto\_091808.pdf`.
3. H. Levesque and G. Lakemeyer, Chapter 24: Cognitive Robotics, in *Handbook of Knowledge Representation*, (Elsevier, Amsterdam, The Netherlands, 2007).
4. M. Ashcraft and G. Radvansky, *Cognition* (Pearson, London, UK, 2013). This is the 6th edition.
5. S. Z. Hashemi, R. Faiez, L. C. Baldor and J. Krauss, Kabul airport attack kills 60 afghans, 13 us troops, *AP News* (2021) `https://apnews.com/article/europe-france-evacuations-kabul-9e457201e5bbe75a4eb1901fedeee7a1`.

6. M. Fitting, Intensional Logic, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2015) https://plato.stanford.edu/entries/logic-intensional.
7. P. Dung, On the Acceptability of Arguments and its Fundamental Fole in Nonmonotonic Reasoning, Logic Programming and N-Person Games, *Artificial Intelligence* **77**, 321 (1995) .
8. H. D. Ebbinghaus, J. Flum and W. Thomas, *Mathematical Logic (second edition)* (Springer-Verlag, New York, NY, 1994).
9. J. Paris and A. Vencovská, *Pure Inductive Logic* (Cambridge University Press, Cambridge, UK, 2015).
10. G. Johnson, *Argument & Inference: An Introduction to Inductive Logic* (MIT Press, Cambridge, MA, 2016).
11. M. Ashcraft, *Human Memory and Cognition* (HarperCollins, New York, NY, 1994).
12. E. B. Goldstein, *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (Cengage Learning, Boston, MA, 2008). This is the 5th edition.
13. S. Bringsjord, N. S. Govindarajulu, J. Licato and M. Giancola, Learning *Ex Nihilo*, in *GCAI 2020. 6th Global Conference on Artificial Intelligence*, , EPiC Series in Computing Vol. 72 (EasyChair Ltd, Manchester, UK, 2020).
14. 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)*, ed. C. Sierra (International Joint Conferences on Artificial Intelligence, 2017).
15. K. Arkoudas and S. Bringsjord, Propositional Attitudes and Causation, *International Journal of Software and Informatics* **3**, 47 (2009) http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf.
16. S. Bringsjord and N. Sundar Govindarajulu, Rectifying the Mischaracterization of Logic by Mental Model Theorists, *Cognitive Science* **44**, p. e12898 (2020) https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12898.
17. S. Bringsjord, M. Giancola and N. S. Govindarajulu, Logic-Based Modeling of Cognition, in *The Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, Cambridge, UK, (forthcoming)) http://kryten.mm.rpi.edu/Logic-basedComputationalModelingOfCognition.pdf.
18. S. Bringsjord, Declarative/Logic-Based Cognitive Modeling, in *The Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, Cambridge, UK, 2008) pp. 127–169. This URL goes to a preprint only. http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf.
19. K. Arkoudas and S. Bringsjord, Vivid: An AI Framework for Heterogeneous Problem Solving, *Artificial Intelligence* **173**, 1367 (2009) http://kryten.mm.rpi.edu/KA_SB_Vivid_offprint_AIJ.pdf.
20. S. Bringsjord, N. S. Govindarajulu and M. Giancola, Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments, *Paladyn, Journal of Behavioral Robotics* **12**, 310 (2021) https://doi.org/10.1515/pjbr-2021-0009.
21. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterdam, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version. .
22. D. Prawitz, The Philosophical Position of Proof Theory, in *Contemporary Philosophy in Scandinavia*, eds. R. E. Olson and A. M. Paul (Johns Hopkins Press, Baltimore, MD, 1972) pp. 123–134 .
23. P. Schroeder-Heister, Proof-Theoretic Semantics, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2018) https://plato.stanford.edu/entries/proof-theoretic-semantics.
24. S. Bringsjord, J. Hendler, N. Govindarajulu, R. Ghosh and M. Giancola, The (Uncomputable!) Meaning of Ethically Charged Natural Language, for Robots, and Us, From Hypergraphical Inferential Semantics, in *Trustworthy Artifical-Intelligent Systems*, ed. I. Ferreira (Springer, Cham, Switzerland, (forthcoming)) This URL goes to a preprint only. http://kryten.mm.rpi.edu/UncomputableNLURobots032421.pdf.
25. N. Francez, *Proof-theoretic Semantics* (College Publications, London, UK, 2015).
26. T. Anderson, D. Schum and W. Twining, *Analysis of Evidence* (Cambridge University Press, Cambridge, UK, 2009). This is the 3rd edition.
27. J. Wigmore, *Science of Judicial Proof, as Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials* (Little, Brown and Co., Boston, MA, 1937).
28. S. E. Preston and R. S. Taylor, *Department of Defense Law of War Manual*, tech. rep., General Counsel of the Department of Defense Washington United States (2016).
29. N. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the Doctrine of Double

Effect: A Formal Model of True Self-Sacrifice, in *Robots and Well-Being*, eds. M. Ferreira, J. Sequeira, G. Virk, M. Tokhi and E. KadarIntelligent Systems, Control and Automation: Science and Engineering (Springer, Basel, Switzerland, 2019) pp. 39–54. The URL here is to a rough preprint. `http://kryten.mm.rpi.edu/NSG\_SB\_RG\_MP\_DDE\_SelfSac\_110617.pdf`.
30. S. Bringsjord, N. Govindarajulu and M. Giancola, Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments, *Paladyn, Journal of Behavioral Robotics* **12**, 310 (2021), The URL here goes to a *rough*, *uncorrected*, *truncated* preprint as of 071421. `http://kryten.mm.rpi.edu/AutomatedArgumentAdjudicationPaladyn071421.pdf`.
31. M. Aikins, C. Koettl, E. Hill and E. Schmitt, NY Times Investigation: "In U.S. Drone Strike, Evidence Suggests No ISIS Bomb", *New York Times* (2021) `https://www.nytimes.com/2021/09/10/world/asia/us-air-strike-drone-kabul-afghanistan-isis.html`.
32. S. D. Said and J. F. Kirby, Pentagon Press Secretary John F. Kirby and Air Force Lt. Gen. Sami D. Said Hold a Press Briefing(November 2021).
33. J. McCarthy, Circumscription—A Form of Non-Monotonic Reasoning, *Artificial Intelligence* **13**, 27 (1980) .
34. J. Pollock, Defasible Reasoning with Variable Degrees of Justification, *Artificial Intelligence* **133**, 233 (2001) .
35. J. L. Pollock, How to Reason Defeasibly, *Artificial Intelligence* **57**, 1 (1992) `citeseer.ist.psu.edu/pollock92how.html`.
36. J. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person* (MIT Press, Cambridge, MA, 1995).
37. R. Reiter, A Logic for Default Reasoning, *Artificial Intelligence* **13**, 81 (1980) .
38. J. Licato, Formalizing Deceptive Reasoning in Breaking Bad: Default Reasoning in a Doxastic Logic, in *2015 AAAI Fall Symposium Series*, 2015.
39. S. Modgil and H. Prakken, The ASPIC$^+$ Framework for Structured Argumentation: A Tutorial, *Argument & Computation* **5**, 31 (2014) .
40. S. Bringsjord and J. Licato, By *Dis*analogy, Cyberwarfare is Utterly New, *Philosophy and Technology* **28**, 339 (2015) `http://kryten.mm.rpi.edu/SB\_JL\_cyberwarfare\_disanalogy\_DRIVER\_final.pdf`.
41. 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*, eds. P. Baroni, D. Gabbay, M. Giacomin and L. Van der Torre, (8) (College Publications, 2017) pp. 2623–2705 .
42. A. Herzig, E. Lorini, J. F. Hübner and L. Vercouter, A logic of trust and reputation, *Logic Journal of the IGPL* **18**, 214 (2010) .
43. R. Feldman and E. Conee, Evidentialism, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* **48**, 15 (1985) .
44. D. Walton, *Witness Testimony Evidence: Argumentation, Artificial Intelligence, and Law* (Cambridge University Press, Cambridge, UK, 2008).
45. D. Walton, *Character Evidence: An Abductive Theory* (Springer, Dordrecht, The Netherlands, 2010).
46. P. Forrest, The Identity of Indiscernibles, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2020) Winter 2020 edn. .
47. 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)*, eds. M. Suda and S. Winkler, Electronic Proceedings in Theoretical Computer Science, Vol. 311 (Open Publishing Association, Waterloo, Australia, 2019) pp. 43–49. The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/. `http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf`.
48. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Ethical Reasoning for Autonomous Agents Under Uncertainty, in *Smart Living and Quality Health with Robots • Proceedings of ICRES 2020*, eds. M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang and A. Kaur (CLAWAR, London, UK, September 2020). Paper vailable at the URL given above. The ShadowAdjudicator system can be obtained here: *https://github.com/RAIRLab/ShadowAdjudicator*.
49. P. Hayes, The Naïve Physics Manifesto, in *Expert Systems in the Microeletronics Age*, ed. D. Mitchie (Edinburgh University Press, Edinburgh, Scotland, 1978) pp. 242–270 .
50. M. Shanahan, An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense

Reasoning, *Artificial Intelligence* **153**, 141 (2004) .
51. M. Shanahan, The Event Calculus Explained, in *Artificial Intelligence Today (LNAI 1600)*, eds. M. Wooldridge and M. Veloso (Springer, New York, NY, 1999) pp. 409–430 .

ISBN: 978-1-7396142-0-1