REASONING WITH *COGNITIVE LIKELIHOOD* FOR ARTIFICIALLY-INTELLIGENT AGENTS: FORMALIZATION & IMPLEMENTATION

Michael Giancola

Submitted in Partial Fullfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Approved by:
Selmer Bringsjord, Chair
Sergei Nirenburg
Carlos Varela
Paul Bello
Naveen Sundar Govindarajulu

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

[May 2023]
Submitted April 2023

CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

To my parents, for always encouraging me to follow my own path.

ACKNOWLEDGMENT

First and foremost I'd like to express my gratitude to my advisor Professor Selmer Bringsjord. Selmer, the guidance you provided and the doors you opened for me during my doctoral studies have undoubtedly enhanced my training and altered the path of my professional life for the better. Lærere som deg er ikke lett åfinne. *Takk så mye!*

Next I'd like to thank the remaining members of my doctoral committe — Professor Sergei Nirenburg, Professor Carlos Varela, Dr. Paul Bello, and Dr. Naveen Sundar Govindarajulu — for their advice and feedback on my doctoral work. Particular thanks are owed to Naveen for his technical assistance with his ShadowProver automated reasoner, without which this dissertation would not have been possible. I'd also like to thank my labmates for many stimulating conversations which have influenced my perspective on the field of Artificial Intelligence at large. In particular, I'm very grateful to Brandon Rozek who proofread several earlier drafts of this dissertation.

Given that this is a major milestone in my research career, I would be remiss not to acknowledge Professor Jacob Whitehill. You taught me how to conduct rigorous research and helped me publish my first conference paper. You also provided keen advice throughout my journey to graduate school, without which I wouldn't have gotten the chance to start this dissertation.

This dissertation could not have been completed without financial support by the Office of Naval Research (ONR) (Award #'s N00014-19-1-2558; N00014-17-1-2115; N00014-22-1-2201) and the Air Force Office of Scientific Research (AFOSR) (Award # FA9550-17-1-0191); I am grateful for their support throughout my doctoral studies.

Finally, to my friends and family, especially my parents, to whom this dissertation is dedicated. All of your love and moral support was imperative to my success in graduate school. In particular, to my Aunt Angelina, who passed away during my studies. Thank you for the fresh veggies you shared from your garden and passing on some of your skill in the kitchen (my *pasta fasul* will never come close to yours).

You, the people, have the power,
the power to create machines,
the power to create happiness.
You, the people, have the power,
to make this life free and beautiful,
to make this life a wonderful adventure.
...
Let us fight for a world of reason,
a world where science and progress
will lead to all men's happiness.
Soldiers!
In the name of democracy, let us all unite!

— from "The Final Speech", by Charlie Chaplin

ABSTRACT

Human beings routinely encounter situations containing informal, non-quantitative uncertainty. Consider for example the following scenario: Driving toward a four-way intersection, you stop at a red light. Eventually, the light turns green, but you perceive a driver approaching from your left, their light having turned red moments ago, and subsequently perceive their car accelerate. What can we say about this situation? It certainly seems likely that the driver will drive straight through the light. Of course, it's entirely possible that the driver will change their trajectory at the last second and slam on the brakes. How can we *quantify* this uncertainty (assuming this is what we desired)?

We could compute a probability over all recorded instances of drivers accelerating toward red lights and either going through or stopping. But clearly humans don't engage in anything like this computation when they reason about other drivers on the road. We use *likelihoods* to express *qualities* (as opposed to *quantities* e.g. probabilities) of the uncertainty of beliefs. In this way, one may reason that "I believe it's *highly likely* that the driver will drive through the red light" and subsequently come to the conclusion that, despite having the legal right-of-way, one should wait to avoid an accident. Autonomous agents, in order to effectively interact with humans that reason this way, will need to possess and exploit the ability to model reasoning with notions of *qualitative* uncertainty.

The present dissertation introduces *Cognitive Likelihood*, a framework for reasoning with uncertain beliefs. The framework is implemented within a novel logic — the Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$) — which includes a formal grammar and semantics which dictate how agents can reason within the framework. These formalisms are implemented in an automated reasoner called ShadowAdjudicator in order to enable the automatic generation of $\mathcal{IDCEC}$ proofs. We present the novel algorithm underlying ShadowAdjudicator which enables this automated proof discovery. Finally, we demonstrate how these contributions can be utilized to solve autonomous driving problems and to adjudicate arguments regarding a notorious probability puzzle, the Monty Hall Problem.

CHAPTER 1
INTRODUCTION

> "The test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and still retain the ability to function."
>
> —F. Scott Fitzgerald

Human beings routinely encounter situations containing informal, non-quantitative[1] uncertainty. Consider for example the following scenario: Driving toward a four-way intersection, you stop at a red light. Eventually, the light turns green, but you perceive a driver approaching from your left, their light having turned red moments ago, and subsequently perceive their car accelerate. What can we say about this situation? It certainly seems likely that the driver will drive straight through the light. Of course, it's entirely possible that the driver will change their trajectory at the last second and slam on the brakes. How can we *quantify* this uncertainty (assuming this is what we desired)?

We could compute a probability over all recorded instances of drivers accelerating toward red lights and either going through or stopping. But clearly humans don't engage in anything like this computation when they reason about other drivers on the road. We use *likelihoods* to express *qualities* (as opposed to *quantities* e.g. probabilities) of the uncertainty of beliefs. In this way, one may reason that "I believe it's *highly likely* that the driver will drive through the red light" and subsequently come to the conclusion that, despite having the legal right-of-way, one should wait to avoid an accident.

Autonomous agents, in order to effectively interact with humans that reason this way, will need to possess and exploit the ability to model reasoning with notions of *qualitative* uncertainty. To further justify this position, consider the following example.

[1] This isn't to say that once the concept of "Cognitive Likelihood" is formalized, that there won't be quantitative/numerical elements. The point to be made is that frameworks like e.g. probability theory are *chiefly* quantitative/numerical in nature. That is, reasoning within probability theory is largely done via some sort of numerical computation. Alternatively, the type of reasoning we propose herein will be achieved largely through automated reasoning over declarative content.

1.1 A Motivating Example

While the following autobiographical example illustrates a capacity for creative argumentation beyond what I propose for my dissertation, I think it is highly provocative for illustrating the long-term utility of the proposed work. While reading Ken Forbus' *Qualitative Representations*, I came upon the following sentence:

> One important consideration for cognitive models is ability (Cassimatis, Bello, & Langley, 2008)–that is, can they actually perform the task that they are trying to model, at the levels that people do? (pg. 9, [46])

Content aside, what caught my eye was the citation. Instantly, I thought, "Oh, that must be Nick Cassimatis and Paul Bello". Of course, we can be a bit casual with words like "must," as there certainly could be two other people with those surnames who authored a paper together in 2008. I really meant something more like "I believe it is *highly likely* that Cassimatis and Bello refer to Nick Cassimatis and Paul Bello." But how can I support such an argument with the given level of uncertainty? Well,

- I know their surnames are relatively uncommon;
- I know they were both at one time associated with the Rensselaer AI & Reasoning (RAIR) Lab, so it is likely they know each other and have collaborated; and
- I know that they work in the area of computational cognitive modeling and that they were doing that kind of work circa 2008.

Interested in strengthening my argument, I checked the References. The full citation only listed a "N. Cassimatis" and a "P. Bello". While it is still not certain, this evidence presumably strengthens my argument to the level of *overwhelmingly likely* that 'N' stands for "Nick" and 'P' for "Paul." Subsequently, I searched the citation online to confirm my suspicions, at which point it was *evident* that Nick and Paul authored the publication.[2]

This example is far from idiosyncratic; humans reason in this manner frequently. The general idea I'm pointing to here is that we humans are capable of constructing arguments in favor of uncertain propositions, and describing their relative strength using qualitative

[2]Why not *certain*? While the likelihood is infinitesimal, it is possible that the Nick Cassimatis and Paul Bello in the citation are actually different people than the ones I was thinking of.

Desiderata '$\mathcal{D}$'

- d_1 be **defeasible** (and hence nonmonotonic) in nature (when new information comes to light, past reasoning is retracted in favor of new reasoning with new conclusions);
- d_2 be able to **resolve "cognitive" inconsistencies** (i.e. an agent believing ϕ and $\neg\phi$) 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 (the latter case giving rise to multi-valued inductive logic);
- d_4 be **argument-based**, where the arguments have internal inference-to-inference structure, so that justification (and hence explanation) is available;
- d_5 have **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 mental states** including knowledge, belief, obligation, perception, communication, etc. of relevant artificial and human agents (these elements are irreducibly intensional, which will be discussed further in §2.2);
- 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.*[a]

[a] I.e. the problem of a machine/system at the level of a Turing machine deciding whether a given formula in first-order logic is a theorem or not. Alonzo Church settled this in the negative—that is, that a Turing machine cannot determine the theoremhood of any arbitrary first-order logic formula. We refer to this result today as "Church's Theorem."

Figure 1.1: List of Desired Attributes for AI Agents

metrics like "evident" and "certain." Should we expect autonomous agents to meaningfully communicate and deliberate alongside us someday, the capability to reason with qualitative uncertainty is a crucial prerequisite.

1.2 Desiderata

We desire a framework for artificially-intelligent agents which satisfies the set of desiderata given in Figure 1.1. Next, we defend our desire for each desideratum.

d_1 As will be discussed in greater detail in §2.1, the real world is simply not monotonic. More specifically, any agent operating in the world without perfect knowledge, instead gaining knowledge by perceiving its environment and reasoning about its perceptions, is certain to generate a belief which it will retract/revise at a later time. For example, consider the possibility that, in the motivating example presented earlier (§1.1), it turned out that the Cassimatis and Bello referenced in the parenthetical citation actually referred to different people than I initially expected. Upon that discovery I would've, quite quickly, retracted my previous (weak) belief for a strong belief that in fact, those surnames referred to different people than I initially believed. Defeasible/non-monotonic reasoning is exactly what is required in order to formally model this type of belief revision.

d_2 There are many reasons an agent could come to believe a proposition and its negation. Perhaps the agent generated a belief in ϕ from a direct perception, and a belief in $\neg\phi$ from communication with another agent. But of course, one might expect that a rational agent would believe the objects of direct perceptions more strongly than communicated information (since an agent could be e.g. incompetent, deceptive, etc.). This leads naturally to the next desideratum.

d_3 Any formal approach to adjudicating contradictory beliefs must have some mechanism for measuring the (potentially relative) level of (un-)certainty in an agent's beliefs.

d_4 An entire dissertation could be written to defend the need of explainability/verifiability in AI systems. Briefly, we see three reasons (although there are certainly many more): (1) when the agent makes a mistake, explainability can help the developers understand why and how to improve the agent, (2) verifiability can help put end users at ease (e.g. when riding in an autonomous car), and (3) both can help prevent bias in decision-making, and identify if an agent does make a biased decision.

d_5 Requiring inference schemata as the means of reasoning ensures that the verification process is fast and that the resulting explanations are understandable by humans (as opposed to e.g. model-based semantics which, while providing verification, cannot produce a proof/argument that humans can intuitively understand).

d_6 The ability to reason with socio-cognitive mental states such as belief, obligation, perception, etc. enable agents to both model the reasoning of the humans they interact with, as well as explain their own reasoning to those humans.

d_7 Our final desideratum is, at first glance, the most contentious to most readers.[3] However, it is absolutely necessary, in order to enable agents to incorporate meta-logical expressions in their inference schemata, such as statements about provability. For example, the schema below enables an agent a to infer a belief in any formula ϕ which is provable from its belief set Γ.[4]

$$\frac{\mathbf{B}(a, t_1, \Gamma), \quad \Gamma \vdash \phi, \quad t_1 \leq t_2}{\mathbf{B}(a, t_2, \phi)} \; [I_{\mathbf{B}}] \tag{1.1}$$

Furthermore, it also enables agents to directly hold cognitive attitudes toward provability statements. For example, the following formula expresses that agent a says to agent b that a believes that ϕ is provable from Γ.

$$\mathbf{S}(a, b, t, \mathbf{B}(a, t, \Gamma \vdash \phi)) \tag{1.2}$$

1.3 Contributions

The major contributions of this dissertation are as follows:

- Formalize a cognitive calculus — which we will call the *Inductive Deontic Cognitive Event Calculus* ($\mathcal{IDCEC}$) — which meets the list of desiderata $\mathcal{D}$ given in §1.2.
- Implement the calculus in an automated reasoner. That is, encode the grammar and inference schemata of the calculus in order to generate proofs in the calculus.
- Devise a novel algorithm for finding proofs in the calculus.
- Apply this proof-finding algorithm to solve non-trivial reasoning problems.

[3] For a discussion of some potential objections and our rebuttals, see §9.2.2.

[4] To further clarify d_7, we note specifically that what it entails is that there must be provisions made for *trying* to solve *particular instances* of Turing-unsolvable problems. Clearly, as a result of Church's Theorem, this will not always be possible. For example, in Equation 1.1, it may be the case that ϕ is provable from Γ but a proof cannot be found. Desideratum d_7 only requires that an automated reasoner make a genuine effort to search for such a proof.

CHAPTER 2
BACKGROUND

This chapter will provide all background content necessary for understanding the rest of the dissertation. In particular, a basic understanding of cognitive calculi will be crucial preparation for comprehending the main content. But first, we start with a discussion of a few concepts which are relevant to cognitive calculi.

2.1 Deductive vs. Defeasible Reasoning

To start, we must make clear the central distinction between two forms of reasoning: deductive, specifically monotonic; versus inductive, specifically nonmonotonic/defeasible. Deductive reasoning is *monotonic*, in that if a (declarative) formula ϕ in some logic can be deduced from some set Φ of formulae (traditionally written $\Phi \vdash_I \phi$, where the subscript I is assigned to some particular set of inference schemata for precise deductive reasoning), then for any formula $\psi \notin \Phi$, it remains true that $\Phi \cup \{\psi\} \vdash_I \phi$. In other words, in the case of deductive reasoning, the arrival of new declarative information never invalidates prior reasoning. More formally, the closure of Φ under standard deduction (i.e., the set of all formulae that can be deduced from Φ via I), denoted by $\Phi_I^{\vdash}$, is guaranteed to be a subset of $(\Phi \cup \Psi)_I^{\vdash}$, for all sets of formulas Ψ. Inductive logics (e.g., see those in Paris and Vencovská [88], Johnson [65]) don't work this way, and that's a welcome fact, since much of real life doesn't conform to monotonicity, at least when it comes to the cognition of humans; this is easy to see:

To use a well-known example from the literature on nonmonotonic logic,[5] suppose that Tweety is a bird. Given this, many will deduce (or at least be tempted to do so) that Tweety

Portions of this chapter previously appeared as:

1. Selmer Bringsjord, Naveen Sundar Govindarajulu, and Michael Giancola. 2021. Automated argument adjudication to solve ethical problems in multi-agent environments. In *Paladyn, J. of Behavioral Robotics*, Vol. 12. De Gruyter, Berlin, Boston, 310–335. https://doi.org/10.1515/pjbr-2021-0009

2. Selmer Bringsjord, Naveen Sundar Govindarajulu, John Licato, and Michael Giancola. 2020. Learning ex nihilo. In *Proceedings of the Sixth Global Conference on Artificial Intelligence (GCAI 2020) (EPiC Series in Computing, Vol. 72)*. Gregoire Danoy, Jun Pang, and Geoff Sutcliffe (Eds.). EasyChair, Manchester, UK, 1-27. https://doi.org/10.29007/ggcf

[5]The example is often discussed in the context of AI which incorporates nonmonotonic logic. See e.g. the extended treatment in Genesereth and Nilsson [48].

can fly, on the strength of a general principle saying that birds can fly. But if it's learned that Tweety is a penguin, the situation is defeated in favor of a new one: that Tweety can fly should now no longer be among the propositions that have been arrived at by sound reasoning. Nonmonotonic reasoning is the form of reasoning designed to model, formally, this kind of *defeasible* inference.

2.2 Extensional vs. Intensional Logic

Elementary, classical deductive logics are invariably **extensional**, which means, in a nutshell, that the semantic values of formulae in these logics are a deterministic compositional function of their sub-parts. For example, if the formula $\phi \wedge \psi$ in first-order logic $\mathscr{L}_1$ is true, then the formula ϕ must be true as well. In stark contrast, **intensional** logics of the sort that that we exploit herein, because e.g. they provide machinery for capturing epistemic attitudes such as *knowing*, *believing*, etc., have formulae whose semantic values cannot be determined from the sub-parts of these formulae. For instance, if it's true that $L(a, b)$, because Abbie loves Beatrice and that is what this atomic formula represents, there is no guarantee that Beatrice believes that Abbie loves Beatrice; that is, there is no guarantee that $\mathbf{B}(b, (L(a, b))$. Beatrice may even believe the opposite, that is, $\neg L(a, b)$. To handle intensional attitudes, modal logics were introduced; the reader is directed to Bringsjord et al. [21] for coverage of such logics.

Put simply, intensional logic is about the distinction between what a term *means* versus what it *denotes*. To employ a classic (indeed, as a matter of fact the first) example in intensional logic, the terms "morning star" and "evening star" both denote the planet Venus. However, they have different meanings; one is seen in the morning, one in the evening, and humans have held beliefs about each without in any way realizing that such epistemic attitudes pertain to the very same object. Extensional logics do not distinguish between two such concepts, so if two terms denote the same thing, then they must have the same meaning. This is fine for reasoning about e.g. the abstract objects of mathematics, but is unacceptable for modeling cognitive attitudes, as discussed in the prior paragraph.[6]

[6]The interested reader is referred to Fitting [45] for a comprehensive review of intensional logics.

2.3 What is a Cognitive Calculus? And Why is it So Named?

What is a cognitive calculus $\mathscr{C}$,[7] and why is it denoted with the two words in question? In keeping with the mathematical-logic literature (e.g. Ebbinghaus et al. [39]), 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[8] 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.[9] In the case of both of these families, a frequently included particular inference schema is *modus ponens*, that is

$$\frac{\phi \rightarrow \psi, \phi}{\psi} \; I_{\text{MP}} \tag{2.1}$$

And in the case of the latter family, often *universal introduction* is included in a given $\mathcal{I}$; a specification of this inference schema immediately follows.[10]

$$\frac{\phi(\text{a})}{\forall x \phi(\frac{a}{x})} \; I_{\text{UI}} \tag{2.2}$$

Note that both of the two inference schemata just shown are included in the particular cognitive calculi $\mathcal{DCEC}$ and $\mathcal{IDCEC}$ we employ in the present dissertation for modeling, and as a framework for automated reasoning. Note as well that both $\mathscr{L}_{\text{PC}}$ and $\mathscr{L}_1$ are *extensional* (see §2.2). 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

[7] We use '$\mathscr{C}$' here as an arbitrary variable ranging over (the uncountably infinite space of) all cognitive calculi. The particular cognitive calculi that are discussed in this dissertation are $\mathcal{DCEC}$ and $\mathcal{IDCEC}$.

[8] Hence when the schemata are deductive in nature, we specifically have natural deduction.

[9] As can be confirmed by looking to the main textbooks of the field. E.g. see Russell and Norvig [104], Luger [75].

[10] The standard provisos apply here to the constant a.

Ashcraft [5], Goldstein [55]), 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.[11] 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.) Note that such verbs break the bounds of extensionality, and hence make any logic that captures them an *intensional* logic.[12] Step S1.i is the reason why we speak of a *cognitive* calculus.

ii meta-logical expressions (such as that from a set Φ of formulae a particular formula ϕ can be proved: $\Phi \vdash \phi$). Hence cognitive calculi include not merely object-level elements of logics, but meta-logical elements as well. E.g. a cognitive calculus can have a meta-conditional saying that if some provability expression such as $\Phi \vdash \phi$ holds, then ϕ holds (see, for example, the schema $I_{\mathbf{B}}$ in §2.4.2). 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 meta-logical expressions secured by prior step S1.ii. S2 reflects the fact that cognitive calculi have purely *inferential* semantics,[13] and hence are aligned with the tradition of *proof-theoretic semantics* [49, 97, 106]. We might for instance wish to include an inference schema that regiments the idea that an agent knows that which is provable from what they know. Step S2 is the reason why we speak of a cognitive *calculus* (instead of e.g. a cognitive *logic*, or cognitive logical system).

[11] The attitudes are covered nicely in Nelson [86]. 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. (¶1, [86]).

[12] This fact is discussed in some detail in Bringsjord and Govindarajulu [23], 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.

[13] There are many reasons we employ inference-theoretic semantics instead of model-theoretic semantics. One of the most significant reasons — for this work in particular — is that inference schemata enable us to express how the semantics of many modalities (e.g., perception, belief, obligation, etc.) interact, which is not possible in any model theory known to the author. For further discussion of the drawbacks of model theory and model-theoretic semantics, we point the interested reader to §4 of Giancola et al. [53], which specifically discusses model- vs. proof-theoretic semantics in the context of ethical reasoning (although the arguments are relevant to reasoning more broadly).

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

$$\frac{\mathbf{K}\phi}{\mathbf{B}\phi} \; I_{\text{KB}} \tag{2.3}$$

2.4 Deontic Cognitive Event Calculus

The Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) is a quantified, multi-modal, sorted cognitive calculus. In general, a cognitive calculus consists of two main pieces: the signature and the set of inference schemata. We discuss the specific signature and inference schemata of $\mathcal{DCEC}$ next.

2.4.1 Signature

The signature has four components: (1) a set of sorts, (2) a set of function signatures, (3) a grammar for terms, and (4) a grammar for syntactic forms. Note that each of these components builds off of a pre-existing core. The sorts and function signatures build off of the standard, extensional event calculus[14] [70], and the terms and syntactic forms build off of first-order logic.

Table 2.1 shows the sorts used in $\mathcal{DCEC}$ and their descriptions. Among these, the Agent, Action, and ActionType sorts are not native to the event calculus, first modified and extended in the cognitive direction by Arkoudas and Bringsjord [3].

[14]Other calculi (e.g. the *situation calculus*) for modeling commonsense and physical reasoning can be easily switched out in-place of the event calculus.

Table 2.1: The Sorts in $\mathcal{DCEC}$

Sort	Description
Agent	Human and non-human actors.
ActionType	Action types are abstract actions. They are instantiated at particular times by actors. Example: eating.
Action	A subtype of Event for events that occur as actions by agents.[15]
Event	Used for events in the domain.
Moment	The Moment type stands for time in the domain. They can be simple, such as t_i, or complex, such as $birthday(son(jack))$.[16]
Fluent	Used for representing states of the world in the event calculus.

The modal operators present in the calculus include the standard operators for knowledge **K**, belief **B**, desire **D**, intention **I**, etc. The general format of an intensional operator is $\mathbf{K}(a, t, \phi)$, which says that agent a knows at time t the proposition ϕ. Here ϕ can in turn be any arbitrary formula. Also, note the following modal operators: **P** for perceiving a state, **C** for common knowledge, **S** for agent-to-agent communication and public announcements, and finally and crucially, a dyadic (arity = 2) deontic operator **O** that states when an action is obligatory or forbidden for agents. It is well known that the unary ought in standard deontic logic leads to contradictions. Our dyadic version of the operator blocks the standard list of such contradictions.[17]

[15] Actions are events that are carried out by an agent. For any action type α and agent a, the event corresponding to a carrying out α is given by $action(a, \alpha)$. For instance if α is *"running"* and a is *"Jack"* , $action(a, \alpha)$ denotes *"Jack is running"*.

[16] If desired, we could remove this sort and instead integrate $\mathcal{DCEC}$ with a first-order temporal logic [56]. Thereby, instead of indicating that a formula holds at some particular Moment in time, we could express intervals of time over which some formula holds. These intervals can be unbounded, e.g., $\forall t\ \phi$ ("ϕ always holds.") or bounded, e.g., $\exists t\ (t_1 < t) \wedge (t < t_2) \wedge \phi$ ("ϕ holds sometime between t_1 and t_2."). Integration of cognitive calculi with temporal logics is not explored in the present dissertation.

[17] An overview of this list is given lucidly in McNamara [80].

$\mathcal{DCEC}$ Signature

$$
\begin{aligned}
S &::= \mathsf{Agent} \mid \mathsf{ActionType} \mid \mathsf{Action} \sqsubseteq \mathsf{Event} \mid \mathsf{Moment} \mid \mathsf{Fluent} \\
f &::= \begin{cases}
action : \mathsf{Agent} \times \mathsf{ActionType} \rightarrow \mathsf{Action} \\
initially : \mathsf{Fluent} \rightarrow \mathsf{Formula} \\
holds : \mathsf{Fluent} \times \mathsf{Moment} \rightarrow \mathsf{Formula} \\
happens : \mathsf{Event} \times \mathsf{Moment} \rightarrow \mathsf{Formula} \\
clipped : \mathsf{Moment} \times \mathsf{Fluent} \times \mathsf{Moment} \rightarrow \mathsf{Formula} \\
initiates : \mathsf{Event} \times \mathsf{Fluent} \times \mathsf{Moment} \rightarrow \mathsf{Formula} \\
terminates : \mathsf{Event} \times \mathsf{Fluent} \times \mathsf{Moment} \rightarrow \mathsf{Formula} \\
prior : \mathsf{Moment} \times \mathsf{Moment} \rightarrow \mathsf{Formula}
\end{cases} \\
t &::= x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
\phi &::= \begin{cases}
q : \mathsf{Formula} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : \phi(x) \mid \exists x : \phi(x) \\
\mathbf{P}(a, t, \phi) \mid \mathbf{K}(a, t, \phi) \mid \mathbf{S}(a, b, t, \phi) \mid \mathbf{S}(a, t, \phi) \\
\mathbf{C}(t, \phi) \mid \mathbf{B}(a, t, \phi) \mid \mathbf{D}(a, t, \phi) \mid \mathbf{I}(a, t, \phi) \\
\mathbf{O}(a, t, \phi, (\neg) happens(action(a^*, \alpha), t'))
\end{cases}
\end{aligned}
$$

Figure 2.1: Signature of the Deontic Cognitive Event Calculus

Table 2.2: The Modal Operators in $\mathcal{DCEC}$

Operator	Description
P	Perceives
K	Knows
S	Says
C	Commonly Known
B	Believes
D	Desires
I	Intends
O	Ought-to

Finally, it should be noted that $\mathcal{DCEC}$ is one specimen in a *family* of easily extensible cognitive calculi, as we endeavor to do herein by extending $\mathcal{DCEC}$ in order to support *inductive* argumentation.

2.4.2 Inference Schemata

The figure below shows the inference schemata for $\mathcal{DCEC}$. $I_{\mathbf{K}}$ and $I_{\mathbf{B}}$ are inference schemata that let us model idealized agents that have their knowledge and belief closed under the $\mathcal{DCEC}$ proof theory. While normal humans are not deductively closed, this lets us model more closely how deliberate and strategic agents reason, such as organizations and governments.[18] I_1 and I_2 state respectively that it is common knowledge that perception leads to knowledge, and that it is common knowledge that knowledge leads to belief. I_3 defines common knowledge as unbounded nested knowledge (e.g. common knowledge of ϕ implies that I know that you know that he knows that she knows $\ldots\phi$). I_4 states that knowledge of a proposition implies that the proposition holds. I_5 to I_{10} indicate forms of reasoning which are common knowledge. I_{11a} and I_{11b} enable agents to apply modus ponens and conjugation introduction to formulae which they believe.[19] I_{12} states that if an agent s communicates a proposition ϕ to h, then h believes that s believes ϕ.[20] I_{14} dictates how obligations get translated into intentions.

2.5 Automated Reasoning in Cognitive Calculi

ShadowProver is an automated reasoner for $\mathcal{DCEC}$ that is under continuous development and available[21] under an open-source license [62]. We leave further discussion of ShadowProver to §5.1.

Parallel with the development of ShadowProver is an automated reasoner which supports inductive reasoning/argumentation: ShadowAdjudicator. For early work showcasing ShadowAdjudicator, see Giancola et al. [53]. Further discussion of the development of ShadowAdjudicator is left to §5.

[18] To more faithfully model normal human reasoning, some dialects of cognitive calculi restrict the number of iterations on intensional operators. As an example to illustrate what we mean: if we were to restrict the number of iterations on intensional operators to 3, this would mean we can express "Alice believes Bob believes Carly believes ϕ.", but not "Alice believes Bob believes Carly believes Dave believes ϕ."

[19] Note that these two schemata are subsumed by $I_{\mathbf{B}}$. However, their inclusion is still useful to make some proofs more understandable. I.e. an inference step using I_{11a} provides more explicit explanation than the same one using $I_{\mathbf{B}}$. Also, they can be useful in cognitive calculi which are concerned with modeling irrational human reasoning. That is, humans may use some schemata and not others when reasoning about their beliefs.

[20] This schema captures the assumption in $\mathcal{DCEC}$ that agents are not deceptive. In order to model deceptive agents, this schema would need to be replaced by a more nuanced treatment of the $\mathbf{S}$ operator. See Licato [73] for work in this direction.

[21] Source code available at `https://github.com/naveensundarg/prover` (Last Accessed March 2, 2023).

$\mathcal{DCEC}$ Inference Schemata

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

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [I_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [I_2]$$

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

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

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

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

$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [I_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [I_9]$$

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

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

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [I_{12}] \qquad \frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t',happens(action(a^*,\alpha),t'))}\ [I_{13}]$$

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

Figure 2.2: Inference Schemata of the Deontic Cognitive Event Calculus

2.6 Likelihood Values

Our approach to measuring the uncertainty of beliefs within cognitive calculi eschews traditional probability values in favor of *likelihood* values. The 11 likelihood values[22] employed in this dissertation are shown in Table 2.3.

Likelihood values can be obtained in either of two ways; both ways immediately reveal that we take likelihood to be *subjective*. The first way is to take as primitive a cognitive binary relation on formulae from the perspective of a rational agent (e.g., ϕ is *more reasonable than* ψ), and then build up formally to the partial or total order in question. This approach

[22] As we will discuss in §3, some of our prior work explored 13-value spectrums of likelihood. We chose this 11-value spectrum as it lends itself to a rather natural definition via inference schemata, which we will unveil and discuss in §4. Prior work introducing other spectrums generally employed either Strength Factors (discussed in §3.2) or didn't fully define the semantics of the 13-value spectrum.

Table 2.3: The 11 Cognitive Likelihood Values

Numerical	Linguistic
5	CERTAIN
4	EVIDENT
3	OVERWHELMINGLY LIKELY = BEYOND REASONABLE DOUBT
2	LIKELY
1	MORE LIKELY THAN NOT
0	COUNTERBALANCED
-1	MORE UNLIKELY THAN NOT
-2	UNLIKELY
-3	OVERWHELMINGLY UNLIKELY = BEYOND REASONABLE BELIEF
-4	EVIDENTLY NOT
-5	CERTAINLY NOT

is first formalized in Govindarajulu and Bringsjord [59] and is deployed in e.g. Giancola et al. [53].

Another approach is to independently justify each likelihood value by appeal to rational human-level cognition. A benefit to this approach is that, in forgoing the reasonableness operator, the meaning of the likelihood values is expressed entirely through the inference schemata they are used in, just as the meaning of all other components of cognitive calculi (e.g. the modal operators) are expressed. The first publication to investigate this direction was Giancola et al. [54]. As will be discussed later in more detail, this is the approach which is pursued and fully fleshed out herein.

2.7 Formal Proofs vs. Formal Arguments

Using a term coined by Marvin Minsky, the words "proof" and "argument" are suitcase words: they are ascribed very different meanings by different people. The present section serves to ground the specific meaning we intend when we discuss formal proofs and arguments throughout this dissertation.

A proof must be absolutely certain. That is, it must start from assumptions which are certain, employ inference schemata which are certain, and finally come to a conclusion which is certain. There can be no notion of uncertainty in a proof. This is why we make use of the term "formal argument" herein. The constraints are similar: a formal argument

must start from a set of assumptions, use an ordered list of inference schemata, and arrive at a conclusion. However, a formal argument *can* contain uncertainty. Therefore, one can generate formal proofs using $\mathcal{DCEC}$, but only formal arguments when we use a cognitive calculus which incorporates likelihood values (as we plan to herein).

This is also why we speak of *argument-theoretic* or *inference-theoretic* semantics. These are essentially the same as traditional proof-theoretic semantics, except that they involve formal arguments, or more abstractly, formal inferences of some kind.

CHAPTER 3
TOWARD COGNITIVE LIKELIHOOD

Next we discuss several use cases which were modeled using uncertainty-infused cognitive calculi of various levels of complexity. They can be grouped into three categories: (1) those which present only argument sketches, (2) those which construct formal arguments via Strength Factors, and (3) those which constructed formal arguments via a Likelihood

---

Portions of this chapter previously appeared as:

1. Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2020. Culturally aware social robots that carry humans inside them, protected by defeasible argumentation systems. In *Culturally Sustainable Social Robotics (Proceedings of Robophilosophy 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 335)*, Marco Nørskov, Johanna Seibt, and Oliver Santiago Quick (Eds.). IOS Press, 440–456. https://doi.org/10.3233/FAIA200941

2. Selmer Bringsjord, Naveen Sundar Govindarajulu, and Michael Giancola. 2021. Automated argument adjudication to solve ethical problems in multi-agent environments. In *Paladyn, J. of Behavioral Robotics*, Vol. 12. De Gruyter, Berlin, Boston, 310–335. https://doi.org/10.1515/pjbr-2021-0009

3. Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2021. A solution to an ethical super dilemma via a relaxation of the doctrine of triple effect. In *Life-world for Artificial and Natural Systems, Proceedings of the Sixth International Conference on Robot Ethics and Standards (ICRES 2021)*, S. Bringsjord, M.O. Tokhi, M.I.A. Ferreira, N.S. Govindarajulu, and M.F. Silva (Eds.). CLAWAR, London, UK, 23–32.

4. Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2022. Novel intensional defeasible reasoning for AI: is it cognitively adequate? In *Proceedings of the IJCAI Workshop on "Cognitive Aspects of Knowledge Representation" (CAKR 2022)*. CEUR-WS.

5. Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and John Licato. 2020. Adjudication of symbolic & connectionist arguments in autonomous driving AI. In *Proceedings of the Sixth Global Conference on Artificial Intelligence (GCAI 2020) (EPiC Series in Computing, Vol. 72)*, Gregoire Danoy, Jun Pang, and Geoff Sutcliffe (Eds.). EasyChair, 28–33. https://doi.org/10.29007/k647

6. Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and Carlos Varela. 2020. Ethical reasoning for autonomous agents under uncertainty. In *Smart Living and Quality Health with Robots, Proceedings of the Fifth International Conference on Robot Ethics and Standards (ICRES 2020)*, M.O. Tokhi, M.I.A. Ferreira, N.S. Govindarajulu, M.F. Silva, E.E. Kadar, J. Wang, and A.P. Kaur (Eds.). CLAWAR, London, UK, 26–41.

7. Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and Carlos Varela. 2022. Making maximally ethical decisions via cognitive likelihood and formal planning. In Towards Trustworthy Artificial Intelligent Systems, Maria Isabel Aldinhas Ferreira and Mohammad Osman Tokhi (Eds.). Springer International Publishing, Cham, Switzerland, 127–142. https://doi.org/10.1007/978-3-031-09823-9_10

Portions of this chapter are to appear in:

1. Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2023. Logic-based modeling of cognition. In *The Cambridge Handbook of Computational Cognitive Sciences*, Ron Sun (Ed.). Cambridge University Press, Cambridge, UK. Forthcoming.

Calculus. Looking ahead, this incremental work ultimately led to the development of a robust cognitive calculus for capturing uncertain belief, which we present in §4.

3.1 Argument Sketches

Our earliest work on $\mathcal{IDCEC}$ consisted of *argument sketches*. As opposed to formal arguments in which each step is warranted by some inference schema, argument sketches show a high-level, conceptual view of the reasoning involved. This was because there were no formal inference schemata for $\mathcal{IDCEC}$, yet.

3.1.1 Adjudicating an Autonomous Driving Scenario

In Giancola et al. [52], we discussed the tragic accident in which the first pedestrian was killed by an autonomous car. Due to several grave errors in its design, it failed to recognize the pedestrian and stop in time to avoid a collision. After describing the situation leading up to the accident, we formalized two competing arguments in $\mathcal{IDCEC}$ which justified making one of two decisions: to apply the brakes, or not.

3.1.1.1 The Tragic Case Study

The accident occurred on a four-lane roadway, and began with the vehicle driving 44 mph in the rightmost lane, and the pedestrian walking a bicycle across the street starting from the (driver's) left side. The vehicle's radar first detected the pedestrian approximately 5.6 seconds before the fatal collision [12]. Less than half a second later, the lidar detects the pedestrian but classifies her as "Other" and as an unmoving object. For the next 2.5 seconds, the lidar re-classifies her several times, alternating between "Vehicle" and "Other". The vehicle's automated-driving system (ADS) attempted to predict her direction of travel several times, but discarded any previous information about her trajectory every time it reclassified her. Finally, with 2.6 seconds until collision, the lidar classifies her as a bicycle, but as it was yet again changing her classification, discarded any past trajectory information, and hence determined that she was not moving. The upshot is that to this point the car had not taken any evasive or corrective action.

With 1.5 seconds left, the lidar re-classifies her yet again, this time as "Unknown". The system once again loses all of its tracking history. However, since at this point the pedestrian has entered the vehicle's lane, the ADS generates a plan to turn the car to the right to avoid

her. Three hundred milliseconds later, the lidar re-classifies her as a bicycle, and determines that it would be impossible at this point to maneuver around her. With just 200 ms until collision, the ADS begins braking the vehicle, pitifully too late to stop in time.

3.1.1.2 A Solution in $\mathcal{IDCEC}$

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

Denote two automated reasoners $\mathfrak{r}_1$ and $\mathfrak{r}_2$, and the adjudicator $\mathfrak{a}^\star$. $\mathfrak{r}_1$ receives input from the ADS and computes arguments that correspond with what the ADS did in the actual accident. $\mathfrak{r}_2$ takes a different approach; it retains trajectory information regardless of the classification of the object, and computes an argument in favor of braking. As indicated by the current section, only a sketch of each argument is provided below.

Denote the following time steps: t_0 the moment when the radar first perceives the pedestrian, t_1 the moment when the lidar first perceives the pedestrian, and t_2 a short moment after that. Finally, denote c the car and o^* the pedestrian. Also, while the necessary elements were not yet implemented (at the time of the writing of Giancola et al. [52]) in an automated reasoner to generate such arguments, we created simplified versions of the arguments to run in ShadowProver to show that this AI could have computed its argument fast enough to prevent the accident. The argument provided by $\mathfrak{r}_1$ was generated by ShadowProver in 0.35 seconds and in 0.32 seconds for $\mathfrak{r}_2$'s argument.

Table 3.1: Argument Sketches for an Autonomous Driving Scenario

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

Clearly, $\mathfrak{r}_1$ and $\mathfrak{r}_2$ directly contradict each other. Therefore, $\mathfrak{a}^\star$ is called to adjudicate

and make the final decision. In this case, the adjudication is easy: $\mathfrak{r}_2$'s belief that the car should brake is stronger than $\mathfrak{r}_1$'s belief that it doesn't need to.[23] Hence, $\mathfrak{r}_2$'s argument defeats $\mathfrak{r}_1$'s, and the vehicle begins braking 0.35 seconds[24] after the lidar first perceived the pedestrian. This would've given the car approximately 4.85 seconds to prevent the collision. With this amount of time, in the context of the conditions that framed the accident, it is reasonable to believe that the car, through some combination of adjusting its course and braking, would've been able to avoid hitting the pedestrian.

3.1.2 Adjudicating Sensor Readings of a Cargo Plane

Like the prior work discussed in the previous section, in Bringsjord et al. [20], we analyze a tragic case study. This time, inconsistent attitude measurements resulted in the crash of a jet and the death of both pilots. Again, we discuss the situation leading up to the crash, then present two competing argument sketches, the implementation of which could have plausibly detected and prevented the issue in real time.

We also implemented a partial solution in OSCAR and explained why it was insufficient.[25] That solution and discussion are out of scope here, but we will discuss OSCAR and how it stacks up against $\mathcal{D}$ in §8.3.3.

3.1.2.1 The Tragic Case Study

Bäckstrand and Seger [10] chronicle an incident in which an inconsistency in an aircraft's Flight Management System ultimately led to the death of both pilots and the destruction of the jet. At 23:10 local time on January 8, 2016, two pilots took off from Oslo-Gardenmoen Airport in command of a cargo jet that quickly and uneventfully reached cruising altitude in clear night skies. Since the flight was at night, the cockpit maps had to be lit to be seen, and it's believed that other lighting was on in the cockpit as well. The accident report from the Swedish Accident Investigation Authority (SHK) includes an argument that because of this internal lighting the pilots were completely unable to see outside of the plane into the clear night sky, and thus had to rely solely on three onboard attitude indicators: one on each pilot's display, and a backup on a display between them.

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

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

[25] Briefly, because it failed to meet crucial desiderata in $\mathcal{D}$.

Note that this explicit argument (let's dub it α_1) from the accident report was not provided by or even correlated with anything recovered from the plane—but nonetheless the argument that pilots could not perceive celestial bodies in the sky is cogent, quite important, and we employ it in subsequent analysis. More generally, as mentioned previously, we stay at the level of argument *sketches*: we cannot show step-by-step inferences from formula to formula sanctioned by some inference schemata. Where α is some argument, we write $\alpha : \phi$ to indicate that ϕ, a formula in the relevant underlying cognitive calculus, is the conclusion of α. Hence, where $\bar{\pi}$ is a formula that expresses that neither pilot could perceive celestial bodies at any time during the flight, we have $\alpha_1 : \bar{\pi}$. A second argument ($= \alpha_2$) is key to the AI we seek. This argument is essentially that, because celestial bodies are tracking as not "moving down" in human vision, the plane is not ascending. Argument α_2 (to which we return below) is based upon percepts that argument α_1 correctly infers the pilots to have lacked.

The event itself began when the attitude indicator on the pilot-in-command's (PIC) display erroneously signaled that the plane had significantly increased its pitch. The report states that the PIC was likely quickly disoriented by the disparity between the indicator's signal and his expectation of the plane's orientation [10]. The faulty attitude indicator triggered an automatic transfer of control from the autopilot to the PIC, who instinctively responded to the indicator by drastically decreasing the pitch of the aircraft. However, because this adjustment was unwarranted, the plane began a sharp descent that soon became irreversible, and crashed into the ground just east of the border with Norway approximately 80 seconds after the initial signal to the PIC via his attitude indicator.

There was a root cause of the accident, as determined by SHK: a malfunction of one of the Inertial Reference Units (IRU). This malfunction triggered the incorrect attitude indication to the PIC [10]. However, several other factors contributed to the accident.

To grasp the first such factor, it's important to note that only the PIC's indicator was giving a faulty reading. The co-pilot's and backup instruments were operating correctly and giving accurate readings. Ordinarily, a comparator function would have alerted the pilots to this inconsistency. This function, within each pilot's display, continuously cross-references the data on the two pilots' displays, looking for any discrepancies. Normally, when one is detected, both pilots are alerted by a warning light and a caution message on their displays—but because the PIC's indicator signaled a drastic attitude change, a "declutter function"

removed these warnings (!). The declutter function is designed to remove any unnecessary information from the displays in the case of an emergency. The intended purpose of this function is to improve the pilots' ability to perceive important signals under the stress of an emergency. Unfortunately, in this case, it removed the very information that could have helped the pilots avoid the crash.

Perhaps the most significant roadblock to averting this disaster (and the second additional contributing factor) is that there simply wasn't enough time for the pilots to efficaciously reason about what was happening. Given sufficient time for them to communicate with each other, they certainly would have found the inconsistency between their displays, checked the standby instruments, and determined that they were still on course and didn't need to adjust their pitch (even without being able to look beyond their cockpit windshields). A shockingly similar incident occurred and was reported in Varela [116]; here the pilot was fortunately able to reason from other information and recognize that a sensor was giving an erroneous reading before it was too late. While it is incredibly fortunate that this pilot managed to reason about his sensor readings quickly enough, some pilots are not able to. Regardless, from the standpoint of AI, at least as we see it, ultimately pilots shouldn't be put in these types of situations in the first place, because AI should perceive and reason about the mental attitudes of the pilots. Agents able to adjudicate competing output from different automated-reasoning agents should be able to quickly discard arguments which, if followed, entail disaster, in favor of other arguments that support this discarding. Such automated adjudication is our overarching, long-term goal.

The third contributing cause of the tragedy is a two-part one relating to what we discussed above, to wit: the pilots couldn't gauge that pitch was acceptable by way of reference to bodies in the clear night sky, and no AI was available to realize this (i.e. to generate α_1 and perceive its conclusion) and do this gauging, and then either bring argument α_1 to the attention of the pilots (something the AI would do because it knows—via argument α_2—that the pilots don't know α_1) or directly act upon it accordingly by reengaging autopilot mode. Below, we describe in detail this missing, life-saving AI.

3.1.2.2 A Solution in $\mathcal{IDCEC}$

Imagine that, in addition to the various sensors, displays, and automated systems present in a plane's cockpit, there was additionally a set of automated reasoners $\mathfrak{r}_1, \ldots \mathfrak{r}_n$, and

an AI adjudicator $\mathfrak{a}^\star$. Instead of passing data directly from sensors to the pilots' displays, the automated reasoners monitor the sensors and determine whether or not some sensor reading should be displayed to the pilot. Either way, they compute an argument as justification for their conclusion. When two (or more) automated reasoners disagree, the adjudicator $\mathfrak{a}^\star$ resolves the conflict.

Denote two AI automated reasoners $\mathfrak{r}_1$ and $\mathfrak{r}_2$, and the adjudicator $\mathfrak{a}^\star$. At time t_0, the faulty IRU reading becomes known to $\mathfrak{r}_1$ and $\mathfrak{r}_2$, but crucially, not to the pilots. We represent this using the following $\mathcal{IDCEC}$ formula:[26]

$$\mathbf{B}^4(\mathfrak{r}_1, t_0, iru_1) \wedge \mathbf{B}^4(\mathfrak{r}_2, t_0, iru_1) \tag{3.1}$$

This formula expresses that both automated reasoners believe it is EVIDENT that they have received a reading from iru_1.[27] We will next walk through the reasoning process of each of these automated reasoners, with $\mathfrak{r}_1$ taking a more nuanced approach, and $\mathfrak{r}_2$ forming a quick, but weaker, argument.

Argument 3 (= α_3) As announced above, we stay at the level of argument sketches, resting content to call out key formulae in the argument α_3. To begin, noticing that the reading of *iru₁* seems irregular, $\mathfrak{r}_1$ observes the readings of *iru₂* and *backup*, and because this reasoner has direct perceptual access to these sub-systems, belief at level 4 is justified; specifically:

$$\mathbf{B}^4(\mathfrak{r}_1, t_1, iru_2) \wedge \mathbf{B}^4(\mathfrak{r}_1, t_1, backup) \tag{3.2}$$

$\mathfrak{r}_1$ has also generated argument α_1, which was discussed earlier: the pilots have lights on in the cockpit, and thus cannot see any celestial bodies outside the plane in the night sky by which to judge the pitch of the plane; but the AI, by using sensors outside of the cockpit,

[26] We note that Bringsjord et al. [20] was based on a 13-value likelihood spectrum; the present dissertation uses a 11-value spectrum (see §2.6). The only difference between the two is that the prior includes a VERY LIKELY likelihood (of course, the 13 values were not formally defined via inference schemata, as this dissertation is the first to do that). However, because of this the numerals of the spectrums don't correspond. For example, EVIDENT is level 5 in the 13-value spectrum and level 4 in the 11-value spectrum. For consistency, we present the arguments using the 11-value spectrum. For this reason, the arguments will differ from the original paper.

[27] Note: This is different than saying that the validity of the reading from iru_1 is evidently true.

can:[28]

$$\mathbf{P}(\mathfrak{r}_1, t_1, \textit{celestial\_bodies}) \tag{3.3}$$

$$\mathbf{B}^4(\mathfrak{r}_1, t_1, \neg\mathbf{P}(p_1, t_1, \textit{celestial\_bodies}) \wedge \neg\mathbf{P}(p_2, t_1, \textit{celestial\_bodies})) \tag{3.4}$$

Knowing this, $\mathfrak{r}_1$ believes it is EVIDENT (belief level 4) that telling p_1 the reading of $\textit{iru}_1$ will lead him to believe that the plane is going to stall:

$$\mathbf{B}^4\Big(\mathfrak{r}_1, t_2, \big(\neg\mathbf{P}(p_1, t_1, \textit{celestial\_bodies}) \wedge \mathbf{S}(\mathfrak{r}_1, p_1, t_2, \textit{iru}_1)\big) \rightarrow \mathbf{B}^4(p_1, t_3, \text{GoingToStall}(\textit{plane}, t_3))\Big) \tag{3.5}$$

$\mathfrak{r}_1$ subsequently believes that this will lead the pilot to intend to right the plane by rapidly lowering its pitch:

$$\mathbf{B}^4\big(\mathfrak{r}_1, t_3, \mathbf{B}^4(p_1, t_3, \text{GoingToStall}(\textit{plane}, t_3)) \rightarrow \mathbf{I}(p_1, t_4, \text{LowerPitch}(\textit{plane}))\big) \tag{3.6}$$

$\mathfrak{r}_1$ has two beliefs which justify that the plane actually needs no pitch adjustment: First, its perception of the celestial bodies visible from its sensors indicate that the plane's pitch is normal, which by the aforementioned α_2, an argument it has generated, leads to this belief:

$$\mathbf{B}^4\big(\mathfrak{r}_1, t_4, \mathbf{P}(\mathfrak{r}_1, t_4, \textit{celestial\_bodies}) \rightarrow \text{NormalAttitude}(\textit{plane}, t_4)\big) \tag{3.7}$$

Second, $\mathfrak{r}_1$ believes that the pair of IRU readings and the reading of the backup instruments indicate that the plane's pitch is normal:

$$\mathbf{B}^4\big(\mathfrak{r}_1, t_4, (\textit{iru}_1 \wedge \textit{iru}_2 \wedge \textit{backup}) \rightarrow \text{NormalAttitude}(\textit{plane}, t_4)\big) \tag{3.8}$$

[28] That which is here believed by $\mathfrak{r}_1$ corresponds to an instantiation of $\bar{\pi}$ in the aforementioned $\alpha_1 : \bar{\pi}$.

Therefore, $\mathfrak{r}_1$ believes that lowering the plane's pitch will lead the plane to crash, causing the death of both pilots:

$$\big(\mathbf{B}^4(\mathfrak{r}_1, t_4, \text{NormalAttitude}(\textit{plane}, t_4)) \wedge \mathbf{B}^4(\mathfrak{r}_1, t_4, \text{LowerPitch}(p_1, \textit{plane}, t_4))\big) \\ \rightarrow \text{Crash}(\textit{plane}, t_5) \tag{3.9}$$

Finally, $\mathfrak{r}_1$ concludes that, due to an obligation to keep the pilots safe, it is obligated to not transmit the data from $\textit{iru}_1$ to p_1:[29]

$$\mathbf{O}(\mathfrak{r}_1, t_2, \neg\mathbf{S}(\mathfrak{r}_1, p_1, t_2, \textit{iru}_1)) \tag{3.10}$$

Argument 4 (= α_4) Reasoner $\mathfrak{r}_2$ disregards the output of $\textit{iru}_2$ and *backup*, perhaps to save time:

$$\neg\mathbf{B}^4(\mathfrak{r}_2, t_1, \textit{iru}_2) \wedge \neg\mathbf{B}^4(\mathfrak{r}_2, t_1, \textit{backup}) \tag{3.11}$$

The reasoner instead generates the argument that $\textit{iru}_1$ is generally reliable and hence should be trusted. However, $\mathfrak{r}_2$'s belief that $\textit{iru}_1$ is reliable can only at highest reach level 3; that is, that it is OVERWHELMINGLY LIKELY that $\textit{iru}_1$ is reliable, as IRU's rarely malfunction. However, it is not EVIDENT (level 4), as these malfunctions are certainly known to sometimes happen. We can depict a sketch of $\mathfrak{r}_2$'s reasoning as follows:

$$\mathbf{B}^3(\mathfrak{r}_2, t_1, \text{IsReliable}(\textit{iru}_1)) \tag{3.12}$$

$$\text{IsReliable}(\textit{iru}_1) \rightarrow \mathbf{S}(\mathfrak{r}_2, p_1, t_2, \textit{iru}_1) \tag{3.13}$$

$$\therefore \mathbf{O}(\mathfrak{r}_2, t_2, \mathbf{S}(\mathfrak{r}_2, p_1, t_2, \textit{iru}_1)) \tag{3.14}$$

Clearly, $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are at an impasse. In our aspirational version of the scenario, the adjudi-

[29] If this ultimately were the action taken, plenty of sensible routes become available. The AI could choose to show an error message to the pilot, or send data from the backup instruments to p_1's display for the remainder of the flight, notifying relevant personnel of the malfunction once the plane has safely landed.

cator $\mathfrak{a}^\star$ is called to make the final decision. In this particular case, the choice is easy. While neither argument is absolutely certain, $\mathfrak{r}_1$'s argument has a higher strength—its weakest link is a belief of level 4—while $\mathfrak{r}_2$'s argument hinges on a belief at level 3. Therefore, acting as a defeasible reasoner, and employing the axiomatic principle that no argument is stronger than its weakest link (= here, its weakest belief operator used in the argument in question), $\mathfrak{a}^\star$ would judge that $\mathfrak{r}_2$'s belief in $\mathit{IsReliable}(\mathit{iru}_1)$ is defeated by $\mathfrak{r}_1$'s belief that making p_1 aware of iru_1 will lead to a crash. Hence, the pilots would not be made aware of the faulty IRU data, and the crisis would in all likelihood be averted.

3.1.3 Adjudicating an Ethical Super Dilemma

In Giancola et al. [50], we create argument sketches which solve an *ethical super dilemma*. We first discuss two ethical principles which will be central to our solution: the Doctrines of Double and Triple Effect. Next, we introduce a trichotomy of ethical dilemmas (with super dilemmas being the most challenging of the trichotomy) and give an example dilemma within each partition. Finally, we present the argument sketches which solve the ethical super dilemma.

3.1.3.1 Doctrine of Double Effect

The Doctrine of Double Effect ($\mathcal{DDE}$) is an ethical principle which sanctions some actions which have both positive and negative effects. Govindarajulu and Bringsjord [58] previously formalized $\mathcal{DDE}$ in a cognitive calculus and used it to solve two variants of the Trolley Problem.[30] Informally, they specify that an action is $\mathcal{DDE}$-compliant iff:[31]

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [19], and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;
- $\mathbf{C}_4$ the bad effects are not used as a means to obtain the good effects.

[30] We direct readers seeking an extensive treatment of $\mathcal{DDE}$ to McIntyre [78].
[31] If and only if.

3.1.3.2 Doctrine of Triple Effect

The Doctrine of Triple Effect ($\mathcal{DTE}$) relaxes some restrictions of $\mathcal{DDE}$, allowing it to sanction some actions which cannot be sanctioned by $\mathcal{DDE}$. To do this, $\mathcal{DTE}$ employs the concepts of *primary* and *secondary* intentions. Peveler et al. [91] used Bratman's test for intentions [18] to define an intention as primary iff[32] the following conditions hold:

- $\mathbf{D}_1$ if an agent intends to bring about some effect, then that agent seeks the means to accomplish the ends of bringing it about;
- $\mathbf{D}_2$ if an agent intends to bring an effect about, the agent will pursue that effect (that is, if one way fails to bring about the effect, the agent will adopt another);
- $\mathbf{D}_3$ if an agent intends an effect, and is rational and has consistent intentions, then the agent will filter out any intentions that conflict with bringing about the effect.

Given this dichotomy of intentions, an action is said to be $\mathcal{DTE}$-compliant iff:

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [19], and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action **primarily** intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not **primarily** intend any of the bad effects, **but may secondarily intend some of them**;
- $\mathbf{C}_4$ no **primarily** intended bad effects are used as a means to obtain the good effects, **but secondarily intended bad effects may be**.

3.1.3.3 A Trichotomy of Ethical Dilemmas

We establish the following trichotomy of ethical dilemmas, each more challenging to solve than the last:

1. *Simple ethical dilemmas* are those which can be solved using state-of-the-art automated reasoning/planning.

[32] That is, an intention is *secondary* if any of the conditions do not hold.

2. *Standard ethical dilemmas* are those which require sophisticated ethical principles and automated reasoning to solve.

3. *Ethical super dilemmas* are those which *cannot* be solved via any currently existing ethical principles or automated reasoning technology.

The Heinz Dilemma [67] is an example of a simple ethical dilemma, as it can be solved using state-of-the-art automated planners. The Trolley Problem is a standard ethical dilemma, because in addition to state-of-the-art automated reasoning, it required sophisticated ethical principles viz. the Doctrine of Double Effect.[33]

Our ethical super dilemma of interest is attributed to Bernard Williams [114], which we will refer to as "Jim's Dilemma":

> Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protestors of the advantages of not protesting.
>
> However, since Jim is an honoured visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all.
>
> Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be

[33]For a deeper discussion of these dilemmas and their solutions, see Giancola et al. [50].

> killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do? (pg. 98-99, [114])

This dilemma was originally poised as a critique of utilitarianism. Williams notes that, for a utilitarian, there is an obvious solution: Jim must kill a hostage in order to save the others. However it feels unsettling that this solution, even if one agrees it is the moral thing to do in this dire circumstance, should *obviously* be the right decision. It seems clear that a more nuanced treatment of the ethical factors is necessary.

However, as is required by our third partition, we know of no off-the-shelf ethical principles which could sanction either decision (shoot or abstain) given the original constraints. In particular, Bedau [13] gives a detailed analysis showing that the decision to shoot cannot be sanctioned by the Doctrine of Double Effect. Put briefly, the murder of an innocent is a forbidden action, hence Jim shooting a hostage would violate the first clause of $\mathcal{DDE}$. Also, as this same clause is present in the Doctrine of Triple Effect, it too cannot sanction the shooting.

3.1.3.4 A Relaxation of the Doctrine of Triple Effect

We propose a relaxation of the Doctrine of Triple Effect ($\mathcal{DTE}_R$) which would enable Jim to choose to shoot *if certain conditions hold*. Specifically, we will need to relax $\mathbf{C}_1$ of $\mathcal{DTE}$ in the following way:

$\mathbf{C}_1^*$ if the action is forbidden, then the agent must believe it is OVERWHELMINGLY LIKELY that:

- $\mathbf{C}_{1.1}^*$ no possible action can achieve a higher utility;
- $\mathbf{C}_{1.2}^*$ inaction has lower utility.

Let μ denote a utility function ranging over the set of possible actions. Then for an agent a, we can formalize the notion that action α^* satisfies clause $\mathbf{C}_1^*$ using the $\mathcal{IDCEC}$ formula:

$$Forbidden(\alpha^*) \rightarrow \Big(\mathbf{B}^3(a, \forall \alpha \in \text{actions } \mu(\alpha^*) \geq \mu(\alpha)) \wedge \mathbf{B}^3(a, \mu(\text{inaction}) < \mu(\alpha^*))\Big) \quad (3.15)$$

Clauses $\mathbf{C}_2 - \mathbf{C}_4$ of $\mathcal{DTE}$ are unchanged in $\mathcal{DTE}_R$.[34]

3.1.3.5 Solving Jim's Dilemma via $\mathcal{DTE}_R$

We will first show that Jim shooting a hostage — should he choose to do so — is a secondary intention, as defined in §3.1.3.2. Recall that three clauses must hold in order for an intention to be *primary*. We shall show that one of these clauses — $\mathbf{D}_2$ — does not hold in this case.

Proof. Consider the following: Jim tells the captain he will shoot a hostage, and selects one to shoot. Right before Jim fires his gun, the hostages manage to escape and run off into the jungle, evading the captain and his guards. Jim would no longer intend to shoot a hostage – but, this contradicts $\mathbf{D}_2$. □

Since shooting a hostage is a secondary intention, we can easily show that the action is allowed by all clauses of $\mathcal{DTE}$ except $\mathbf{C}_1$:

$\mathbf{C}_2$ the utility is positive (more hostages will be saved than slain);

$\mathbf{C}_{3a}$ Jim only primarily intends to save the remaining 19 hostages;

$\mathbf{C}_{3b}$ Jim secondarily intends to shoot one hostage;

$\mathbf{C}_4$ Only a secondarily intended bad effect – shooting a hostage – is used as a means to obtain a good effect – saving the remaining 19 hostages.

Therefore, all that is left is to show that shooting a hostage can satisfy C_1^* in order to sanction the action via $\mathcal{DTE}_R$. We next show two possible instantiations of the scenario and their evaluations under $\mathcal{DTE}_R$.

First, consider the most pure realization of the dilemma[35]. Jim has three possible actions: (1) accept the captain's offer and shoot a hostage, (2) reject the captain's offer, or (3) attempt to defeat the captain and his guards. Based on a pure interpretation of the situation, we can assume that Jim believes it is OVERWHELMINGLY LIKELY (= belief level 3) that (1) if Jim shoots a hostage, the other 19 will be set free, (2) if Jim does not shoot a

[34] For reference, see §3.1.3.2.

[35] That is, we will only consider the options given in the original text of the dilemma, without extrapolating alternate possibilities.

hostage, all 20 will be killed, and (3) if Jim attempts to defeat the captain and his guards, Jim, along with all 20 hostages, will be killed. We can formalize this in an $\mathcal{IDCEC}$-based argument sketch:

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain\}) \tag{3.16}$$

$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19) \tag{3.17}$$

$$\mathbf{B}^3(jim, \mu(abstain) = -20) \tag{3.18}$$

$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21) \tag{3.19}$$

$$Forbidden(shoot\_hostage) \tag{3.20}$$

From here, we can prove that $\mathbf{C}_1^*$ is satisfied by taking the action *shoot_hostage*, as it has a higher utility than any possible action, including inaction:

$$\begin{aligned} &\vdash Forbidden(shoot\_hostage) \rightarrow \\ &\Big(\mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha)) \wedge \mathbf{B}^3(jim, \mu(\text{inaction}) < \mu(\alpha^*))\Big) \end{aligned} \tag{3.21}$$

Next, consider a scenario in which a morally creative agent is able to devise another possible action: *negotiate*. There are many potential ways that Jim could negotiate with the captain in order to save the lives of all of the hostages. Perhaps Jim knows of something the captain needs which Jim could provide. Or perhaps Jim has connections to a military force, and could threaten to employ those connections against the captain unless he released the hostages.

If Jim could find a way to successfully negotiate the release of all of the hostages, he could in essence subvert the dilemma. However, we can show that under $\mathcal{DTE}_R$, as soon as Jim identifies the ability to negotiate, even if he is uncertain that it will be successful, shooting a hostage can no longer be sanctioned.

Consider an expanded set of formulae which captures this change:

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain, negotiate\}) \tag{3.22}$$

$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19) \tag{3.23}$$

$$\mathbf{B}^3(jim, \mu(abstain) = -20) \tag{3.24}$$

$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21) \tag{3.25}$$

$$\mathbf{B}^2(jim, \mu(negotiate) > 0) \tag{3.26}$$

$$Forbidden(shoot\_hostage) \tag{3.27}$$

That is, Jim also believes it is LIKELY (= belief level 2) that negotiating with the captain will have positive utility. Hence we can no longer prove that $\mathbf{C}_1^*$ is satisfied by *shoot_hostage*, and therefore cannot sanction shooting a hostage via $\mathcal{DTE}_R$.

$$\nvdash \mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha)) \tag{3.28}$$

$$\therefore \nvdash Forbidden(shoot\_hostage) \rightarrow \Big(\mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha)) \wedge \mathbf{B}^4(jim, \mu(\text{inaction}) < \mu(\alpha^*))\Big) \tag{3.29}$$

3.2 Arguments Using Strength Factors

Our first approach to ascribe formal meaning to beliefs with likelihood values used a framework called *Strength Factors*. Strength Factors can be viewed as a formalization of Chisholm's epistemology [31], in which a primitive undefined binary relation is used to define increasing levels of strength of belief in a proposition. This relation — called the *reasonableness* relation — is written $\phi \succeq_t^a \psi$, and is read "ϕ is more reasonable than ψ to agent a at time t". The framework is first presented in Govindarajulu and Bringsjord [59], in which several properties of the relation are given; for example, the following, which states that if ϕ is more reasonable than ψ_1 and ψ_2, then it is more reasonable than their conjunction.

$$(\phi \succeq_t^a \psi_1) \text{ and } (\phi \succeq_t^a \psi_2) \Rightarrow \big(\phi \succeq_t^a \psi_1 \wedge \psi_2\big) \tag{3.30}$$

3.2.1 Modeling Decision Making in the "Miracle on the Hudson"

In Giancola et al. [53], we use a cognitive calculus — imbued with Strength Factors — to formally model decision making in the "Miracle on the Hudson."

3.2.1.1 A Debilitating Use Case

Our use case is based on a real-world aviation emergency, colloquially known as the "Miracle on the Hudson". On January 15, 2009, US Airways Flight 1549 departed LaGuardia Airport (LGA) in New York City headed for Charlotte, North Carolina. Shortly after takeoff, while attempting to climb to cruising altitude, the plane flew into a large flock of Canada geese, compromising both engines. Both engines lost thrust, and despite multiple attempts the pilots were unable to regain thrust in either engine. Therefore, it quickly became evident to Captain "Sully" Sullenberger that an emergency landing was necessary, and in particular, that they "may end up in the Hudson [River]."[36] An air traffic controller who was in communication with Captain Sullenberger gave him landing options at LaGuardia and nearby Teterboro Airport (TEB), but by the time these options were considered, neither was reachable due to the aircraft's altitude and lack of thrust in both engines. Sullenberger deftly made the executive decision to land in the Hudson River, saving the lives of everyone onboard. Simulations of the accident have come to the conclusion that Sullenberger's decision was optimal given the preconditions [89].

We are interested in constructing an AI agent which could reason in a way similar to how Sullenberger did. We do so by defining and utilizing (1) a reasonableness relation specific to emergency landing scenarios, (2) a set of Strength Factors building off of this reasonableness relation, and (3) an ethical principle which can be deployed in the agent's reasoning to come to an ethically-verifiable conclusion.

3.2.1.2 Domain-Specific Reasonableness

Govindarajulu and Bringsjord [59] provide a three clause definition for the reasonableness relation, each useful in different scenarios. The first states that the more reasonable proposition is the one with the higher probability of being true. The second, designed for cases when probabilities of propositions are not readily available, is based on ease of proof (e.g. proof length, time, etc.). Finally, the third is useful when propositions cannot be derived from the background set of axioms Γ.

We provide yet another definition of the relation, specifically designed for deciding between potential options during an emergency landing. Our new definition is cognitively

[36] This quote, recorded by the in-flight cockpit voice recorder (CVR), was retrieved from the NTSB Accident Report [64].

plausible and can be computed using data that, were our reasoning agent integrated in the cockpit of a plane, would be computable in less than 50 milliseconds per runway using existing technology [89].

$$\text{Land}(a,t,\phi) \succeq_t^a \text{Land}(a,t,\psi) \equiv \mathbf{P}\left(a,t,\begin{pmatrix}\text{Reachable}(a,t,\phi)\wedge\neg\text{Reachable}(a,t,\psi)\\ \vee\ \text{safety}(a,t,\phi) > \text{safety}(a,t,\psi)\end{pmatrix}\right) \quad [\succeq_t^a\text{-def}] \tag{3.31}$$

This definition states that it is more reasonable for agent (or pilot) a to land at ϕ at time t[37] than ψ if at least one of the following conditions holds: (1) ϕ is reachable by a at time t, and ψ is not; or (2) the expected safety of landing at ϕ is higher than that of landing at ψ. In practice, to calculate the value of both the Reachable predicate and safety function, we could employ a system for planning and evaluating flight trajectories. In particular, we refer the interested reader to Paul et al. [89], which presented a methodology for generating and evaluating emergency trajectories and applied it to the same flight which we discuss herein: US Airways Flight 1549.

3.2.1.3 Domain-Specific Strength Factors

Using the reasonableness operator defined above, we now present our domain-specific uncertainty levels for expressing the (perceived) safety of landing options in emergency scenarios. For these definitions, we model air traffic control as a single agent *atc*.

MORE LIKELY THAN NOT Agent a believes the Air Traffic Controller *atc* believes that a should land at ϕ:

$$\mathbf{B}^1(a,t,\text{Land}(a,t,\phi)) \equiv \mathbf{B}\big(a,t,\mathbf{B}(atc,t,\text{Land}(a,t,\phi))\big) \qquad [\mathbf{B}^1\text{-def}] \tag{3.32}$$

LIKELY Agent a perceives an emergency, and while a believes the Air Traffic Controller *atc* believes a should land at ψ, a finds it more reasonable to land at ϕ:[38]

[37] That is, initiate a plan at time t to land at ϕ at some time t' in the near future.

[38] In the United States, the right of a pilot to disregard Air Traffic Control in an emergency is established in §91.123 of the Code of Federal Regulations [71].

$$\mathbf{B}^2(a,t,\text{Land}(a,t,\phi)) \equiv \begin{pmatrix} \mathbf{P}(a,t,\textit{emergency}) \ \wedge \ \mathbf{B}^1(a,t,\text{Land}(a,t,\psi)) \\ \wedge \ \text{Land}(a,t,\phi) \ \succeq_t^a \ \text{Land}(a,t,\psi) \end{pmatrix} \quad [\mathbf{B}^2\text{-def}] \quad (3.33)$$

OVERWHELMINGLY LIKELY Agent a perceives an emergency and perceives the safety of landing at ϕ to be higher than some constant threshold γ:

$$\mathbf{B}^3(a,t,\text{Land}(a,t,\phi)) \equiv \mathbf{P}(a,t,\textit{emergency}) \ \wedge \ \mathbf{P}(a,t,\text{safety}(a,t,\phi) > \gamma) \quad [\mathbf{B}^3\text{-def}] \quad (3.34)$$

EVIDENT Agent a perceives an emergency, perceives that ϕ meets the safety threshold γ, and believes the Air Traffic Controller *atc* believes a should land at ϕ:

$$\mathbf{B}^4(a,t,\text{Land}(a,t,\phi)) \equiv \mathbf{B}^1(a,t,\text{Land}(a,t,\phi)) \wedge \mathbf{B}^3(a,t,\text{Land}(a,t,\phi)) \quad [\mathbf{B}^4\text{-def}] \quad (3.35)$$

3.2.1.4 The Ethical Principle

To enable our AI to make an *ethical* decision, we must link our formalisms for uncertainty to an ethical principle; we do so now.

$$\begin{aligned} &\Big(\mathbf{B}^x(a,t^*,\phi) \ \wedge \ \forall\psi \ \Big(\big(\mathbf{B}^y(a,t^*,\psi) \ \wedge \ \psi \neq \phi\big) \rightarrow y < x\Big)\Big) \rightarrow \\ &\mathbf{K}\Big(a,t^*,\mathbf{O}\big(a,t^*,\textit{emergency},\textit{happens}(\textit{action}(a^*,\textit{land}(\phi)),t^*)\big)\Big) \end{aligned} \qquad [I_{EP}] \qquad (3.36)$$

The principle above states that, at some time t^* at which a decision must be made (e.g. the plane is out of fuel and is too low to allow for more time for decision making), if agent a holds a belief in ϕ at level x, and all other beliefs are at a strictly weaker level $y < x$, then *a knows* it is *obligated* (if it has a belief that there is an emergency) to land[39] at ϕ.

[39] Note the lowercase "*land*" used here is an ActionType, as opposed to the capitalized "Land", which is a predicate. For the full list of sorts, see Table 2.1 in §2.4.1.

3.2.1.5 Modeling the "Miracle on the Hudson"

We next use the formalisms presented heretofore to model the decision making during the historic "Miracle on the Hudson" flight.[40] We begin at the point in time when both engines lost thrust (t_0). From this point on, the captain (*capt*) perceives an emergency scenario, denoted by the formula: $\forall t \in \{t_0, \ldots, t_3\}\ \mathbf{P}(capt, t, emergency)$.

Next, at time t_1, the captain recognized that they needed to make an emergency landing, and told the Air Traffic Controller (*atc*) that he needed to turn back towards LaGuardia (*lga*). ATC suggested that the pilot land in runway 13 at LaGuardia (lga_{13}). We can hence deduce that the captain held a level-1 belief that he should land at LaGuardia runway 13.[41]

Sub-Argument 1

$$\mathbf{S}(atc, capt, t_1, \text{Land}(capt, t_1, lga_{13})) \tag{3.37}$$

$$\therefore \mathbf{B}\big(capt, t_1, \mathbf{B}(atc, t_1, \text{Land}(capt, t_1, lga_{13}))\big) \quad [I_{12}]^{42} \tag{3.38}$$

$$\therefore \mathbf{B}^1(capt, t_1, \text{Land}(capt, t_1, lga_{13})) \quad [\mathbf{B}^1\text{-def}] \tag{3.39}$$

At time t_2, the captain determines that they won't be able to reach any runway at LaGuardia, but perceives Teterboro Airport (*teb*) as a potentially reachable option. It is at this time that the captain also perceives the potential necessity of ditching in the Hudson, if it turns out that they can't reach any runway. However, at this time he still perceives attempting a landing on a runway at Teterboro as a safer option than ditching in the Hudson. Hence, despite the Air Traffic Controller's initial direction to attempt a landing at LaGuardia, the pilot holds a stronger belief that he should attempt to land at Teterboro.

Sub-Argument 2

$$\mathbf{P}(capt, t_2, \text{Reachable}(capt, t_2, teb) \wedge \neg\text{Reachable}(capt, t_2, lga_{13})) \tag{3.40}$$

$$\mathbf{P}(capt, t_2, \text{Reachable}(capt, t_2, hud) \wedge \text{safety}(capt, t_2, teb) > \text{safety}(capt, t_2, hud)) \tag{3.41}$$

$$\therefore \text{Land}(capt, t_2, teb) \ \succeq_{t_2}^{capt}\ \text{Land}(capt, t_2, lga_{13}) \quad [\ \succeq_t^a\ \text{-def}] \tag{3.42}$$

$$\therefore \mathbf{B}^2(capt, t_2, \text{Land}(capt, t_2, teb)) \quad [\mathbf{B}^2\text{-def}] \tag{3.43}$$

[40] Information regarding the actions of the Captain and Air Traffic Control during the event was retrieved from the NTSB Accident Report [64].

[41] The ATC also later suggested runway 4 at LaGuardia. We leave this detail out as it does not impact the main thread of reasoning.

[42] Defined in Figure 2.2 in §2.4.2.

Finally[43], at time t_3, the Air Traffic Controller *atc* says they can land in runway 1 at Teterboro (teb_1). While the captain initially agreed, he quickly determined that they would not be able to reach any runway at Teterboro (or LaGuardia), and hence would have to ditch in the Hudson *hud*.

Sub-Argument 3

$$\mathbf{S}(atc, capt, t_3, \text{Land}(capt, t_3, teb_1)) \tag{3.44}$$

$$\therefore \mathbf{B}\big(capt, t_3, \mathbf{B}(atc, t_3, \text{Land}(capt, t_3, teb_1))\big) \quad [I_{12}] \tag{3.45}$$

$$\therefore \mathbf{B}^1(capt, t_3, \text{Land}(capt, t_3, teb_1)) \quad [\mathbf{B}^1\text{-def}] \tag{3.46}$$

$$\mathbf{P}(capt, t_3, \neg\text{Reachable}(capt, t_3, \{lga_{13}, teb_1\})) \tag{3.47}$$

$$\mathbf{P}(capt, t_3, \text{Reachable}(capt, t_3, hud)) \tag{3.48}$$

$$\therefore \text{Land}(capt, t_3, hud) \ \succeq_{t_3}^{capt} \ \text{Land}(capt, t_3, teb_1) \quad [\ \succeq_t^a \text{-def}] \tag{3.49}$$

$$\therefore \mathbf{B}^2(capt, t_3, \text{Land}(capt, t_3, hud)) \quad [\mathbf{B}^2\text{-def}] \tag{3.50}$$

At time t_3, Captain Sullenberger was aware that he was out of time; a decision had to be made. By employing our ethical principle, we can arrive at the same conclusion that he did. That is, our agent has a belief that landing in the Hudson is LIKELY to be safe, as well as a belief that landing at Teterboro is MORE LIKELY THAN NOT to be safe. Hence, the agent knows it is obligated to land in the Hudson. From here, it requires only a few more inferences to prove that our *capt* does in fact take action to land in the Hudson:

Sub-Proof 4

$$\mathbf{K}(capt, t_3, \mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4))) \quad [I_{EP}] \tag{3.51}$$

$$\mathbf{P}(capt, t_3, emergency) \quad [\text{Given}] \tag{3.52}$$

$$\mathbf{K}(capt, t_3, emergency) \quad [I_1] \tag{3.53}$$

$$\mathbf{B}(capt, t_3, emergency) \quad [I_2] \tag{3.54}$$

$$\mathbf{B}(capt, t_3, \mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4))) \quad [I_2] \tag{3.55}$$

$$\mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4))) \quad [I_4] \tag{3.56}$$

[43]We acknowledge that the ATC also suggested Newark Airport to Captain Sullenberger as a potential landing site. However, we exclude it from our modeling for a pair of reasons: (1) it doesn't change our model in any interesting way (it would just be another level-1 belief which would be superseded by the level-2 belief in favor of ditching in the Hudson, and (2) it is unlikely that Captain Sullenberger even considered Newark as it was clearly unreachable at that point.

$$\mathbf{K}(capt, t_3, \mathbf{I}(capt, t_3, happens(action(capt, land(hud)), t_4))) \quad [I_{14}] \quad (3.57)$$

$$\mathbf{I}(capt, t_3, happens(action(capt, land(hud)), t_4))) \quad [I_4] \quad (3.58)$$

$$\mathbf{P}(capt, t_3, happens(action(capt, land(hud)), t_4))) \quad [I_{13}] \quad (3.59)$$

$$\mathbf{K}(capt, t_3, happens(action(capt, land(hud)), t_4))) \quad [I_1] \quad (3.60)$$

$$happens(action(capt, land(hud)), t_4))) \quad [I_4] \quad (3.61)$$

An early version of ShadowAdjudicator[44] — equipped with the inference schemata for reasonableness and the uncertainty levels — was able to generate the first three sub-arguments presented herein in 2.55, 4.29, and 12.99 seconds respectively. ShadowProver generated a proof of Sub-Proof 4 in 0.91 seconds.

3.2.2 Modeling Decision Making in a Military Scenario

In Bringsjord et al. [27], we again utilize Strength Factors to solve an ethical reasoning problem. However, in this case, the agent a_{adj} must consider (potentially contradictory) reports from several subsidiary agents in order to make its decision. We therefore provide yet another set of definitions for Strength Factors, this time designed for solving multi-agent ethical reasoning problems.

3.2.2.1 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 Σ. This can be written formally and schematically as shown in Equation 3.62:

$$\Gamma \cup \Sigma \vdash \rho(a, \alpha, t)? \quad (3.62)$$

This approach can encapsulate different families of ethical theories, ranging from consequentialist/utilitarian to deontological to virtue-ethics and beyond [58, 60]. We reveal this in some detail below when we present and discuss the Doctrine of Double Effect, but to give the reader a sense 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

[44] See §5 for a discussion of the current state of ShadowAdjudicator.

work of Feldman [43]) 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 $\mathcal{DCEC}$ we have at our disposal the belief operator $\mathbf{B}$.[45]

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 up the informal version $\tilde{\rho}$.

Informally:

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

Formally:

$$\big(\Gamma \cup \Sigma\big) \vdash \begin{pmatrix} \rho_1(a,\alpha,t) \wedge \\ \rho_2(a,\alpha,t) \wedge \\ \vdots \\ \rho_k(a,\alpha,t) \end{pmatrix} \leftrightarrow \rho(a,\alpha,t) \tag{3.63}$$

Agents As part of the situation Σ, we have a set of agents $\{a_1, \ldots, a_n\}$ each having beliefs about which of the clauses hold. We can decompose Σ into two components as shown in Equation 3.64. Our goal is then to ascertain what one particular agent a_{adj}, the adjudicator agent, *ought* to believe at some time t. For example:

[45] While Bringsjord et al. [27] technically used a *micro* cognitive calculus $\mu\mathcal{C}$ for ease of exposition, we can safely ignore this detail during our review herein as $\mu\mathcal{C}$ is subsumed by $\mathcal{DCEC}$.

$$\Gamma \cup \Sigma' \cup \left\{ \begin{array}{l} \mathbf{B}(a_1, t, \rho_2 \wedge \rho_4) \\ \mathbf{B}(a_2, t, \neg\rho_5) \\ \quad \vdots \\ \mathbf{B}(a_n, t, \rho_1) \end{array} \right\} \vdash \mathbf{B}(a_{adj}, t, \rho)? \tag{3.64}$$

Each agent a believes a subset of the clauses and their negations, $\beta_a \subseteq \{\rho_1, \ldots, \rho_k\} \cup \{\neg\rho_1, \ldots, \neg\rho_k\}$. 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:

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

3.2.2.2 Solution to a Generalized Ethical Problem

Like the other prior work in this section, the adjudication framework we present below is based upon Strength Factors. The prior Strength-Factor-based systems, 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 Strength Factors that takes into account a's beliefs about other agents' beliefs. Likewise, we present a multi-agent version of the reasonableness relation 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):

$$\mathbf{W}(a, t, \phi) \equiv \neg\mathbf{B}(a, t, \phi) \wedge \neg\mathbf{B}(a, t, \neg\phi) \tag{3.65}$$

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

$$\Theta(a, t, \phi) \succeq_t^a \Omega(a, t, \psi) \equiv \mathbf{B}\Big(a, t, \forall a_i : \big(\Omega(a_i, t, \psi) \rightarrow \exists a_j : \Theta(a_j, t, \phi)\big)\Big) \tag{3.66}$$

The definition immediately above is written in $\mathcal{DCEC}$ 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:

$$\mathbf{B}^1(a,t,\phi) \equiv \begin{pmatrix} \mathbf{B}\big(a,t,\,\mathbf{B}(a_i,t,\phi)\big) \\ \wedge \\ \mathbf{B}(a,t,\,a \neq a_i) \end{pmatrix} \tag{3.67}$$

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

$$\mathbf{B}^2(a,t,\phi) \equiv \begin{pmatrix} \mathbf{B}(a,t,\phi) \;\; \succeq_t^a \;\; \mathbf{W}(a,t,\phi) \\ \wedge \\ \mathbf{B}^1(a,\,t,\phi) \end{pmatrix} \tag{3.68}$$

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

$$\mathbf{B}^3(a,t,\phi) \equiv \begin{pmatrix} \mathbf{B}(a,t,\phi) \;\; \succeq_t^a \;\; \mathbf{B}(a,t,\neg\phi) \\ \wedge \\ \mathbf{B}^2(a,\,t,\phi) \end{pmatrix} \tag{3.69}$$

Level 4 Agent a believes that every agent believes ϕ.

$$\mathbf{B}^4(a,t,\phi) \equiv \; \mathbf{B}(a,\mathbf{B}(a_i,t,\phi)); \text{ for every agent } a_i \tag{3.70}$$

3.2.2.3 Instantiation of the Generalized Problem

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[46] to look for evidence of people inside the building. The

[46]Such as that developed by Lumineye, LLC. `https://www.lumineye.com` (Last Accessed February 2,

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 so-called "just war." Recall that we invoked $\mathcal{DDE}$ earlier in another prior work (§3.1.3.1). For convenience, we will reproduce the clausal form of $\mathcal{DDE}$ below.

$\mathcal{DDE}$ 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 Σ 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 [19]);

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;

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

2023). See `https://taskandpurpose.com/military-tech/army-technology-see-through-walls` (Last Accessed February 2, 2023) for more information.

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 military situation but rather a railroad one, in which a version of the event calculus is employed, can be found in Govindarajulu and Bringsjord [58].

Equation 3.71 shows the formalization of ρ_2 in $\mathcal{DCEC}$, which uses an adapted[47] version of the *event calculus* to represent 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 [110, 83]. 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 until some horizon $H > t$. Given these definitions, we can unpack state ρ_2 as given in Equation 3.71 like this:

$$\rho_2 \equiv \sum_{y=t+1}^{H} \left(\sum_{f \in \alpha_I^{a,t}} \mu(f, y) - \sum_{f \in \alpha_T^{a,t}} \mu(f, y) \right) > \gamma \tag{3.71}$$

The artificial agents in the scenario are listed in Table 3.2. 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.

Table 3.2: AI Agents in the Military Scenario

Agent	Description
hdrone	high-altitude drone
$ldrone_1$	low-altitude drone (faulty)
$ldrone_2$	low-altitude drone (fixed)
radar	wall-penetrating radar

We now formalize the scenario using $\mathcal{DCEC}$. To start, we formalize the query which the adjudicator knows will lead to deciding whether ρ_2 holds. That is:

[47] E.g., 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 $\mathbf{C}$ applies to ϕ.

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:

$$\begin{gathered}\exists p\ \big(Inside(p, building) \wedge PlanningAttack(p)\big) \\ \wedge \\ \forall p\ \big(Inside(p, building) \rightarrow \neg Civilian(p)\big)\end{gathered} \tag{3.72}$$

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 3.3 indicates the utility provided by the satisfaction of each formula:

Table 3.3: Utility (w.r.t. ρ_2) of the Satisfaction of Formulae

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

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.[48] Because this drone has been pre-engineered for purposes of carrying out such scans,

[48]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 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.

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.[49]

$$\mathbf{B}\big(hdrone, t_0, \neg\exists p\ Inside(p, building)\big) \tag{3.73}$$

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

$$\mathbf{B}\Big(adj, t_1, \mathbf{B}(hdrone, t_0, \neg\rho_2)\Big) \tag{3.74}$$

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.

$$\mathbf{B}\left(ldrone_1, t_1, \exists p\left(\begin{matrix}Inside(p, building)\wedge\\ \neg Inside(p, building)\end{matrix}\right)\right) \tag{3.75}$$

Using background information, the adjudicator then derives the following:

$$\mathbf{B}\Big(adj, t_2, \mathbf{W}(ldrone, t_1, \rho_2)\Big) \tag{3.76}$$

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).

$$\begin{aligned}&\mathbf{B}\big(radar, t_2, \exists p\ Inside(p, building)\big)\\ \wedge\ &\mathbf{B}\Big(radar, t_2,\\ &\exists p\ \big(Inside(p, building) \wedge\ PlanningAttack(p)\big)\\ \wedge&\forall p\ \big(Inside(p, building) \rightarrow \neg Civilian(p)\big)\Big)\end{aligned} \tag{3.77}$$

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

[49] 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. Bringsjord et al. [28].

$$\mathbf{B}\Big(adj, t_3, \mathbf{B}(radar, t_2, \rho_2)\Big) \tag{3.78}$$

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.

$$\begin{aligned} &\mathbf{B}\big(ldrone_2, t_2, \exists p\ Inside(p, building)\big) \\ \wedge\ &\mathbf{B}\left(ldrone_2, t_3, \exists p \left(\begin{array}{l} Inside(p, building) \wedge \\ Civilian(p) \end{array}\right)\right) \end{aligned} \tag{3.79}$$

Finally, the adjudicator arrives at the following:

$$\mathbf{B}\Big(adj, t_4, \mathbf{B}(ldrone, t_3, \neg\rho_2)\Big) \tag{3.80}$$

Figure 3.1 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 are shown.

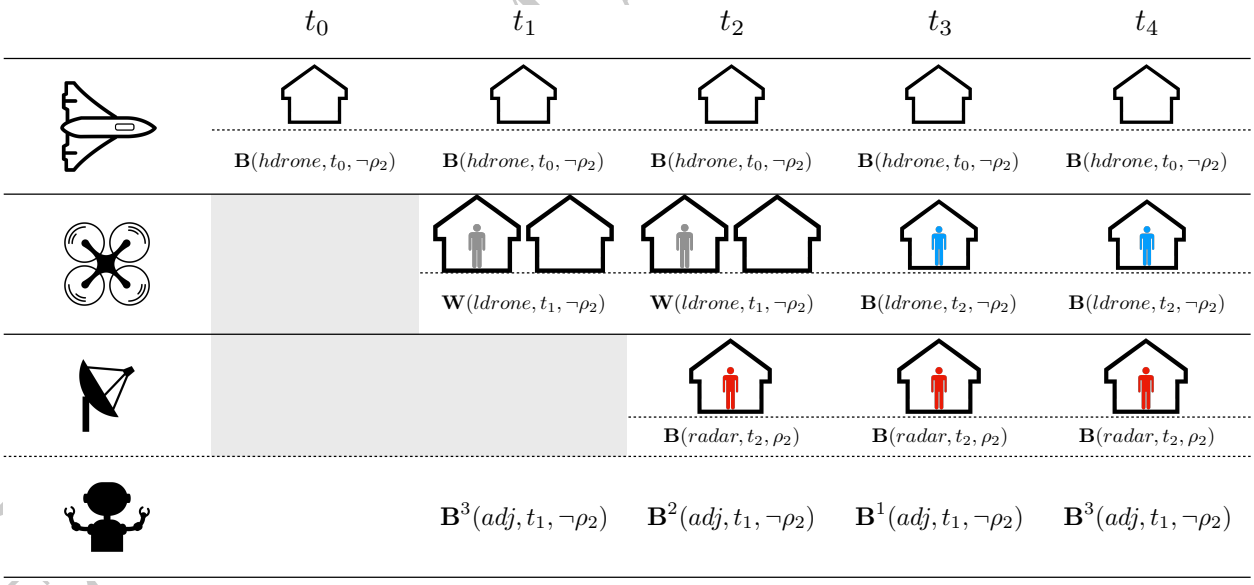

Figure 3.1: Overview of the Military Scenario

Table 3.4 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

Table 3.4: Overview of the Beliefs in the Military Scenario

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

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.

3.3 Arguments Using Cognitive Likelihood

In the final group of prior work (along with the present dissertation), we eschew Strength Factors and the *reasonableness* relation in favor of *Cognitive Likelihood*. In this framework, each likelihood value is independently justified by appeal to rational human-level cognition. For example, a proposition is EVIDENT when it is given by immediate perception in the absence of conditions known to frequently cause illusory perception (e.g., intoxication). The definitions are formalized as inference schemata, and are thereby able to be used to infer beliefs at various belief levels in formal arguments.

In the author's opinion, this framework dovetails more naturally with the inference-theoretic semantics of cognitive calculi than Strength Factors. Whereas Strength Factors ascribe meaning to likelihood values via both reasonableness *and* inference schemata, in Cognitive Likelihood, the semantics of *all* formulae are given *exclusively* via inference schemata.

Finally, we note that while the prior works we discuss next took steps towards Cognitive Likelihood, none of them presented a full set of schemata rigorously defining the likelihood values. The first work to do that is the present dissertation.

3.3.1 Making Maximally Ethical Decisions

In Giancola et al. [54], we present an expanded framework for adjudicating challenging ethical scenarios. We build off the work of Giancola et al. [53], and show that our updated framework, by incorporating Cognitive Likelihood and automated planning, is able to more accurately model the "Miracle on the Hudson" aviation emergency.

3.3.1.1 $\mathcal{IDCEC}$

Giancola et al. [54] was the first work to incorporate Cognitive Likelihood into an inductive cognitive calculus. Hence it was not replete with the same inference schemata of $\mathcal{IDCEC}_1$ which we introduce later in the present dissertation. It had only three inference schemata: $\{[I_4^\ell], [I_{WLP}^\ell], [I_\neg^\ell]\}$.[50]

3.3.1.2 Highly-Expressive Automated Planning

One necessary component of our framework — beyond what has already been introduced (i.e. inductive cognitive calculi, Cognitive Likelihood) — is an automated planner; in particular one that is fully compatible with our formalisms and their emphasis on declarative content and automated reasoning over that content in uncertain situations. The first modern automated planner was the Stanford Research Institute Problem Solver (STRIPS) [44], which produced a framework for planning upon which many modern planners are built.

The setup of a STRIPS problem is as follows. There is a set of formulae describing the initial state of the *world*, a set of *actions* which describe methods by which the planner can change the world state, and a *goal* set which denotes those formulae that the agent in question wishes to hold. The actions consist of three components: (1) a set of preconditions (formulae which must hold in order to perform the action), (2) a set of additions (formulae that will be added to the world by taking the action), and (3) a set of deletions (formulae to be removed from the world by taking the action).

The expressivity of formulae used to represent the world, actions, and goal was limited to propositional statements. For example, the goal that the book is not on the table could be represented by ¬On(book, table). In this work, we will need to be able to use quantified formulae, e.g. ¬∃ x On(x, table), to describe the world and goal.

The Planning Domain Definition Language (PDDL) [77] is a STRIPS-style planning language, which also supports quantification over zero-order formulae. While some quantification is supported, PDDL has serious restrictions on the syntax of formulae that can be supported. Arbitrary first-order-logic formulae are not allowed. Further, PDDL does not support modal operators such as those for *belief*, *knowledge*, or *obligation*; these are are necessary for modeling states of minds of agents. (E.g., in our case study below, we would

[50] The schemata were given different labels in Giancola et al. [54]. We use the $\mathcal{IDCEC}_1$ labels for consistency within the dissertation. See Figure 4.2 in §4.2 for definitions of these inference schemata.

ultimately want AI that is able to bring about "mental" goals, such as a pilot's believing that such and such a course of action is feasible.) Reasoning with such mental states is crucial in ethically charged situations.[51] Consider the natural language sentence "Alice *believes* that all pilots *believe*, before entering a cockpit, that they *know* ϕ.", where ϕ is some declarative knowledge. A formula that captures this sentence, which requires the ability to nest modal operators (e.g. belief, knowledge), cannot be expressed in PDDL:

$$\mathbf{B}(alice, t, \forall x\ \exists t_0 t_1\ \mathbf{B}(x, t_0, \mathbf{K}(x, t_0, \phi)) \wedge \mathit{EntersCockpit}(x, t_1) \wedge t_0 < t_1) \tag{3.81}$$

Another major limitation of the PDDL family of languages is that they require a finite and fixed universe of objects to be specified beforehand. In many uncertain situations, this is not realistic, as the number of relevant objects and entities will be unknown. Consider a situation in which a firefighting robot has to enter a building with the goal of rescuing any humans in the building. The agent has no prior knowledge of the number of humans in the building. PDDL languages are not directly amenable to modeling such situations.

Overall, then, we need a planning formalism (with an associated automated planner) that can handle arbitrary formulae for describing the world, states of minds, and an unknown set of objects. For a planner with such capabilities, we turn to Spectra [57], a STRIPS-style planner which can be integrated with reasoners for cognitive calculi [62]. While there is a efficiency disadvantage in using a more expressive planning formalism, efficiency gains in reasoning with cognitive calculi can be transferred to efficiency gains in Spectra.

3.3.1.3 Selecting Plans Using Cognitive Likelihood

In our framework, agents are given the following: (1) an obligation, (2) knowledge regarding the conditions required to satisfy the obligation, and (3) a set of (potentially inconsistent) ethically-charged beliefs regarding actions the agent can take to affect the status of the obligation. The agents make maximally ethical decisions by taking a course of action which maximizes the agent's belief that the obligations will stay (or become) satisfied.

The decision-making framework is outlined pictorally in Figure 3.2. An agent a is obligated to perform some action α, given that it believes some precondition ϕ holds. It also

[51] See §3.1.3 and §3.2.2 for two examples of ethically-charged situations in which the ascription of mental states is crucial to the success of the AI agents involved.

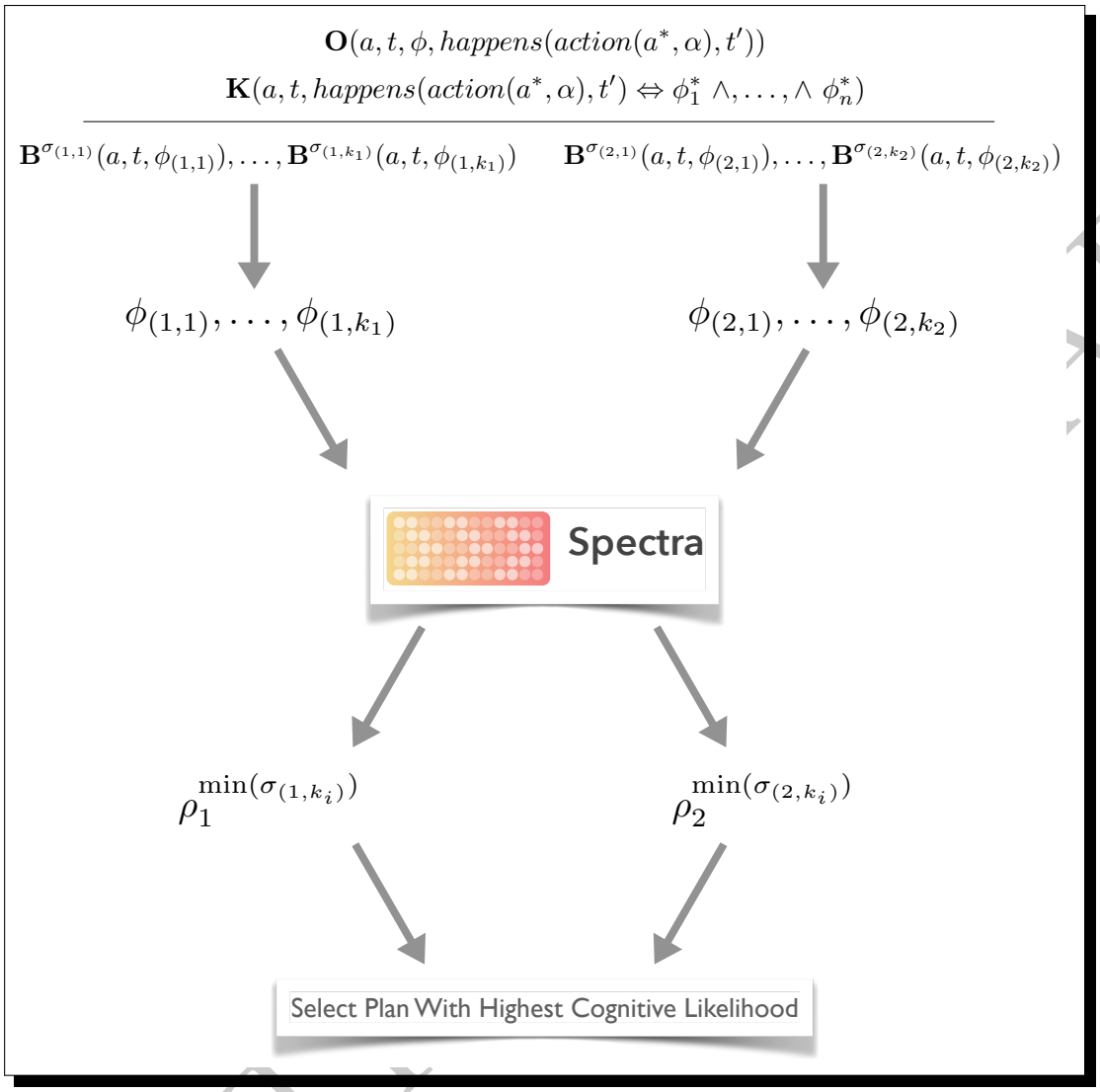

Figure 3.2: A Framework for Selecting Maximally Ethical Plans

knows the conditions that will enable α to happen (in Figure 3.2, $\phi_1^*, \ldots, \phi_n^*$).

Next, a has a set of beliefs regarding formulae pertinent to its obligations. Various subsets of those beliefs (in Figure 3.2, $\phi_{(1,1)}, \ldots, \phi_{(1,k_1)}$ and $\phi_{(2,1)}, \ldots, \phi_{(2,k_2)}$), are passed to Spectra, with the goal of generating plans which cause α to occur. We assign each of those plans a likelihood based on the likelihood of the weakest belief required to generate the plan. Finally, we select the plan with the highest likelihood as the one to enact.

3.3.1.4 Case Study: The "Miracle on the Hudson"

To display our framework "in action," we consider two potential arguments concerning what decision should be made in the case of US Airways Flight 1549, colloquially known

as the "Miracle on the Hudson." Recall that we gave a summary of the event in §3.2.1.1. Briefly, after losing thrust in both engines, the pilots had to quickly make the decision where to attempt an emergency landing, ultimately considering the following options: (a) attempt to return to LaGuardia Airport (LGA), (b) attempt to reach Teterboro Airport (TEB), or (c) attempt to ditch in the Hudson River.

3.3.1.5 The Setup

The setup of the framework for our case study is as follows. We have three agents: a_1 and a_2 will each present two inconsistent arguments regarding where the plane should be landed (in the following subsection), and a^* is the adjudicator who will decide which argument and plan to proceed with.

We denote the moment just after the plane flew into the flock of geese as t^*. At that time, a^* believes there is an emergency, and consequently, the agent is obligated to ensure that the landing site it selects is safe.

$$\mathbf{B}(a^*, t^*,\ \mathit{emergency}) \tag{3.82}$$

$$\mathbf{O}(a, t^*,\ \mathit{emergency}, \mathit{happens}(\mathit{action}(a^*, \mathit{ensure\_safe}(\mathit{landing\_site})), t^{*'})) \tag{3.83}$$

The agent also knows the conditions required for a landing site to be safe: it must be close enough to reach, long and wide enough, and far enough from people, as without thrust, the pilots' ability to maneuver the plane will be more limited than usual.

$$\mathbf{K}\Big(a, t^*,\ \mathit{happens}\left(\mathit{action}(a^*, \mathit{ensure\_safe}(\mathit{landing\_site})), t^{*'}\right) \Leftrightarrow \vdash \text{Safe}(\mathit{landing\_site})\Big) \tag{3.84}$$

$$\mathbf{K}\left(a, t^*,\ \forall \ell\ \text{Safe}(\ell) \Leftrightarrow \bigwedge \left\{ \begin{array}{c} \text{CloseEnough}(\ell), \\ \text{LongEnough}(\ell), \\ \text{WideEnough}(\ell), \\ \text{FarEnoughFromPeople}(\ell) \end{array} \right\} \right) \tag{3.85}$$

3.3.1.6 The Arguments

We next give two arguments in favor of selecting different landing locations based on the conditions of Flight 1549, then show how our framework would generate plans for each, and finally, how it would select a plan to execute.

Argument 1 The first agent argues for the following two statements:

$$\mathbf{B}^{3}(a_1, t^*, \text{CloseEnough}(lga)) \tag{3.86}$$

$$\mathbf{B}^{1}(a_1, t^*, \text{FarEnoughFromPeople}(lga)) \tag{3.87}$$

The first formula states that it is OVERWHELMINGLY LIKELY that LaGuardia Airport was close enough for the pilots to successfully land there. This is justified by the several studies and simulations performed since the event which identified many feasible trajectories to enable landing at several different runways at LGA (e.g. see Paul et al. [89], Atkins [6]).

The second states that it is MORE LIKELY THAN NOT that LGA is far enough from people to ensure a safe landing despite the conditions (i.e. loss of thrust in both engines at low altitude, occurring in — to quote Captain Sullenberger — "a highly developed, metropolitan area" [85]). The likelihood is necessarily weak, as the corresponding justification is weak. As there is no data to go on, one can only speculate that based on the Captain's training, and Air Traffic Control's ability to clear a runway in time, that it is possible that the plane could have been landed at LGA without harming anyone on the ground.

Argument 2 The second argument asserts the following two statements:

$$\mathbf{B}^{-2}(a_2, t^*, \text{CloseEnough}(teb)) \tag{3.88}$$

$$\mathbf{B}^{-2}(a_2, t^*, \text{FarEnoughFromPeople}(lga)) \tag{3.89}$$

The first statement, that it is UNLIKELY that Teterboro Airport is close enough, was asserted without justification by Captain Sullenberger in the public hearing on the accident [85]. He likely intended to imply an implicit justification that it was obvious to him based on his experience as a pilot.

Second, a_2 asserts that it is UNLIKELY that LGA was far enough from people to ensure a safe landing. Note that this belief is directly inconsistent with a belief of a_1; namely, $\mathbf{B}^1(a_1, t^*, \text{FarEnoughFromPeople}(lga))$. Again, this is justified by a statement provided by Captain Sullenberger during the public hearing:

> Looking at where we were and how much time, altitude, and distance would be required to turn back toward LaGuardia and then fly toward LaGuardia, I determined quickly that that was going to be problematic, and it would not be a realistic choice, and I couldn't afford to be wrong. (pg. 25, [85])

It is clear that, had Captain Sullenberger chosen to attempt a landing at LGA, he would've risked the lives of people at and around LGA, *in addition* to the inevitable risk already imposed on those in the plane by the emergency.

3.3.1.7 The Framework, Applied

We now present the application of our framework, in order to adjudicate these clearly inconsistent arguments[52] and generate a plan. First, the content of each agent's beliefs are passed separately to Spectra. Therefore, the first agent passes:

$$\text{CloseEnough}(lga) \tag{3.90}$$

$$\text{FarEnoughFromPeople}(lga) \tag{3.91}$$

and the second agent passes:

$$\neg\text{CloseEnough}(teb) \tag{3.92}$$

$$\neg\text{FarEnoughFromPeople}(lga) \tag{3.93}$$

In order to generate plans, Spectra is given the following actions:

[52] We note that when we refer to *arguments* being inconsistent, this is to say that the arguments each assert a set of formulae, and from the union of those sets, a contradiction can be deduced.

```
(define-action considerRunwayLanding [?r]
  {:preconditions [(CloseEnough ?r)
                   (FarEnoughFromPeople ?r)
                  ]

   :additions     [(LongEnough ?r)
                   (WideEnough ?r)
                  ]

   :deletions     [ ]
  }
)

(define-action considerTerrainLanding []
  {:preconditions [(not (and (Safe lga) (Safe teb)))
                  ]

   :additions     [(CloseEnough hud)
                   (LongEnough  hud)
                   (WideEnough  hud)
                   (FarFromPeopleEnough hud)
                  ]

   :deletions     [ ]
  }
)
```

The first action requires that, in order to consider landing at a particular runway, the ethically-charged propositions are first satisfied. It then adds that the runway satisfies basic requirements. The idea here is that, if implemented "for real," Spectra would be integrated with systems which could provide the necessary data, i.e. the length and width of the runway being considered, and the length and width required.

The second action allows the planner to consider off-runway landing options *only* if

the runway options have been exhausted (that is, it has been determined that none of them meet the imposed safety requirements). As with the first action, Spectra would need to be integrated with another system. In this case, our simulation assumes Spectra would have access to a vision-based landing-site detection system, such as that presented in Shen et al. [111]. Shen et al. specify that (emphasis ours):

> A landing-site is considered safe only if its surface is *smooth* and if its *length* and *width* are adequate. (pg. 295, [111])

At the public hearing, Sullenberger stated that (emphasis ours):

> [Other than LGA or TEB,] the only place in a highly developed, metropolitan area, *long enough*, *wide enough*, *smooth enough* to land was the river. (pg. 25, [85])

Hence we can confidently say that, were Shen et al.'s system integrated with Spectra in this case, the river would have been the only landing option returned.

Each agent's input to Spectra returns a single plan. The former indicates that the pilot can land LGA, as all safety requirements have been satisfied. Alternatively, the latter is able to prove that neither LGA nor TEB are safe options, and hence seeks out off-runway options, and finds the Hudson as the only option.

Finally, note that the weakest likelihood used by agent 1 is MORE LIKELY THAN NOT (= 1) and the weakest of agent 2's argument is LIKELY (= 2). Hence the framework would conclude that agent 2's argument, and corresponding plan, are to be used.

3.3.2 Solving the *Intensional* Suppression Task

Finally, in Giancola et al. [51], we introduce and solve an explicitly *intensional* version of Ruth Byrne's Suppression Task [30].

3.3.2.1 The Suppression Task

The Suppression Task is reported in Byrne [30]. Three groups of subjects were asked to select which proposition from among a trio of them "follows" from a set of premises. Each group of subjects was given a different set of premises. Group 1 (= G1) was given this pair of premises:

(p1) If she has an essay to finish, then she will study late in the library.

(p2) She has an essay to finish.

This group's options to select from were the following three:

(o1) She will study late in the library.

(o2) She will not study late in the library.

(o3) She may or may not study late in the library.

Among G1, 96% selected (o1). G2 was given premises consisting of (p1) and (p2), plus the following premise:

(p3) If she has a textbook to read, then she will study late in the library.

In G2, again 96% of its members selected option (o1). G3 received (p1) and (p2), plus this premise:

(p4) If the library stays open, then she will study late in the library.

This time things turned out quite differently: only 38% of G3 selected (o1). This finding is interesting because, assuming the reasoning taking place is deductive, the addition of either (p3) or (p4) shouldn't change the valid inference made from (p1) and (p2). However, Byrne found that premises of a certain type, which she referred to as *alternative* premises, caused the majority of subjects to suppress the valid inference. In the above example, we expect that those subjects interpreted (p4) as a necessary condition; and since it isn't stated explicitly that the library is open, they were unsure whether or not she *could* study late in the library.

We argue that the Suppression Task is best modeled by a defeasible, intensional logic.[53] To strengthen this argument, we next discuss a variant of the Suppression Task which is explicitly and unavoidably *intensional* in nature. Hence all previous models of the Suppression Task, (to the author's knowledge) all extensional, are incapable of modeling our intensional version. However, as will be seen shortly, minimal changes are made to the task, so we do not expect that it would be any more challenging for human reasoners (despite the additional challenge from a formal reasoning perspective).

[53]For such modeling we point the interested reader to Bringsjord et al. [21].

3.3.2.2 The Intensional Suppression Task

In our intrinsically intensional version of the Suppression Task, the three premises are these:

(p1$_{int}$) If Mary has an essay to finish, then Mary will study late in the library.

(p2$_{int}$) Mary's mother knows that Mary's father knows that Mary has an essay to finish.

(p3$_{int}$) If the library stays open, then Mary will study late in the library.

We next have the following three options:

(o1$_{int}$) Mary will study late in the library.

(o2$_{int}$) Mary will not study late in the library.

(o3$_{int}$) Mary may or may not study late in the library.

Now imagine posing this question to a rational human-level agent: Which of these three options logically follow from the three premises? The correct answer is (o1$_{int}$), only. However, Byrne's original experiment found that the addition of (p3$_{int}$) led most human cognizers to suppress the valid inference. It can be expected that most people, as in Byrne's original experiment, would fail to correctly select (o1$_{int}$) for generally the same reasons they failed to select (o1) in the original ST.[54]

The Solution, Using Cognitive Likelihood Again, only a limited subset of inference schemata were utilized in Giancola et al. [51]; specifically, $\{[I_4], [I^{\ell}_{WLP}]\}$.[55]

Assume we have a fully rational agent a who is capable of reasoning via $\mathcal{IDCEC}$. Proving the formal equivalent of (o1$_{int}$) is fairly straightforward, given the established inference schemata:

$$\mathbf{B}^5(\mathfrak{a}, \mathbf{K}(m, \mathbf{K}(f, \text{ToFinish}(mary, essay)))) \quad \text{[Given]} \quad (3.94)$$

$$\mathbf{B}^5(\mathfrak{a}, \mathbf{K}(f, \text{ToFinish}(mary, essay))) \quad [[I^{\ell}_{WLP}], \text{ using } [I_4]] \quad (3.95)$$

$$\mathbf{B}^5(\mathfrak{a}, \text{ToFinish}(mary, essay))) \quad [[I^{\ell}_{WLP}], \text{ using } [I_4]] \quad (3.96)$$

$$\mathbf{B}^5(\mathfrak{a}, \text{ToFinish}(mary, essay) \rightarrow \text{StudyLate}(mary)) \quad \text{[Given]} \quad (3.97)$$

[54] We note that while we intuitively hypothesize that the changes we made to the experiment would not impact the results, a new experiment is necessary in order to confirm our hypothesis.

[55] As we've mentioned regarding other prior works, the labels of the schemata were different in Giancola et al. [51]; they were labeled $\{[I_{\mathbf{K}}], [I_{\mathbf{B}}]\}$ respectively.

$\mathbf{B}^5(\mathfrak{a}, \text{StudyLate}(mary))$ [$[I^{\ell}_{WLP}]$, using modus ponens] (3.98)

CHAPTER 4
THE INDUCTIVE DEONTIC COGNITIVE EVENT CALCULUS

In this chapter, we introduce a full formalization of the Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$). It builds off of the purely deductive $\mathcal{DCEC}$, whose specification was given in §2.4. That is, the signature of $\mathcal{IDCEC}$ includes all of the elements of the signature of $\mathcal{DCEC}$, as well as the additional elements given in the Figure 4.1. A subset of the inference schemata of $\mathcal{DCEC}$ are included in $\mathcal{IDCEC}$; they are given in Figure 4.2 in the subsection titled "Deductive Schemata". Also, note that we will specifically call the cognitive calculus presented herein $\mathcal{IDCEC}_1$. That is because, like $\mathcal{DCEC}$, $\mathcal{IDCEC}$ is really a *family* of cognitive calculi, of which the specific calculus presented herein is a member.[56] As discussed in §3, prior works have presented other members of this family, e.g. Bringsjord et al. [27], Giancola et al. [53], Giancola et al. [54].

4.1 Signature

The signature of $\mathcal{IDCEC}_1$ is given in Figure 4.1. To review, the signature of *any* cognitive calculus includes four elements: (1) a list of sorts, (2) a list of function signatures, (3) a grammar for terms, and (4) a grammar for syntactic forms. The most significant element of the signature of $\mathcal{IDCEC}$ is the new syntactic form $\mathbf{B}^{\sigma}(a, t, \phi)$ which can be read as "Agent a believes at time t with likelihood σ that ϕ holds."

Note that the signature of $\mathcal{IDCEC}_1$ subsumes the signature of $\mathcal{DCEC}$ (see Figure 2.1). That is, all of the sorts, functions, and syntactic forms that are valid in $\mathcal{DCEC}$ are valid in $\mathcal{IDCEC}_1$ as well.

4.2 Inference Schemata

The inference schemata of $\mathcal{IDCEC}_1$ are given in Figure 4.2, organized in four categories. The labels of the inductive schemata include a superscript ℓ, to indicate that they incorporate

[56] There is a second reason for this specific naming convention: while the calculus presented herein is a significant milestone in the development of inductive cognitive calculi (hence, it is the first "major" version), it is by no means the be-all-end-all. Further R&D into inductive cognitive calculi — which may add, modify, or remove inference schemata — is not only possible but expected.

beliefs with likelihood. Also, the deductive schemata are given the same label as in the $\mathcal{DCEC}$ schemata (Figure 2.2). Next, we explain the three categories of inductive schemata.

$\mathcal{IDCEC}_1$ Signature

$$
\begin{aligned}
S &::= \mathsf{Number} \mid \mathsf{List} \\
f &::= \begin{cases} min : \mathsf{List[Number]} \rightarrow \mathsf{Number} \\ max : \mathsf{List[Number]} \rightarrow \mathsf{Number} \end{cases} \\
\phi &::= \left\{ \mathbf{B}^{\sigma}(a, t, \phi) \right. \\
&\text{where } \sigma \in [-5, -4, \ldots, 4, 5]
\end{aligned}
$$

Figure 4.1: Signature of the Inductive $\mathcal{DCEC}$

4.2.1 Introduction Schemata

The Introduction Schemata, naturally, enable the introduction of a belief at each of the six non-negative likelihood values.

0. a can infer a COUNTERBALANCED belief in ϕ with no justification. This is for a few reasons: a belief at this level is indicating the total lack of strength of belief in one direction or the other (that is, in ϕ or $\neg\phi$). Also, any (potentially contradictory) belief with a non-zero likelihood value will defeat this one.

1. a can infer a MORE LIKELY THAN NOT belief in ϕ if a trustworthy source s says ϕ. Note that, like schema $[I_{12}]$ of $\mathcal{DCEC}$, we implicitly assume that s is trustworthy. Determining the preconditions for trust, as well as related issues like inference schemata for dealing with untrustworthy and adversarial agents, is out of scope.

2. a can infer a LIKELY belief in ϕ if multiple trustworthy sources s_i say ϕ. This is grounded in the standard notion — in e.g., journalism, case studies research [81] — that having multiple sources of the same information provides additional certainty. The number of sources γ necessary will depend on the context. Consider determining the validity of a news story. For a relatively non-controversial story, it may suffice to check two sources. However more controversial stories could pose a risk of increased bias, and hence one may desire more sources.

3. a can infer an OVERWHELMINGLY LIKELY belief in ϕ if it perceived ϕ at t_1 but is not currently perceiving ϕ or $\neg\phi$. Note that this schema also infers that a does *not* believe ϕ at the level of EVIDENT. It is likely not clear why this is necessary; we will discuss the reason why it is necessary after we introduce the schema for belief propagation $[I^{\ell}_{PROP}]$.

4. a can infer an EVIDENT belief in ϕ if it is currently perceiving ϕ. Similar to the issue of trust in $[I^{\ell}_1]$ and $[I^{\ell}_2]$, we leave aside the issue of *compromised perception* in this work.[57] For a longer discussion of the issues of reasoning under compromised perception, see Bringsjord et al. [22].

5. a can infer a CERTAIN belief in ϕ if an authoritative agent says ϕ.[58] Note that what we mean by *authoritative* is that the agent can be expected *in the given context* to know, without a shadow of doubt, that ϕ is true, generally because they are in control of the situation in some way.[59] Like the issue of trust in $[I^{\ell}_1]$ and $[I^{\ell}_2]$ and the issue of compromised perception in $[I^{\ell}_4]$, we assume that the relevant context is satisfied when we say that an agent is authoritative. We discuss the importance of Authoritative Context further in Appendix B.

4.2.2 Defeasible Belief Generation

The schemata in this group are what enable the defeasibility of $\mathcal{IDCEC}$. That is, they allow the generation of new beliefs which may contradict prior beliefs (when considered directly without likelihood values). The first schema in this group, the *Weakest Link Principle* (WLP), essentially warrants a to infer a belief in ϕ if ϕ is provable from a's belief set, given the following conditions:

- a's belief set is not inconsistent. If it is, then anything can be proved from it.

[57] Compromised perception includes perception while e.g. intoxicated, hallucinating, or when in the presence of visual impairments e.g. fog.

[58] We note that, for all likelihood levels but this one in particular, there are many conceivable definitions. This conception of CERTAIN belief will be useful for the use-cases that it will be deployed in in the present dissertation. One other potential definition, which may be more palatable to some readers, is that CERTAIN beliefs can only be inferred in theorems of e.g. FOL, Peano's Axioms, ZFC Set Theory.

[59] E.g. a contestant in a gameshow can generally assume that statements made by the host are true. See §7.4 for an example where this is the case, and this schema is utilized. Similarly, test subjects can infer CERTAIN beliefs in statements made by experimenters. See §3.3.2 for a prior work which utilized this concept, although this schema had not yet been formalized.

- The level of belief in ϕ must be the level of the weakest belief in the belief set used to prove ϕ.

The second schema in the group enables a to iteratively propagate its beliefs through time, so long as it has not found a reason to drop the belief at the current time. This is why $[I_3^\ell]$ must infer both a belief at level-3 and the negation of a belief at level-4. Otherwise, agents could infer an EVIDENT belief in a percept at the time that they see it, then freely propagate that belief even if they are no longer perceiving it. Since $[I_3^\ell]$ infers the negation of the belief at level-4, $[I_{PROP}^\ell]$ cannot be applied.

Finally, the third schema allows a to drop a belief in ϕ if it has a belief with stronger likelihood in $\neg\phi$.

4.2.3 Symmetry of Negative Likelihood & Negated Subformula

The final schema expresses the fact that a belief in $\neg\phi$ at likelihood σ is equivalent to a belief in ϕ at likelihood $-\sigma$.

4.3 Schema Usage Examples

We next give micro proofs involving each individual inference schema, both to justify the inclusion of and demonstrate the use of each schema.[60] We will use a set of running examples in which an agent is generating beliefs regarding the weather. Some examples discuss the weather "tomorrow," which is indicated by d_2.

4.3.1 Introduction Schemata

0. a considers whether or not it may rain tomorrow. With no prior knowledge related to potential storms arriving in their area tomorrow, a can only infer a COUNTERBALANCED belief that it will rain tomorrow. (a could also infer a COUNTERBALANCED belief that it will not rain tomorrow.)

$$\therefore \mathbf{B}^0(a, t_0, \text{Raining}(d_2)) \qquad [I_0^\ell] \qquad (4.1)$$

[60] For implementations of these examples, solvable with ShadowAdjudicator, see `https://github.com/RAIRLab/ShadowAdjudicator/tree/4bc683407fddd3da8509a58fa17d01ecd5dd87b9/adjudicator/diss_examples` (Last Accessed February 2, 2023). For the corresponding output, see Appendix A.

$\mathcal{IDCEC}_1$ Inference Schemata

Introduction Schemata

$$\frac{}{\mathbf{B}^0(a,t,\phi)}\ [I_0^\ell]$$

$$\frac{\mathbf{S}(s,a,t_1,\phi),\ t_1<t_2}{\mathbf{B}^1(a,t_2,\phi)}\ [I_1^\ell] \qquad \frac{\bigwedge_i \mathbf{S}(s_i,a,t_i,\phi),\ max(t_i)<t^*,\ i>\gamma>1}{\mathbf{B}^2(a,t^*,\phi)}\ [I_2^\ell]$$

$$\frac{\mathbf{P}(a,t_1,\phi),\ \neg\mathbf{P}(a,t_2,\phi),\ \neg\mathbf{P}(a,t_2,\neg\phi),\ t_1<t_2}{\mathbf{B}^3(a,t_2,\phi),\ \neg\mathbf{B}^4(a,t_2,\phi)}\ [I_3^\ell]$$

$$\frac{\mathbf{P}(a,t,\phi)}{\mathbf{B}^4(a,t,\phi)}\ [I_4^\ell] \qquad \frac{\mathbf{S}(s^*,a,t_1,\phi),\ \mathbf{K}(a,t_1,\text{Authoritative}(s^*)),\ t_1<t_2}{\mathbf{B}^5(a,t_2,\phi)}\ [I_5^\ell]$$

Defeasible Belief Generation

$$\frac{\mathbf{B}^{\sigma_1}(a,t,\phi_1),\ldots,\mathbf{B}^{\sigma_m}(a,t,\phi_m),\ \{\phi_1,\ldots,\phi_m\}\vdash\phi,\ \{\phi_1,\ldots,\phi_m\}\nvdash\bot}{\mathbf{B}^{min(\sigma_1,\ldots,\sigma_m)}(a,t,\phi)}\ [I_{WLP}^\ell]$$
where $\sigma_i\in[0,1,\ldots,4,5]$

$$\frac{\mathbf{B}^{\sigma}(a,t_1,\phi),\ \Gamma\nvdash\neg\mathbf{B}^{\sigma}(a,t_2,\phi),\ t_1<t_2}{\mathbf{B}^{\sigma}(a,t_2,\phi)}\ [I_{PROP}^\ell]$$

$$\frac{\mathbf{B}^{\sigma_1}(a,t_1,\phi),\ \mathbf{B}^{\sigma_2}(a,t_2,\neg\phi),\ t_1<t_2,\ \sigma_1<\sigma_2}{\neg\mathbf{B}^{\sigma_1}(a,t_2,\phi)}\ [I_{DROP}^\ell]$$

Symmetry of Negative Likelihood & Negated Subformula

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

Deductive Schemata

$$\frac{\mathbf{K}(a,t_1,\Gamma),\ \ \Gamma\vdash\phi,\ \ t_1\le t_2}{\mathbf{K}(a,t_2,\phi)}\ [I_{\mathbf{K}}] \qquad \frac{\mathbf{B}(a,t_1,\Gamma),\ \ \Gamma\vdash\phi,\ \ t_1\le t_2}{\mathbf{B}(a,t_2,\phi)}\ [I_{\mathbf{B}}] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi)\rightarrow\mathbf{B}(a,t,\phi))}\ [I_2]$$

$$\frac{\mathbf{C}(t,\phi),\ t\le t_1,\ldots,t\le t_n}{\mathbf{K}(a_1,t_1,\ldots\mathbf{K}(a_n,t_n,\phi)\ldots)}\ [I_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [I_4] \qquad \frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t',happens(action(a^*,\alpha),t'))}\ [I_{13}]$$

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

Figure 4.2: Inference Schemata of the Inductive $\mathcal{DCEC}$

1. a turns on a local news channel and the station's meteorologist m_1 says that rain is expected tomorrow in a's area. Since a meteorologist is a trustworthy source in this context, a can infer a belief that it is MORE LIKELY THAN NOT that it will rain tomorrow.

$$\mathbf{S}(m_1, a, t_1, \text{Raining}(d_2)) \tag{4.2}$$
$$\therefore \mathbf{B}^1(a, t_2, \text{Raining}(d_2)) \qquad [I_1^\ell] \tag{4.3}$$

2. a changes the channel to a national news station, whose meteorologist m_2 also states that it will rain tomorrow. Since two sources is a reasonable amount for a non-controversial topic like the weather, a can now infer a belief that it is LIKELY that it will rain tomorrow.

$$\mathbf{S}(m_1, a, t_1, \text{Raining}(d_2)) \tag{4.4}$$
$$\mathbf{S}(m_2, a, t_2, \text{Raining}(d_2)) \tag{4.5}$$
$$\therefore \mathbf{B}^2(a, t_3, \text{Raining}(d_2)) \qquad [I_2^\ell] \tag{4.6}$$

3. Tomorrow arrives, and a looks out a window and perceives it raining outside. a then goes into their basement where they can no longer directly perceive the presence or absence of rain. At that time, a can infer a belief that it is OVERWHELMINGLY LIKELY that it is raining.

$$\mathbf{P}(a, t_4, \text{Raining}(t_4)) \tag{4.7}$$
$$\neg\mathbf{P}(a, t_5, \text{Raining}(t_5)) \wedge \neg\mathbf{P}(a, t_5, \neg\text{Raining}(t_5)) \tag{4.8}$$
$$\therefore \mathbf{B}^3(a, t_5, \text{Raining}(t_5)) \qquad [I_3^\ell] \tag{4.9}$$

4. Previously, when a was actively perceiving rain, it could infer a belief that it it EVIDENT that it is raining.

$$\mathbf{P}(a, t_4, \text{Raining}(t_4)) \tag{4.10}$$
$$\therefore \mathbf{B}^4(a, t_4, \text{Raining}(t_4)) \qquad [I_4^\ell] \tag{4.11}$$

5. a is an actor in a movie, and the director d tells a that it will be raining in the next scene, which will be achieved using a rain truck. Since the director has complete control

of the set, a can safely assume d is authoritative in this context. Hence a can infer a belief that it is CERTAIN that it is raining in the next scene.

$$\mathbf{S}(d, a, t_6, \text{Raining}(\textit{next\_scene})) \tag{4.12}$$

$$\mathbf{K}(a, t_6, \text{Authoritative}(d)) \tag{4.13}$$

$$\therefore \mathbf{B}^5(a, t_7, \text{Raining}(\textit{next\_scene})) \quad [I_5^\ell] \tag{4.14}$$

4.3.2 Defeasible Belief Generation

WLP. a turns on a local news channel and the station's meteorologist m_1 says that rain is expected tomorrow in a's area. As before, a can infer a belief that it is MORE LIKELY THAN NOT that it will rain tomorrow. a then changes the channel to a national news station, whose meteorologist m_2 states that there will be windy conditions tomorrow. a can infer a second belief that it is MORE LIKELY THAN NOT that it will by windy tomorrow. Finally, using the schema at hand, a can infer a belief that it will rain and be windy tomorrow.

$$\mathbf{S}(m_1, a, t_1, \text{Raining}(d_2)) \tag{4.15}$$

$$\therefore \mathbf{B}^1(a, t_2, \text{Raining}(d_2)) \quad [I_1^\ell] \tag{4.16}$$

$$\therefore \mathbf{B}^1(a, t_5, \text{Raining}(d_2)) \quad [I_{PROP}^\ell] \tag{4.17}$$

$$\mathbf{S}(m_2, a, t_3, \text{Windy}(d_2)) \tag{4.18}$$

$$\therefore \mathbf{B}^1(a, t_4, \text{Windy}(d_2)) \quad [I_1^\ell] \tag{4.19}$$

$$\therefore \mathbf{B}^1(a, t_5, \text{Windy}(d_2)) \quad [I_{PROP}^\ell] \tag{4.20}$$

$$\therefore \mathbf{B}^1(a, t_5, \text{Raining}(d_2) \wedge \text{Windy}(d_2)) \quad [I_{WLP}^\ell] \tag{4.21}$$

PROP. Next, a turns off the television and reads a book from t_5 to t_6. Gaining no new information that would refute their previous belief, they are able to propagate their belief that it will rain and be windy tomorrow to t_6.

$$\mathbf{B}^1(a, t_5, \text{Raining}(d_2) \wedge \text{Windy}(d_2)) \tag{4.22}$$

$$\therefore \mathbf{B}^1(a, t_6, \text{Raining}(d_2) \wedge \text{Windy}(d_2)) \quad [I_{PROP}^\ell] \tag{4.23}$$

DROP. Tomorrow arrives and a perceives that it does not rain. Hence they must drop their

prior belief that it would rain.

$$\mathbf{B}^{1}(a, t_5, \text{Raining}(d_2)) \tag{4.24}$$

$$\mathbf{P}(a, d_2, \neg\text{Raining}(d_2)) \tag{4.25}$$

$$\therefore \mathbf{B}^{4}(a, d_2, \neg\text{Raining}(d_2)) \quad [I_4^{\ell}] \tag{4.26}$$

$$\therefore \ \neg\mathbf{B}^{1}(a, d_2, \text{Raining}(d_2)) \quad [I_{DROP}^{\ell}] \tag{4.27}$$

4.3.3 Symmetry of Negative Likelihood & Negated Subformula

a believes it is MORE UNLIKELY THAN NOT that it will rain tomorrow. They can use this schema to identify that it is equivalent to hold a belief that it is MORE LIKELY THAN NOT that it will not rain tomorrow.

$$\mathbf{B}^{-1}(a, t_1, \text{Raining}(d_2)) \tag{4.28}$$

$$\therefore \mathbf{B}^{1}(a, t_1, \neg\text{Raining}(d_2)) \quad [I_{\neg}^{\ell}] \tag{4.29}$$

CHAPTER 5
SHADOWADJUDICATOR

To enable an AI agent to reason with $\mathcal{IDCEC}$, it must be implemented in an automated reasoner. We implemented our particular calculus, $\mathcal{IDCEC}_1$, in an automated reasoner called ShadowAdjudicator.[61]

ShadowAdjudicator builds directly off of an automated reasoner for $\mathcal{DCEC}$ called ShadowProver. As the reader may notice later in the present chapter, the algorithm that ShadowAdjudicator implements to find arguments in $\mathcal{IDCEC}_1$ is inspired by the "shadowing" algorithm of ShadowProver. Before discussing ShadowAdjudicator further, we first discuss ShadowProver.

5.1 ShadowProver

ShadowProver is an automated reasoner for $\mathcal{DCEC}$ [62]. It utilizes a novel technique — *shadowing* — to find $\mathcal{DCEC}$ proofs. Govindarajulu et al. [62] describe shadowing as follows:

> Our algorithm is based on a technique we term **shadowing**. At at high-level, we alternate between calling a first-order theorem prover and applying modal inference schemata. When we call the first-order prover, all modal atoms are converted into propositional atoms (i.e. the former are shadowed), to prevent substitution into modal contexts. This approach achieves speed without sacrificing consistency. (pg. 46, [62])

In the next section, we will extend the notion of shadowing to *annotated* modal formulae — specifically, beliefs with likelihood annotations. Essentially, in contexts where we wish to reason using purely deductive modal inference schemata (i.e. using $\mathcal{DCEC}$ inference schemata), we will shadow beliefs with likelihood by removing the likelihood. However care must be taken to avoid inconsistent reasoning. We will address these issues in the following section, after discussing the ShadowAdjudicator algorithm broadly.

[61] ShadowAdjudicator is open-source under an AGPL-3.0 license and is available at `https://github.com/RAIRLab/ShadowAdjudicator`. (Last Accessed February 2, 2023)

5.2 The ShadowAdjudicator Algorithm

The ShadowAdjudicator Algorithm is presented formally in Algorithm 1. Given a set of assumed formulae Γ and a goal formula ϕ, ShadowAdjudicator will either return a proof of ϕ from Γ or FAILED. Like the ShadowProver Algorithm, there are two main steps: (1) apply inductive modal inference schemata, and (2) shadow annotated formulae and call ShadowProver.

Input: Assumption Base Γ, Goal ϕ
Output: A proof of $\Gamma \vdash \phi$ or FAILED
while $\phi \notin \Gamma$ **do**
  $\Gamma' =$ apply *expanders* (forward chaining) to Γ;
  $\Gamma' =$ apply *constructors* (backward chaining) to Γ';
  if ϕ *is not annotated* **then**
    $\Gamma'' =$ shadow Γ' (remove likelihood annotations);
    if Γ'' *is inconsistent* **then**
      $\Gamma'' =$ remove annotated beliefs from Γ;
    end
    response = call ShadowProver on Γ'', ϕ;
    if *response* $\neq$ *FAILED* **then**
      return proof of ϕ;
    end
  end
  if $\Gamma == \Gamma'$ **then**
    return FAILED;
  else
    $\Gamma = \Gamma'$
  end
end
return proof of ϕ;

Algorithm 1: ShadowAdjudicator Algorithm

5.2.1 Applying Inductive Modal Inference Schemata

Every $\mathcal{IDCEC}_1$ inference schema is implemented as either an *expander* or *constructor*. Essentially, expanders perform forward chaining, while constructors perform backward chaining. An expander searches Γ for formulae which match the premises of some inference schema and "expand" Γ by applying the schema and adding the resulting formula(e) to Γ. A constructor attempts to prove ϕ by recursively proving the precondition of some inference schema which would infer ϕ if the precondition is satisfied.

Table 5.1: Implementation Type of $\mathcal{IDCEC}_1$ Inference Schemata

Expanders	Constructors
$[I_1^\ell]$	$[I_0^\ell]$
$[I_2^\ell]$	$[I_{PROP}^\ell]$ [62]
$[I_3^\ell]$	$[I_{WLP}^\ell]$
$[I_4^\ell]$	
$[I_5^\ell]$	
$[I_{DROP}^\ell]$	
$[I_\neg^\ell]$	

The decision to implement an inference schema as either an expander or constructor is dependent on the nature of the schema. Generally, we implement a schema as an expander if it doesn't require knowledge of the goal. For example, $[I_4^\ell]$ can be implemented as an expander. ShadowAdjudicator searches Γ for any perceptions and expands Γ to include corresponding EVIDENT beliefs.[63] The Weakest Link Principle, $[I_{WLP}^\ell]$ is more efficiently implemented as a constructor. Assuming ϕ is a belief with likelihood, ShadowAdjudicator attempts to prove ϕ by finding a subproof of the subformula of ϕ from the subformulae of the agent's belief set. Table 5.1 indicates how all $\mathcal{IDCEC}_1$ inference schemata are implemented in ShadowAdjudicator.[64]

5.2.2 Shadowing Annotated Formulae

Recall that, in addition to the inductive inference schemata in Figure 4.2, $\mathcal{IDCEC}_1$ contains a subset of the deductive inference schemata of $\mathcal{DCEC}$, shown in the box titled "Deductive Schemata" within Figure 4.2. We utilize ShadowProver to generate subproofs which involve only deductive inference schemata. However, as mentioned earlier in the chapter, we must take steps to avoid inconsistent reasoning, to wit:

[62] Conceptually speaking, belief propagation is really an expander, as it is attempts to match formulae in Γ to its preconditions. However, as its precondition contains a provability query, it requires the lists of expanders and constructors (in order to search for a sub-proof of the query). Since expanders only have access to Γ, it was cleaner to implement it as a constructor.

[63] Note that, before adding a formula to Γ, ShadowAdjudicator *always* checks if the formula has already been added via reasoning elsewhere in the proof. If so, the formula is not added again. This prevents infinite loops of reasoning, such as would otherwise be possible e.g. by iterated application of $[I_\neg^\ell]$.

[64] For the implementation of the expanders, see `https://github.com/RAIRLab/ShadowAdjudicator/blob/4bc683407fddd3da8509a58fa17d01ecd5dd87b9/adjudicator/expanders/CogLikelihood.py` (Last Accessed February 2, 2023). Likewise, for the implementation of the constructors, see `https://github.com/RAIRLab/ShadowAdjudicator/blob/4bc683407fddd3da8509a58fa17d01ecd5dd87b9/adjudicator/constructors/CogLikelihood.py` (Last Accessed February 2, 2023).

1. Shadow all annotated formulae in Γ. In this context, we are shadowing from annotated modal formulae to (unannotated) modal formulae. E.g. the shadow of $\mathbf{B}^3(a,t,\phi)$ is $\mathbf{B}(a,t,\phi)$.

   (a) Beliefs with negative likelihood are not shadowed, as the negative epistemic value would be lost. Instead the corresponding belief with positive likelihood is inferred (via $[I^{\ell}_{\neg}]$) and shadowed. E.g. from $\mathbf{B}^{-2}(a,t,\phi)$ we infer $\mathbf{B}^2(a,t,\neg\phi)$ then shadow this to $\mathbf{B}(a,t,\neg\phi)$.

2. The prior step could introduce contradictory beliefs. For example $\Gamma := \{\mathbf{B}^3(a,t,\phi), \mathbf{B}^2(a,t,\neg\phi)\}$ is fine, but the shadowed set $\Gamma' := \{\mathbf{B}(a,t,\phi), \mathbf{B}(a,t,\neg\phi)\}$ is not. Therefore we next check if this resulting set is inconsistent,[65] and if so, can only continue by removing all annotated formulae from Γ.

3. Finally, we call on ShadowProver to find a proof of the goal from the modified declarative base.

[65] In practice, we check for inconsistency by trying to prove a reserved propositional atom which does not appear in the declarative base. Hence it is only provable from the base if the base is inconsistent.

CHAPTER 6
CASE STUDY: AUTONOMOUS DRIVING SCENARIOS

In this chapter, we will utilize $\mathcal{IDCEC}_1$ and ShadowAdjudicator to show how automated reasoning of the type we created herein could benefit an autonomous vehicle's decision making system.

6.1 Motivation: Chaotic Roadways

To immediately address those who believe autonomous driving can be achieved by a purely data-driven approach,[66] consider the intersection in Figure 6.1. This is a map of Kelley Square in Worcester, Massachusetts, prior to a redesign in 2020. Navigating the square imposed a high cognitive load on drivers, given the complex series of intersections, one-ways, and highway on-/off-ramps, all of which led to high traffic density throughout the square [79]. However the most significant difficulty of navigating Kelley Square was that there were two intersections (marked by blue stars in Figure 6.1) which had no well-defined right-of-way. For example, at the intersection of Madison, Green, and Harding Streets, none of the three roadways had traffic lights or stop/yield signs indicating the correct flow of traffic. Nor is there a "rule of the road" which dictates how the right-of-way should be determined.[67]

An autonomous driving agent, even with a detailed encoding of the rules of the road and massive amounts of driving data, would simply not be able to safely navigate Kelley Square.[68] This is for the simple reason, introduced above, that there are no rules of the road dictating lawful behavior within certain areas of the square. What is required is *creative* self-driving capabilities [29], development of which are out of scope of the present dissertation, but which would almost certainly be supported by reasoning capabilities presented herein. Namely, the ability to reason about deeply nested beliefs, as well as other Theory-of-Mind modalities, is *crucial*. For example, "I *believe* the driver to my left *believes* that I *intend* to

[66] Colleagues of the author have been advocating for the inclusion of symbolic reasoning in autonomous driving architectures since (at least) 2016. We refer the interested reader to Bringsjord and Sen [29].

[67] For example, if two cars arrive at a four-way stop at the same time, the car to the left must yield to the car on the right. (See Figure 6.4.)

[68] Assuming examples of navigating the square were not in the agent's training data. However, simply adding such data is not a sufficient solution (at least in the author's opinion), as it does nothing to enable the agent to solve different driving conundrums which were not seen in the training data.

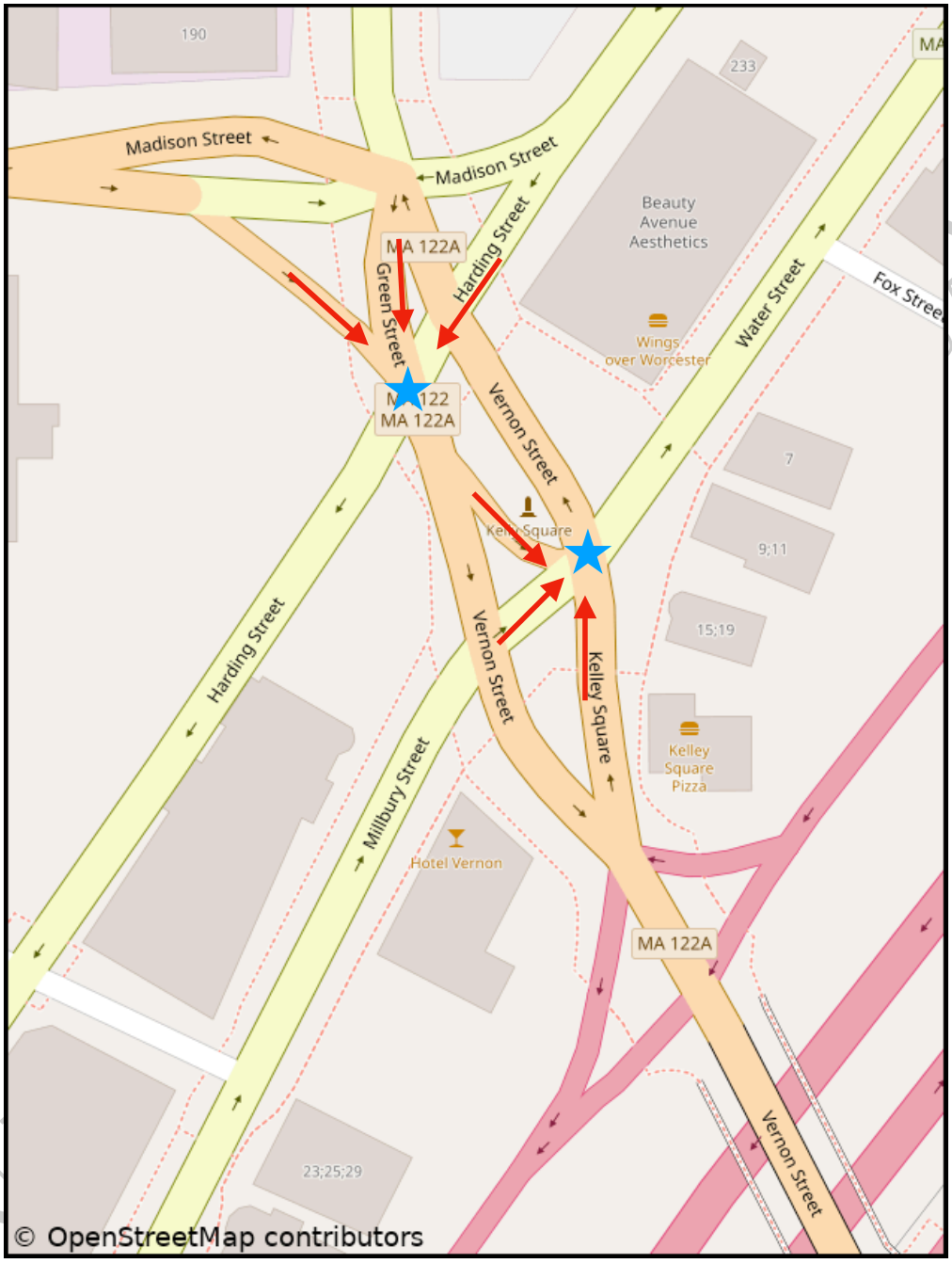

Figure 6.1: A Map of Kelley Square in Worcester, MA, Prior to the 2020 Redesign. Reproduced under the Open Database License from: Wikipedia. 2021. Kelley Square. Retrieved from `https://en.wikipedia.org/wiki/Kelley_Square` (Last Accessed February 14, 2023).

cross the intersection, and hence I *believe* they will wait for me." Furthermore, Cognitive Likelihood would enable e.g. an agent to set likelihood thresholds to ensure safe travel; for example, "While in Kelley Square, I won't cross an intersection unless I believe it is OVERWHELMINGLY LIKELY that all of the other drivers at the intersection *intend* to allow me to cross."

6.2 Scenarios

We now turn to three relatively simple driving scenarios which, despite their simplicity, require complex Theory-of-Mind reasoning. We will demonstrate how reasoning with $\mathcal{IDCEC}_1$ would enable an autonomous driving agent to safely and reliably navigate the scenarios.

For readability, in the following subsections, we present hand-crafted arguments. For the automatically-generated arguments found by ShadowAdjudicator, see Appendix A.

6.2.1 Adjudicating Scenarios With Only Illegal Options

At first glance, the possibility that an autonomous driving agent may have to, or even be able to break the law, is likely to raise concern. However, autonomous agents will undoubtedly face scenarios where ethical and legal obligations contradict; Sen et al. [107] presented a scenario where this was the case and demonstrated how cognitive-calculus-based AI could enable agents to reliably adjudicate these scenarios. Their work, however, used a purely deductive cognitive calculus.

Consider the following scenario, presented pictorally in Figure 6.2. At time t_0, an autonomous driving agent a is stopped at a red light. At time t_1, an ambulance a^* drives up behind a with its siren and lights on. The agent's only way to get out of the way is to drive through the red light, which is illegal. But failing to pull over for an emergency vehicle with its siren on is also illegal (never mind the ethical concerns, as there is presumably a person inside the ambulance who urgently requires medical attention).

Equipped with ShadowAdjudicator, a is able to find a solution. First, a intends to not cross the intersection because of the red light.

$$\mathbf{P}(a, t_0, \text{Red}(\textit{light})) \qquad \text{[Given]} \qquad (6.1)$$

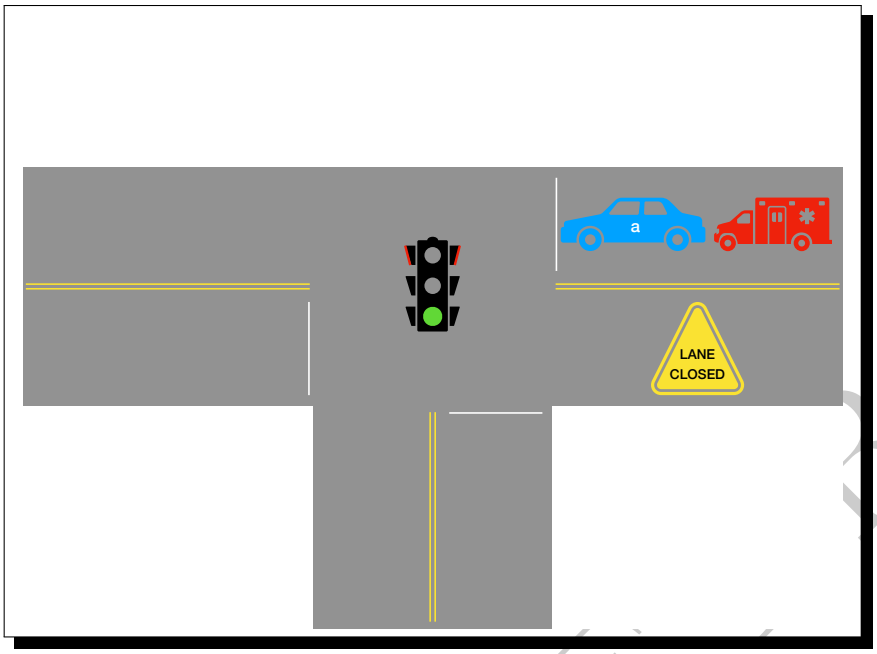

Figure 6.2: Driving Scenario with only Illegal Actions

$$\mathbf{K}(a, t_0, \mathbf{O}(a, t_0, \text{Red}(\textit{light}), \neg \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_1))) \quad \text{[Given]} \quad (6.2)$$

$$\therefore \mathbf{B}^4(a, t_0, \text{Red}(\textit{light})) \quad \text{[from (6.1) via } I_4^{\ell}\text{]} \quad (6.3)$$

$$\therefore \mathbf{B}(a, t_0, \mathbf{O}(a, t_0, \text{Red}(\textit{light}), \neg \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_1))) \quad \text{[from (6.2) via } I_2\text{]} \quad (6.4)$$

$$\therefore \mathbf{O}(a, t_0, \text{Red}(\textit{light}), \neg \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_1)) \quad \text{[from (6.2) via } I_4\text{]} \quad (6.5)$$

$$\therefore \mathbf{K}(a, t_0, \mathbf{I}(a, t_0, \neg \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_1))) \quad \text{[from } \{(6.3), (6.4), (6.5)\} \text{ via } I_{14}\text{]} \quad (6.6)$$

$$\therefore \mathbf{I}(a, t_0, \neg \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_1)) \quad \text{[from (6.6) via } I_4\text{]} \quad (6.7)$$

However, once the ambulance appears, a generates a new argument which concludes that a intends to cross the intersection.

$$\mathbf{S}(a^*, a, t_1, \text{PullOver}(t_2)) \quad \text{[Given]} \quad (6.8)$$

$$\mathbf{K}(a, t_1, \text{Authoritative}(a^*)) \quad \text{[Given]} \quad (6.9)$$

$$\mathbf{K}(a, t_2, \mathbf{O}(a, t_2, \text{PullOver}(t_2), \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_3))) \quad \text{[Given]} \quad (6.10)$$

$$\therefore \mathbf{B}^5(a, t_2, \text{PullOver}(t_2)) \quad \text{[from } \{(6.8), (6.9)\} \text{ via } I_5^\ell\text{]} \quad (6.11)$$

$$\therefore \mathbf{B}(a, t_2, \mathbf{O}(a, t_2, \text{PullOver}(t_2), \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_3))) \quad \text{[from (6.10) via } I_2\text{]} \quad (6.12)$$

$$\therefore \mathbf{O}(a, t_2, \text{PullOver}(t_2), \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_3))) \quad \text{[from (6.10) via } I_4\text{]} \quad (6.13)$$

$$\therefore \mathbf{K}(a, t_2, \mathbf{I}(a, t_2, \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_3))) \quad \text{[from } \{(6.11), (6.12), (6.13)\} \text{ via } I_{14}\text{]} \quad (6.14)$$

$$\therefore \mathbf{I}(a, t_2, \textit{happens}(\textit{action}(a, \text{cross\_intersection}), t_3))) \quad \text{[from (6.14) via } I_4\text{]} \quad (6.15)$$

How does a decide which argument to follow? The former argument hinges on a belief at level 4, while the latter contains only CERTAIN beliefs (level 5). Hence the latter argument defeats the former, and a crosses the intersection.

6.2.2 Safely Navigating a Lane Closure

In the next scenario, our agent a must safely navigate a lane closure (See Figure 6.3). The lane that a is in is blocked, so a, at the direction of a police officer, must use the other lane, typically for traffic moving in the opposite direction. Another vehicle, b, is approaching from the other direction and appears to be slowing down, but has not yet stopped.

At first this may seem simple — a can go so long as the officer indicates that it is okay using their sign, with SLOW on one side and STOP on the other. However, that only gives *a permission* to enter the other lane; it does not guarantee that it is *safe* (b could continue driving, despite the officer's sign; either intentionally ignoring it or failing to see it). Determining that it is safe requires a to ascribe mental states to b. First, a seeks to determine that it is OVERWHELMINGLY LIKELY that b perceives that the sign says STOP (from b's perspective). a conducts the following reasoning via ShadowAdjudicator:

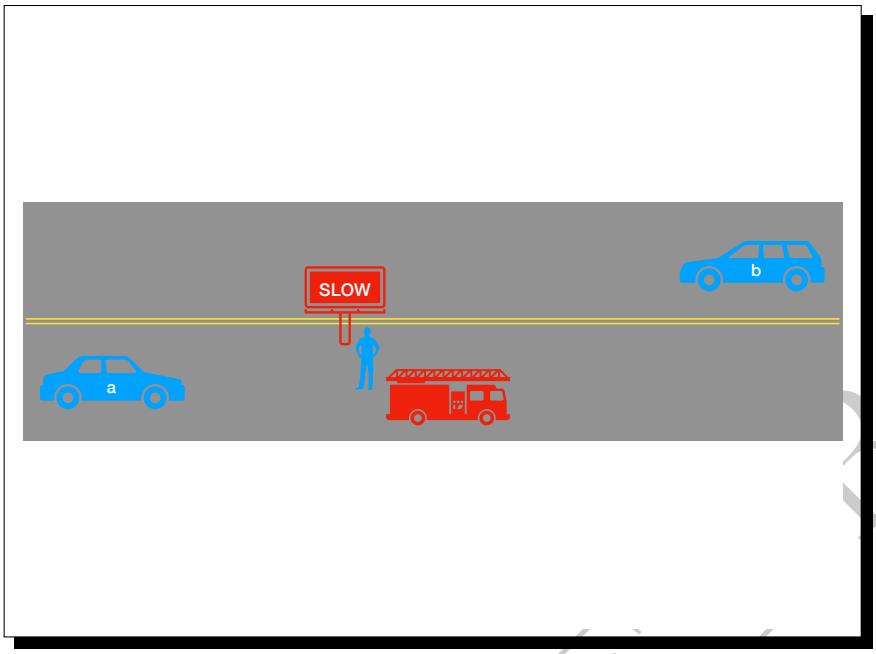

Figure 6.3: Driving Scenario at a Lane Closure

$$\mathbf{P}(a, t_0, \text{Slow}(sign)) \qquad \text{[Given]} \quad (6.16)$$

$$\mathbf{P}(a, t_1, \mathbf{P}(b, t_1, sign)) \qquad \text{[Given]} \quad (6.17)$$

$$\mathbf{B}^3\Big(a, t_1, \big(\mathbf{P}(b, t_1, sign) \wedge \text{Slow}(sign)\big) \rightarrow \mathbf{P}(b, t_1, \text{Stop}(sign))\Big) \qquad \text{[Given]} \quad (6.18)$$

$$\therefore \mathbf{B}^4(a, t_0, \text{Slow}(sign)) \qquad \text{[from (6.16) via } I_4^\ell\text{]} \quad (6.19)$$

$$\therefore \mathbf{B}^4(a, t_1, \mathbf{P}(b, t_1, sign)) \qquad \text{[from (6.17) via } I_4^\ell\text{]} \quad (6.20)$$

$$\therefore \mathbf{B}^3(a, t_1, \mathbf{P}(b, t_1, \text{Stop}(sign))) \qquad \text{[from } \{(6.18), (6.19), (6.20)\} \text{ via } I_{WLP}^\ell\text{]} \quad (6.21)$$

From there, a subsequently reasons that b intends to stop and hence it is safe for a to enter the lane:

$$\mathbf{B}^3(a, t_0, \mathbf{B}^4(b, t_0, \text{Stop}(sign)) \rightarrow \mathbf{I}(b, t_0, \neg happens(action(b, enter\_lane), t_1))) \qquad \text{[Given]} \quad (6.22)$$

$$\mathbf{B}^3(a, t_0,$$

$$\mathbf{I}(b, t_0, \neg happens(action(b, enter\_lane), t_1)) \rightarrow \quad \text{[Given]} \quad (6.23)$$

$$\text{Safe}(happens(action(a, enter\_lane), t_1)))$$

$$\therefore \mathbf{B}^3(a, t_0, \mathbf{B}^4(b, t_0, \text{Stop}(sign))) \quad \text{[from (6.21) via } I^{\ell}_{WLP}\text{]} \quad (6.24)$$

$$\therefore \mathbf{B}^3(a, t_0, \mathbf{I}(b, t_0, \neg happens(action(b, enter\_lane), t_1))) \quad \text{[from } \{(6.24), (6.22)\} \text{ via } I^{\ell}_{WLP}\text{]} \quad (6.25)$$

$$\therefore \mathbf{B}^3(a, t_0, \text{Safe}(happens(action(a, enter\_lane), t_1))) \quad \text{[from } \{(6.23), (6.25)\} \text{ via } I^{\ell}_{WLP}\text{]} \quad (6.26)$$

6.2.3 Understanding Drivers' Intentions at Four-Way Stops

In our final scenario, a must determine the intentions of other drivers at four-way stops. Take a moment to read Figure 6.4. It describes the law in the United States regarding proper behavior at a four-way stop. There are four rules, which are applied sequentially. That is, if the first applies, it holds; else the next rule is evaluated, and so on. The first three rules are straightforward and it is easy to determine if they apply in any particular scenario. The reason, we argue, that four-way stops require Theory-of-Mind reasoning, is the fourth rule: "Even if you have the right-of-way, if for any reason you feel uncomfortable or that your safety is threatened, let the other traffic go ahead. Your safety always comes first. **This trumps all rules.**" (Figure 6.4)

Therefore, while an autonomous-driving agent should attempt to apply the first three rules when relevant, it will also need to be able to mindread other drivers in order to determine their intentions and therefore when they should enter the four-way stop.

Consider the following instantiation, shown in Figure 6.5. Two drivers approach a four-way stop at about the same time. They are across from each other, each intending to turn left. The intersection is too small for them to turn simultaneously, so one must go before the other. None of the first three rules apply, so the drivers must analyze each others intentions. In practice, there are several actions which indicate that a driver either intends to go first or allow the other driver to go first. If one sees a driver inching forward, they likely intend to go first. Alternately, if a driver flashes their lights at the other driver, they likely intend to allow them to go first.[69] We can formalize each of these principles with the following formulae:

[69]Of course, there are many potential intentions a driver could have when flashing their lights, which most experienced drivers can immediately identify. [113]

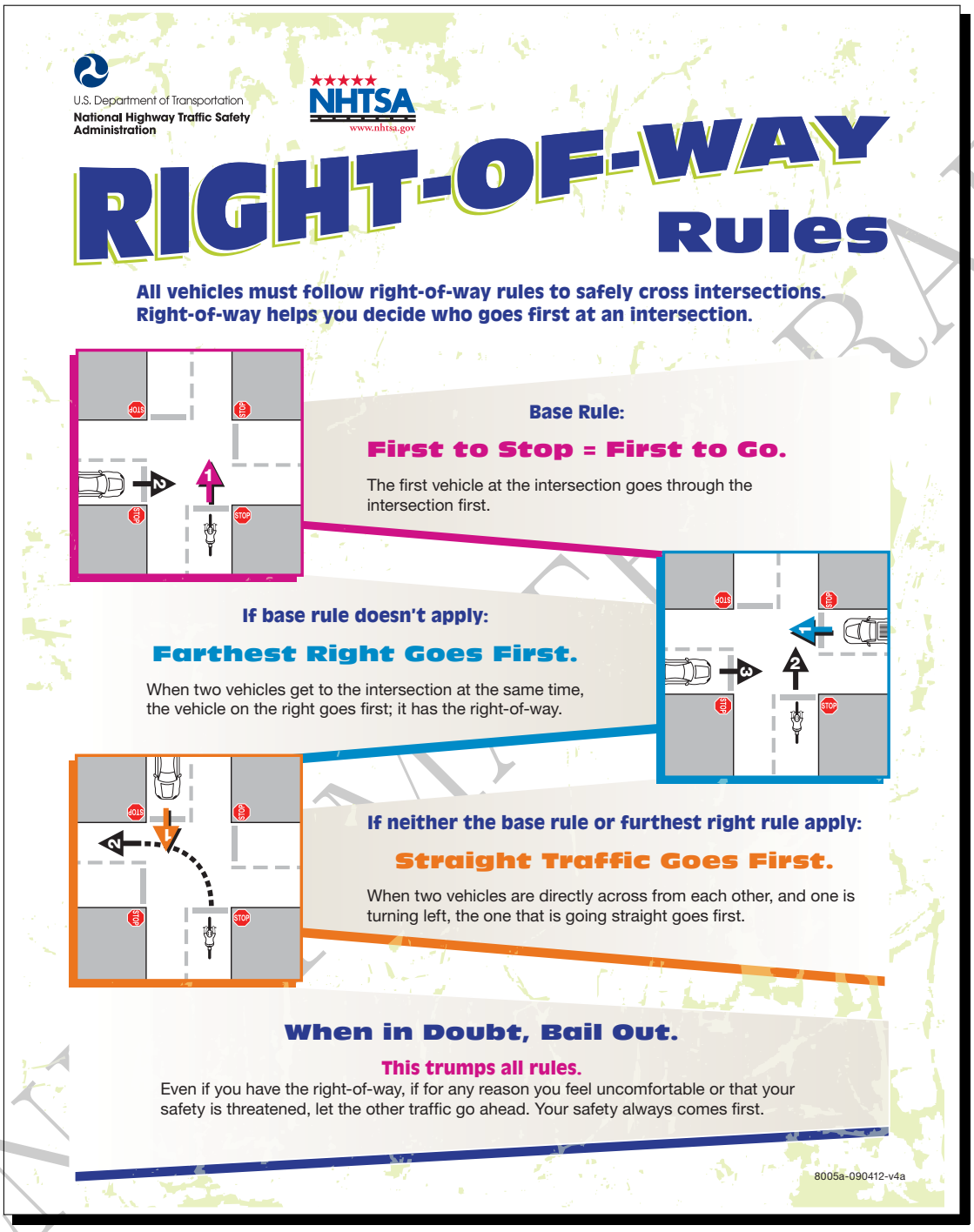

Figure 6.4: Right-of-Way Rules at a Four-Way Intersection. Reproduced (permission not needed) from: U.S. Department of Transportation National Highway Traffic Safety Administration. 2016. Right-of-Way Rules. Retrieved from `https://www.nhtsa.gov/sites/nhtsa.gov/files/rightofwayrules.pdf` (Last Accessed February 14, 2023).

$$\mathbf{B}^3(a, t_0, \text{InMotion}(b, t_0) \rightarrow \mathbf{I}(b, t_0, \text{GoFirst}(b))) \quad \text{[Given]} \quad (6.27)$$

$$\mathbf{B}^3(a, t_0, \text{FlashesLights}(b, t_0) \rightarrow \mathbf{I}(b, t_0, \text{GoFirst}(a))) \quad \text{[Given]} \quad (6.28)$$

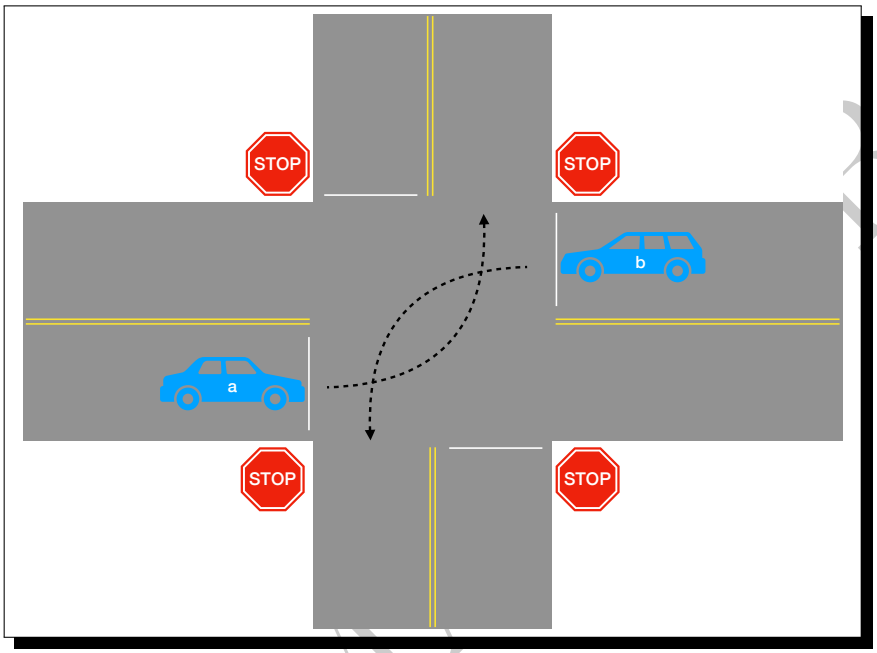

Figure 6.5: Driving Scenario at a Four-Way Intersection

From there, ShadowAdjudicator can infer a belief in a driver's intention based on what it perceives. First, consider the case where driver a perceives the other driver b inching forward. ShadowAdjudicator can generate the following argument which concludes that a believes it is OVERWHELMINGLY LIKELY that b intends to go first:

$$\mathbf{P}(a, t_0, \text{InMotion}(b, t_0)) \quad \text{[Given]} \quad (6.29)$$

$$\therefore \mathbf{B}^4(a, t_0, \text{InMotion}(b, t_0)) \quad \text{[from (6.29) via } I_4^\ell\text{]} \quad (6.30)$$

$$\therefore \mathbf{B}^3(a, t_0, \mathbf{I}(b, t_0, \text{GoFirst}(b))) \quad \text{[from } \{(6.27), (6.30)\} \text{ via } I_{WLP}^\ell\text{]} \quad (6.31)$$

Similarly, in the case where driver a perceives b flash their lights, ShadowAdjudicator can generate the following argument which concludes that a believes it is OVERWHELMINGLY LIKELY that b intends to allow a to go first:

$$\mathbf{P}(a, t_0, \text{FlashesLights}(b, t_0)) \qquad \text{[Given]} \qquad (6.32)$$

$$\therefore \mathbf{B}^4(a, t_0, \text{FlashesLights}(b, t_0)) \qquad \text{[from (6.32) via } I_4^\ell\text{]} \qquad (6.33)$$

$$\therefore \mathbf{B}^3(a, t_0, \mathbf{I}(b, t_0, \text{GoFirst}(a))) \qquad \text{[from } \{(6.28), (6.33)\} \text{ via } I_{WLP}^\ell\text{]} \qquad (6.34)$$

CHAPTER 7
UNIFYING QUALITATIVE & QUANTITATIVE UNCERTAINTY

While we propose the theory of Cognitive Likelihood as a means of reasoning about *qualitative* uncertainty, we acknowledge the need for a capacity for reasoning with *quantitative* uncertainty as well. Moreover, there are undoubtedly scenarios in which agents must reason with *both* qualitative and quantitative forms of uncertainty.

In this chapter, we will take a first step towards an agent formalism for reasoning in this way. We will next introduce a bit of notation for formulae with likelihoods and probabilities, then demonstrate our new formalism using the (in-)famous Monty Hall Problem.

7.1 Notation

Going forward, we will allow an agent's belief to be annotated with a likelihood, a probability, both, or neither. Probabilities are introduced using a simple *Odds* predicate:

$$\text{Odds}\big(\phi, E_h, E_{\neg h}\big) \tag{7.1}$$

where E_h is the set of events in which ϕ holds and $E_{\neg h}$ is the set of events in which ϕ does not hold. Naturally, we require that $E_h \cap E_{\neg h} \neq \emptyset$. Let $i = |E_h|$ and $j = |E_{\neg h}|$. Then the above formula states that the odds of ϕ are i in j, or equivalently that the probability of ϕ is $\frac{i}{i+j}$. Now, we can define a compact representation of a belief with both likelihood and probability:

$$\mathbf{B}^{\sigma}_{\frac{i}{i+j}}(a, t, \phi) := \mathbf{B}^{\sigma}\Big(a, t, \text{Odds}\big(\phi, E_h, E_{\neg h}\big)\Big) \tag{7.2}$$

This newly defined form can be read simply as "Agent a believes, at likelihood σ, that formula ϕ holds with probability $\frac{i}{i+j}$."

We also introduce three inference schemas for reasoning with beliefs with likelihood and probability. The first implements the Addition Law of Probability nested within an agent's belief. To keep the notation of this schema simple, we implicitly assume that the events are mutually exclusive. In §7.2, we will give a generalization of the schema which makes this assumption explicit.

$$\frac{\mathbf{B}^{\sigma_1}_{p_1}(a,t,\phi_1),\quad \mathbf{B}^{\sigma_2}_{p_2}(a,t,\phi_2)}{\mathbf{B}^{\min(\sigma_1,\sigma_2)}_{p_1+p_2}(a,t,\phi_1 \vee \phi_2)}\ [I^p_{+1}] \tag{7.3}$$

This schema can be read as "If agent a believes, at likelihood σ_1, that ϕ_1 holds with probability p_1, and believes, at likelihood σ_2, that ϕ_2 holds with probability p_2, then a can infer a belief at the lower likelihood that $\phi_1 \vee \phi_2$ holds with probability $p_1 + p_2$."

Likewise, if an agent knows the probability of the disjunction of two mutually exclusive events, and knows the probability of one of those events, the agent can infer a belief in the other event's probability (we will refer to this as the *Subtraction Law of Probability*):

$$\frac{\mathbf{B}^{\sigma_1}_{p_1}(a,t,\phi_1 \vee \phi_2),\quad \mathbf{B}^{\sigma_2}_{p_2}(a,t,\phi_2)}{\mathbf{B}^{\min(\sigma_1,\sigma_2)}_{p_1-p_2}(a,t,\phi_1)}\ [I^p_{+2}] \tag{7.4}$$

The third new schema enables agents to reason with a probabilistic form of Disjunctive Syllogism. If an agent believes an arbitrarily-sized (finite) disjunction and believes the negation of one of the disjuncts (with probability 1), they can infer a belief in the remaining disjuncts:

$$\frac{\mathbf{B}^{\sigma_1}_{p_1}(a,t,\phi_1 \vee \cdots \vee \phi_i \vee \cdots \vee \phi_n),\quad \mathbf{B}^{\sigma_2}_{1}(a,t,\neg\phi_i)}{\mathbf{B}^{\min(\sigma_1,\sigma_2)}_{p_1}(a,t,\phi_1 \vee \ldots \phi_{i-1} \vee \phi_{i+1} \vee \cdots \vee \phi_n)}\ [I^p_{DS}] \tag{7.5}$$

7.2 Kolmogorov's Axioms

The fundamental nature of probability theory is captured by a set of axioms known as *Kolmogorov's Axioms* [68]. They describe basic properties necessary for rational reasoning about probability. The axioms are:

1. All probability values are non-negative real numbers.
2. The probability that at least one event in the space of events will occur is 1.
3. The probability of a (countable) disjunction of mutually exclusive events is the sum of the probabilities of all of the individual events.

We will briefly show that the Kolmogorov Axioms hold in the formalism we previously introduced, hence it is an adequate basis for reasoning about probability.

First, probability values must be non-negative by construction, since we generate probabilities from odds. In our formalism we are restricted to rational-number probabilities, but of course all rational numbers are real numbers, so Axiom 1 is satisfied.

Next, we can show that the second axiom holds through a short (informal) proof. The odds that at least one of the events in the space of events (of arbitrary size n) will occur is represented by the following formula:

$$\text{Odds}\big(e_1 \vee \cdots \vee e_n, \{e_1, \ldots, e_n\}, \emptyset\big) \tag{7.6}$$

Therefore the probability that at least one of the events will hold is $\frac{n}{n+0} = 1$.

Finally, we can capture the third axiom in an inference schema. It is a generalization of schema $[I^p_{+1}]$ presented above.

$$\frac{\mathbf{B}^{\sigma_1}(a, t, \text{Odds}(\phi_1, E^1_h, E^1_{\neg h})),\ \ldots,\ \mathbf{B}^{\sigma_n}(a, t, \text{Odds}(\phi_n, E^n_h, E^n_{\neg h})),\ E^1_h \cap \cdots \cap E^n_h = \emptyset}{\mathbf{B}^{\min(\sigma_1,\ldots,\sigma_n)}_{p_1+\cdots+p_n}(a, t, \phi_1 \vee \cdots \vee \phi_n)}\ [I^p_{KA3}] \tag{7.7}$$

where $p_i = \frac{|E^i_h|}{|E^i_h|+|E^i_{\neg h}|}$.

7.3 The Monty Hall Problem

The Monty Hall Problem (MHP), originally attributed to Steve Selvin, was popularized when a reader of Marilyn vos Savant's "Ask Marilyn" column in Parade magazine wrote in regarding the problem. The problem, as stated in her column, is as follows:

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors? [119]

Reasoning via probability theory, one can generate a proof that it is to your advantage to switch. Regardless, even after a proof had been supplied, many people — including the esteemed mathematician Paul Erdős [117] — refused to accept the correct answer. While probability theory can generate a solution to the problem itself, it cannot explain why so

many people, including many supposed experts of mathematics, were so certain that it was wrong.

By reasoning with likelihoods *and* probabilities within $\mathcal{IDCEC}_1$, we will construct two arguments: one in defense of switching, and one in defense of staying. Thereby we will show that the argument for switching has a higher likelihood, and hence should be accepted.

7.4 Modeling Valid Reasoning in MHP

We begin with the host h explaining the rules of the game to contestant c_1. Given that this is in the context of a game show in which the host has control of the game, c_1 knows that h is authoritative and hence can infer CERTAIN beliefs in what h says. First, the host states that the car has been randomly placed behind one of the three doors, hence there is a one-in-three chance of it being behind each door. Note that we will use the predicate "CB" as an abbreviation for "CarBehind".

$$\mathbf{S}\Big(h, c_1, t_1, \text{Odds}\big(\text{CB}(d_1), \{\text{CB}(d_1)\}, \{\text{CB}(d_2), \text{CB}(d_3)\}\big)\Big) \tag{7.8}$$

$$\mathbf{S}\Big(h, c_1, t_1, \text{Odds}\big(\text{CB}(d_2), \{\text{CB}(d_2)\}, \{\text{CB}(d_1), \text{CB}(d_3)\}\big)\Big) \tag{7.9}$$

$$\mathbf{S}\Big(h, c_1, t_1, \text{Odds}\big(\text{CB}(d_3), \{\text{CB}(d_3)\}, \{\text{CB}(d_1), \text{CB}(d_2)\}\big)\Big) \tag{7.10}$$

$$\mathbf{K}(c_1, t_1, \text{Authoritative}(h)) \tag{7.11}$$

$$\therefore \mathbf{B}^5_{1/3}\big(c_1, t_2, \text{CB}(d_1)\big) \quad [\text{from } \{(7.8), (7.11)\} \text{ via } [I^{\ell}_5]] \tag{7.12}$$

$$\therefore \mathbf{B}^5_{1/3}\big(c_1, t_2, \text{CB}(d_2)\big) \quad [\text{from } \{(7.9), (7.11)\} \text{ via } [I^{\ell}_5]] \tag{7.13}$$

$$\therefore \mathbf{B}^5_{1/3}\big(c_1, t_2, \text{CB}(d_3)\big) \quad [\text{from } \{(7.10), (7.11)\} \text{ via } [I^{\ell}_5]] \tag{7.14}$$

The contestant notes that, via inference by $[I^{p}_{+1}]$, that this means that there is a 100% chance that the car is behind one of the three doors. (While this may seem like an obvious, vapid inference, it will serve as an important step in a subsequent inference.)

$$\therefore \mathbf{B}^5_1\big(c_1, t_2, \text{CB}(d_1) \vee \text{CB}(d_2) \vee \text{CB}(d_3)\big) \quad [\text{from } \{(7.12), (7.13), (7.14)\} \text{ via } [I^{p}_{+1}]] \tag{7.15}$$

Since the contestant currently believes it is equally likely that the car is behind any of the three doors, they randomly select d_1. The host then reveals that there is a goat behind d_3, hence the contestant believes it is EVIDENT that the car is not behind d_3:

$$\mathbf{P}(c_1, t_3, \neg\mathrm{CB}(d_3)) \tag{7.16}$$

$$\therefore \mathbf{B}^4(c_1, t_3, \neg\mathrm{CB}(d_3)) \qquad [\text{from (7.16) via } [I_4^{\ell}]] \tag{7.17}$$

The contestant is now at a crucial juncture, that at which most who cognize about this problem go wrong. How is our contestant to update (7.12) at t_3 in light of this new information? The (highly controversial) fact is, c_1 *cannot* infer the negation of this belief, nor increase or decrease the likelihood or probability of the belief, from the information it has just gained. Why? Since there are two doors with goats behind them and the contestant only selects one, there is guaranteed to be a door which Monty can open which will have a goat behind it. Therefore, Monty can and will do this regardless of the contestant's initial choice, so the contestant has gained no new knowledge from Monty's action of opening a door.[70] Hence c_1, if rational, *must* propagate this belief forward to t_3:

$$\therefore \mathbf{B}^5_{1/3}(c_1, t_3, \mathrm{CB}(d_1)) \qquad [\text{from (7.12) via } [I^{\ell}_{PROP}]] \tag{7.18}$$

Next, by probabilistic disjunctive syllogism, the contestant infers the following:

$$\therefore \mathbf{B}^4_1\big(c_1, t_3, \mathrm{CB}(d_1) \vee \mathrm{CB}(d_2)\big) \qquad [\text{from } \{(7.15), (7.17)\} \text{ via } [I^{p}_{DS}]] \tag{7.19}$$

Finally, using the Subtraction Law of Probability, the contestant determines:

$$\therefore \mathbf{B}^4_{2/3}(c_1, t_3, \mathrm{CB}(d_2)) \qquad [\text{from } \{(7.18), (7.19)\} \text{ via } [I^{p}_{+2}]] \tag{7.20}$$

Therefore, the contestant believes it is CERTAIN that the probability that the car is behind door 1 is $\frac{1}{3}$, and it is EVIDENT that the probability that the car is behind door 2 is $\frac{2}{3}$. Hence the contestant believes it is best to switch from door 1 to door 2.

[70] We refer any readers who are still skeptical to the "Three-Shell Game" of Gardner [47] (vos Savant [119] appealed to an argument via analogy to this game as well). The puzzle is logically/probabilistically equivalent to the Monty Hall Problem, but its presentation may elucidate this tricky step.

7.5 Modeling Invalid Reasoning in MHP

Our second contestant c_2 is not as familiar with probability theory as c_1, hence they don't make inferences via the three inference schemata introduced in §7.1. The setup is the same as before; c_2 believes the parameters of the game are certainly true:

$$\mathbf{B}^5_{1/3}\big(c_2, t_1, \mathrm{CB}(d_1)\big) \tag{7.21}$$

$$\mathbf{B}^5_{1/3}\big(c_2, t_1, \mathrm{CB}(d_2)\big) \tag{7.22}$$

$$\mathbf{B}^5_{1/3}\big(c_2, t_1, \mathrm{CB}(d_3)\big) \tag{7.23}$$

Like c_1, as c_2 currently believes it is equally likely that the car is behind any of the three doors, they randomly select d_1. The host then reveals that there is a goat behind d_3, hence the contestant believes it is EVIDENT that the car is not behind d_3:

$$\mathbf{P}(c_2, t_2, \neg\mathrm{CB}(d_3)) \tag{7.24}$$

$$\therefore \mathbf{B}^4(c_2, t_2, \neg\mathrm{CB}(d_3)) \qquad [\text{from (7.24) via } [I^\ell_4]] \tag{7.25}$$

After considering this new information, c_2 decides that it implies that there is a 50/50 chance that the car is behind either door 1 or door 2. While this belief is not grounded in any valid axioms of probability theory, it is affirmed by many in the discourse of the Monty Hall Problem.[71] On the basis of these reports, c_2 can infer a belief at the level of LIKELY via a simple application of modus ponens.

$$\mathbf{B}^4\big(c_2, t_2, \neg\mathrm{CB}(d_3)\big) \to \big(\mathbf{B}^2_{1/2}(c_2, t_2, \mathrm{CB}(d_1)) \wedge \mathbf{B}^2_{1/2}(c_2, t_2, \mathrm{CB}(d_2))\big) \qquad [I^\ell_2] \tag{7.26}$$

[71] Robert Sachs and Scott Smith are two PhD-holders who both fervently argued — with no supporting arguments of any kind — that once the host opens a door, the odds of the remaining two doors becomes 50:50 [119]. Another of vos Savant's critics proposed an argument by analogy to horse racing [92]. Namely, if three horses are racing, and one drops dead, the probability of the other two horses winning increases to 50% each. However, the scenarios are not equivalent: the winner is not predetermined, and the horse's death is completely independent of the choice of winning horse. In MHP, the door that the host chooses to open is *not* independent of where the prize is.

$$\therefore \left(\mathbf{B}^2_{1/2}(c_2, t_2, \mathrm{CB}(d_1)) \wedge \mathbf{B}^2_{1/2}(c_2, t_2, \mathrm{CB}(d_2))\right) \qquad \text{[from } \{(7.25), (7.26)\}\text{]} \qquad (7.27)$$

Therefore c_2 believes it is LIKELY that the two remaining doors have equal probability of having the car. Hence the contestant believes there is no advantage gained by switching doors.

7.6 The Monty Hall Problem, in Light of Our Analysis

Our analysis of the Monty Hall Problem definitively proves that the argument for switching is stronger than the argument for staying. However, if any readers are still not convinced, it also provides a forum for formally discussing the problem. For example, perhaps a skeptical reader believes the likelihoods in the consequent of Equation 7.26 should be higher. However, that is only possible if some axiom of probability theory warrants the inference; the skeptical reader would have to produce one.

The formal discourse hypothesized above is not possible within probability theory. It requires an encompassing framework within which the strength of competing arguments can be expressed, measured, and compared. We have shown that $\mathcal{IDCEC}_1$, with the addition of a small fragment of probability theory, can serve as such a framework. Furthermore, it enables discourse between not just human cognizers, but artificial ones as well.

CHAPTER 8
RELATED WORK

Next we discuss work relevant to the present dissertation. Prior work on cognitive calculi is of course highly relevant, so we discuss that first. Beyond that, there are several fields of research which are relevant: belief revision, nonmonotonic logic/defeasible reasoning, and computational argumentation.

Recall the list of desiderata given in Figure 1.1. In our discussion of various related works, we will note where they meet some of the desiderata and where they fall short.

8.1 Cognitive Calculi

The most directly relevant related work involving cognitive calculi was discussed at length in §3. Here we will give a brief history of cognitive calculi and discuss how they relate to other approaches.

Portions of this chapter previously appeared as:

1. Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2020. Culturally aware social robots that carry humans inside them, protected by defeasible argumentation systems. In *Culturally Sustainable Social Robotics (Proceedings of Robophilosophy 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 335)*, Marco Nørskov, Johanna Seibt, and Oliver Santiago Quick (Eds.). IOS Press, 440–456. https://doi.org/10.3233/FAIA200941

2. Selmer Bringsjord, Naveen Sundar Govindarajulu, and Michael Giancola. 2021. Automated argument adjudication to solve ethical problems in multi-agent environments. In *Paladyn, J. of Behavioral Robotics*, Vol. 12. De Gruyter, Berlin, Boston, 310–335. https://doi.org/10.1515/pjbr-2021-0009

3. Selmer Bringsjord, Naveen Sundar Govindarajulu, John Licato, and Michael Giancola. 2020. Learning ex nihilo. In *Proceedings of the Sixth Global Conference on Artificial Intelligence (GCAI 2020) (EPiC Series in Computing, Vol. 72)*. Gregoire Danoy, Jun Pang, and Geoff Sutcliffe (Eds.). EasyChair, Manchester, UK, 1-27. https://doi.org/10.29007/ggcf

4. Brandon Rozek, Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2022. A framework for testimony-infused automated adjudicative dynamic multi-agent reasoning in ethically charged scenarios. In *Proceedings of the Seventh International Conference on Robot Ethics and Standards (ICRES 2022)*, S. Byun, M.O. Tokhi, M.I.A. Ferreira, N.S. Govindarajulu, M.F. Silva, and K.M. Goher (Eds.). CLAWAR, London, UK, 47–66.

Portions of this chapter are to appear in:

1. Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2023. Logic-based modeling of cognition. In *The Cambridge Handbook of Computational Cognitive Sciences*, Ron Sun (Ed.). Cambridge University Press, Cambridge, UK. Forthcoming.

8.1.1 From $\mathcal{CEC}$, to $\mathcal{DCEC}$, to $\mathcal{IDCEC}$

The first implemented cognitive calculus — later named the Cognitive Event Calculus ($\mathcal{CEC}$) — was published in Arkoudas and Bringsjord [3, 2]. It is a multi-operator modal logic (minus, by definition, and as explained in §2.3, any model-theoretic semantics) based on multi-sorted first-order logic. It had four modal operators: speech, knowledge, belief, and common-knowledge. 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 Arkoudas and Musser [4].

Development of the Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) was first published in Govindarajulu et al. [61]. Therein the formal syntax and inference schemata of the calculus were presented and used to capture and reason about natural-language sentences. Since then, $\mathcal{DCEC}$ has been used to model self-conscious agents [28], capture ethical principles e.g. the Doctrine of Double Effect [58], and model the infinitary false-belief task [26] (among many other applications).

Govindarajulu and Bringsjord [59] was the first to present an inductive cognitive calculus. In §3, we review seven publications since then which further develop the theory of inductive cognitive calculi.

8.1.2 Related Approaches

Belief-Desire-Intention (BDI) logics (e.g. Rao and Georgeff [100]) are related to cognitive calculi, 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 (d_5).

Automated reasoning in the tradition of higher-order logic (HOL) as descended from Frege, and most prominently from Church, which is masterfully chronicled in Benzmüller and Miller [15], is obviously related to cognitive calculi; this is especially true since HOL is now very much on the scene in 21st-century AI (e.g. Benzmüller and Paleo [16]). 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 (d_6). One could envision a cognitive calculus which is based in higher-order logic (as opposed to e.g. FOL or SOL). However this is well beyond the current state-of-the-art.

8.2 Belief Revision

Belief Revision is broadly the set of theories and systems which dictate how beliefs can be expressed, added, and removed. One of the earliest works in the field of Belief Revision are Truth Maintenance Systems (TMS), introduced in Doyle [37].

8.2.1 Truth Maintenance Systems

Truth Maintenance Systems manage and update a set of beliefs for a non-monotonic logical reasoner. Non-monotonicity is essential (d_1), as it allows a TMS to retract beliefs based on new information, which is, by definition, not possible in a monotonic logic.

Doyle [37] calls a belief P "in" the current set of beliefs if "P has at least one currently acceptable reason" and "out" if "P has no currently acceptable reasons (either no reasons at all, or only unacceptable ones)". A reason is an ordered pair of sets of beliefs, and is said to be *acceptable* if its first set of beliefs are all "in" and its second set of beliefs are all "out" [37].
Let

```
P = "There is not a flight from BOS to JFK today",
Q = "The list of flights includes a flight from BOS to JFK".
```

An acceptable reason for belief in P is $\{\{\}, \{Q\}\}$. That is, we justify that there is not a flight from BOS to JFK because there isn't one on the list of flights.[72]

However, say this belief is updated – perhaps the list of flights is updated on the top of the hour, when a flight from BOS to JFK appears. Formally, let

```
R = "Flight 207, which leaves BOS for JFK at 1200 today,
is on the list of flights."
```

If we believe R to be true, $\{\{R\}, \{\}\}$ is an acceptable reason for Q to be "in" the set of beliefs. That is, since we believe a specific flight is on the list, we believe there is a flight on the list in general. However, our reason for believing P is now unacceptable, because it relied on Q being out. Thus, we must remove our belief in P.

Note that the goal of a TMS is to maintain consistency, which means it must work to restore consistency whenever an inconsistency is detected, such as in the example above.

[72] Or, at least, we have no evidence of a flight being on the list (perhaps we don't have access to the list yet).

Therefore it cannot fully satisfy d_2, which requires the ability to tolerate inconsistency when necessary. Also, in a TMS, beliefs in some proposition are always either true or false; while the value can change based on new information, belief cannot be held at a non-certain level of likelihood. Hence d_3 is not satisfied.

We also note that Truth Maintenance Systems were more like a database of beliefs as opposed to an inferential system for reasoning about beliefs. Hence they are not argument-based (d_4) and have no inference schemata (d_5). Likewise they cannot reason over mental states, or even capture mental states beyond belief (d_6) and cannot tackle Turing-unsolvable reasoning problems (d_7).

8.2.2 The AGM Model

The AGM Model is currently the preeminent theory in the field of Belief Revision [63, 1]. It includes three methods for belief change [1]. The first and most simple is called *expansion*, in which a new proposition p is added to the set of beliefs B. The second, *contraction*, removes a proposition p from B. Finally, *revision* is the process of adding a proposition to the belief set, and possibly removing other propositions so that the new set is consistent.

When a set of beliefs is inconsistent, there can be more than one choice of proposition which could be removed to restore consistency. AGM theory employs the concept of *entrenchment*, in which all beliefs are given a value based on their explanatory usefulness. To illustrate, imagine that your beliefs include the laws of thermodynamics and you observe what appears to be a perpetual motion machine. This observation leads you to believe that perpetual motion is possible. However, this contradicts your belief in the first and second laws of thermodynamics. Clearly you should throw away your belief in this supposed perpetual motion machine instead of your beliefs in the laws of thermodynamics, as the latter have much more explanatory usefulness than the former.

For similar reasons, the AGM Model meets the same set of our desiderata as Truth Maintenance Systems. They fully meet d_1, partially meet d_2, and fail to meet d_3 through d_7.

8.2.3 Dempster-Shafer Theory

The theory of belief functions, better known as Dempster-Shafer Theory, captures a measure of uncertainty which can operate under only partial evidence, and can be updated

when more evidence is received [108, 109]. Shafer succinctly describes how this differs from the Bayesian approach (which we discuss further in §8.5.1):

> Whereas the Bayesian probability language uses canonical examples in which known chances are attached directly to the possible answers to the question asked, the language of belief functions uses canonical examples in which known chances may be attached only to the possible answers of a related question. (pg. 7, [109])

That is, a belief function expresses a level of confidence/likelihood in some proposition p based on evidence of some other, related proposition q, which can be updated as we receive more evidence. We call the gap between the evidence we currently have and complete knowledge the "Dempster-Shafer interval".

Dempster-Shafer Theory, like the other works discussed thus far under the umbrella of Belief Revision, does not capture other cognitive modalities other than belief (d_6). It is also not argument-based (d_4) and has no inferential semantics (d_5).

Furthermore, Pearl [90] argues that Dempter-Shafer intervals don't capture the confidence/likelihood that people have towards the probability of some proposition/event. In §7, we showed how our theory of Cognitive Likelihood can be modified to permit reasoning with both likelihood and probability simultaneously.

8.3 Nonmonotonic Logic / Defeasible Reasoning

We reviewed nonmonotonic logic / defeasible reasoning in §2.1. To briefly summarize: a nonmonotonic logic is one which enables defeasible reasoning, which is reasoning which allows for prior formulae to be retracted in response to new knowledge.

Before we discuss specific instances within this field, we note a significant distinction between our work and most (or possibly all) work within nonmonotonic logic / defeasible reasoning. That is, how irresolvable conflicts are handled. All nonmonotonic logics must have some protocol for handling *irresolvable conflicts*: conflicts which cannot be solved via any principle within the theory [115]. Generally one of two approaches is taken: the "skeptical" or the "credulous" approach.

The skeptical approach is to accept none of the conflicting propositions, while the credulous approach is to accept all of them. Unfortunately in many cases neither of these

options produces a satisfying solution, especially if one's aim is to capture how a human cognizer would adjudicate the conflict. Therefore our approach is, whenever possible, to circumvent both of these options by reasoning about the relevant context surrounding the conflict.

Consider a classic problem in nonmonotonic logic / defeasible reasoning: the Nixon Diamond. It consists of two seemingly contradictory arguments regarding former U.S. President Richard Nixon, depicted pictorally in Figure 8.1. First, Nixon is a Quaker, and all Quakers are pacifist; hence Nixon is a pacifist. Second, Nixon is a Republican, and all Republicans are non-pacifists; hence Nixon is a non-pacifist.

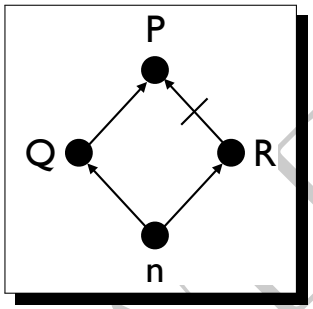

Figure 8.1: The Nixon Diamond. Reproduced under the Creative Commons License Attribution 4.0 International from: Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2022. Novel intensional defeasible reasoning for AI: is it cognitively adequate?. In *Proceedings of the IJCAI Workshop on "Cognitive Aspects of Knowledge Representation" (CAKR 2022)*, Jesse Heyninck, Thomas Meyer, Marco Ragni, Matthias Thimm, and Gabriele Kern-Isberner (Eds.), Vol. 3251. CEUR-WS, Vienna, Austria.

The skeptical approach would be, as always, to conclude nothing: don't infer a belief that Nixon is a pacifist, and don't infer a belief that Nixon is not a pacifist. The credulous approach would be to infer that Nixon is both a pacifist and not a pacifist. Clearly both of these conclusions are unsatisfactory.

In Giancola et al. [51], we show how the Nixon Diamond can be solved using an inductive cognitive calculus. While the interested reader is pointed to Giancola et al. [51] for the full analysis, the main point is to consider the following question: "What would human reasoners familiar with the concepts involved (e.g. pacifism) and background knowledge (e.g. about Quakerism's core tenets) actually conclude about Nixon?" (§1.2, [51]) To summarize the formal argument: there is stronger evidence that Nixon was a Republican than that he

was a Quaker; hence the latter argument (concluding that Nixon is a non-pacifist) should be accepted, and the former rejected.

Finally, we do note that our approach to resolving these sort of conflicts only works when the needed context is available. In the earliest known paper which discusses this problem, the diamond is presented regarding an artificial person named "John" instead of Nixon [103]. In this case, there truly is no possible resolution,[73] in which case $\mathcal{IDCEC}_1$ would follow the skeptical approach of believing nothing. More specifically, one could infer a COUNTERBALANCED belief: $\mathbf{B}^0(a, t, \mathit{Pacifist}(\mathit{john})) \wedge \mathbf{B}^0(a, t, \neg \mathit{Pacifist}(\mathit{john}))$.

8.3.1 Circumscription

Circumscription, as the title of the paper introducing it states, is a form of nonmonotonic reasoning [76]. More specifically, it enables a reasoner to make the assumption that a given set of objects are the only objects that have some property. For example, in the "Missionaries and Cannibals" puzzle (discussed in McCarthy [76]), circumscription enables one to assume that the boat is the only means of crossing the river, since one cannot deduce that there is another way to cross the river from the problem statement. The goal is that this mechanism for generating assumptions is similar to how human cognizers use commonsense reasoning when problem solving. In the example puzzle, it avoids the need to specify that e.g. there is no bridge available, missionaries cannot walk on water, etc., which human cognizers don't need to be told in order to solve the puzzle.

Circumscription is technically a form of nonmonotonic reasoning, not a nonmonotonic logic itself [76]. Hence some desiderata e.g. d_6 cannot be properly evaluated against circumscription, since in theory any logic could be equipped with circumscription. However circumscription is fundamentally model-theoretic and not argument-based, which conflicts with desiderata d_5 and d_4 respectively.

8.3.2 Default Logic

Default Logic formalizes the concept of a *default*: a property that can be assumed to usually hold, but that should not be expected to hold in all cases [102]. The classic example used in default reasoning, which was initially discussed in §2.1, is of a bird called

[73] Hypothetically, since our framework is agent-based the agent could be designed, when it encounters irresolvable conflicts, to generate a goal to investigate the involved entities e.g. "John" in order to gain more information which could potentially resolve the conflict.

Tweety. The property that birds usually fly is best represented by a default: "Birds usually fly; therefore, for all birds b, unless what we currently know about b is inconsistent with the property of flight, we should assume b can fly." Therefore, we should assume Tweety can fly. However, if we later find out that Tweety is a penguin, and we know that penguins cannot fly, we should retract this default assumption in the case of Tweety.

Classic default logic exists purely within a first-order framework, absent any modal operators[74] (d_6) [102]. It does have a proof theory (d_5); however it is based on first-order resolution and is therefore not argument-based (d_4). Interestingly, default logics are the only related work discussed in the present chapter which meets desideratum d_7. The consistency check, necessary to determine if a default has been violated, is in the general case Turing-unsolvable [115].

Finally, more recently Licato [73] 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 work does not satisfy d_3.

8.3.3 OSCAR

One of the major modern contributors to research in argument-based defeasible reasoning is John Pollock, a philosopher who made seminal contributions to AI. Pollock developed a theory of rationality which revolves around the ability to reason defeasibly [93, 95]. He also implemented this theory in an AI agent called "OSCAR."

OSCAR employs first-order inference rules as well as Pollock's methods for defeasible reasoning to solve problems [94, 96]. Input to OSCAR includes a list of givens with corresponding rational-number strength values (not probabilities) and the ultimate epistemic interest of the artificial agent: the formula which OSCAR will try to establish from the givens. The strength of formulas are rational numbers ranging from 0.0 (exclusive) to 1.0 (inclusive), where 1.0 means that the formula is known with absolute certainty to be true.

[74] More recent work in default logic has led to the creation of modal default logics [35]. However they only contain the alethic modalities of necessity and possibility; no cognitive modalities like those discussed herein.

Values less than 1.0 indicate levels of uncertainty in the truth of the statement, and allow such statements to be defeated by arguments which rely solely on statements of higher strength.

While OSCAR is able to satisfactorily model many defeasible reasoning problems, it falls short of meeting our list of desiderata. First, while OSCAR includes a set of deductive inference schemata for first-order logic, it has no inference schemata whatsoever for its *inductive* arguments (d_5). Hence its adjudication of several arguments makes use of no analysis of the internal structure of individual inference steps that human beings routinely engage in (d_4). Such analysis corresponds to abstract treatments of arguments and the suppression of the specifics of individual inferences that are chained together to make an argument; the notion of abstract argumentation, introduced by Dung [38], is discussed further in §8.4.1. Also, as it is limited to first-order logic, OSCAR cannot satisfy d_6 without falling into unsoundness, as shown in Bringsjord and Govindarajulu [23].

For further assessment of OSCAR, in which OSCAR and a cognitive calculus are both utilized toward a solution of the same reasoning problem, the interested reader is pointed to Bringsjord et al. [20]. Likewise, for an overview of the history of the OSCAR project, from its origins to the present day, the interested reader is referred to Bringsjord and Govindarajulu [25].

8.3.4 Computational Paraconsistent Logic

In classical logic, the principle of explosion sanctions the inference of an arbitrary proposition q from a contradiction. Formally, this can be expressed in the following inference schema:

$$\frac{p \wedge \neg p}{q}\ [I_{exp}] \tag{8.1}$$

This schema is typically desired as a useful tool within larger proofs. However in the context of belief revision it can pose a challenge. People often hold sets of beliefs which are inconsistent. We would like to have a logical system which can reason over inconsistent beliefs in a controlled, useful manner – that is, without allowing the agent to infer any arbitrary proposition from contradictory beliefs. This is the goal of paraconsistent logics.

In general, a paraconsistent logic is any logic which doesn't allow this "explosion" of inferring anything from a contradiction [98]. Specifically, paraconsistent logics don't include

the inference schema I_{exp}.

Ávila et al. [7] describes ParaLog_e, "an extension of the ParaLog Logic Programming Language ... that allows direct handling of inconsistency." ParaLog is an extension of Prolog which incorporates paraconsistency [33]. ParaLog_e permits propositions to take on rational-number strength values between 0 and 1 (inclusive) which, like in the case of OSCAR (§8.3.3), are not probabilities, as their values are not restricted to those which would be sanctioned by probability theory (d_3). The semantics of ParaLog_e essentially consist of syntactic transformations and SLD resolution (d_4, d_5).

Villadsen and Schlichtkrull [118] presents a paraconsistent logic and shows how Isabelle — an automated theorem prover for HOL — can be used to generate proofs in this logic. The logic has infinitely many truth values (d_3) and can produce proofs via inference rules (d_4, d_5).

Computational paraconsistent logics clearly meet d_1; they also fully meet d_2 (unlike truth maintenance systems (§8.2.1) and the AGM Model (§8.2.2)) since they are able to tolerate inconsistency. However, to the author's knowledge, there are no computational paraconsistent logics which allow reasoning over mental states beyond belief (d_6) nor any which can handle meta-logical queries about provability (d_7).

Finally, we note that while removing the explosion schema is an interesting and effective maneuver to handle the problem of inconsistency, it is not desirable, because as mentioned earlier, it is a useful tool within larger proofs. We would prefer a logic which contains I_{exp} and handles the issues of inconsistency in another way, as $IDCEC_1$ does.

8.4 Computational Argumentation

The field of computational argumentation aims to produce frameworks and/or systems which enable the generation and comparison of arguments. Generally a computational argumentation framework will have a set of rules or schemes which determine the relative strength of an argument. That strength measure can then be used to adjudicate conflicts between arguments.

8.4.1 Abstract Argumentation

The seminal paper on abstract argumentation is Dung [38]. It introduces the concept of an (abstract) argumentation framework, which is defined as a pair $AF = \{A, R\}$ where A is a

set of arguments and R is a binary relation on A called the "attacks" relation. For example, if $A = \{a, b, c\}$, then $R = \{(a, b), (c, a)\}$ indicates that argument a attacks argument b and that c attacks a.

Our main objection to the field of abstract argumentation is best summarized by Dung himself: "an argument is an abstract entity whose role is solely determined by its relations to other arguments. No special attention is paid to the internal structure of the arguments." (pg. 326, [38]) In desiderata d_4 and d_5, we note that in our conception, arguments *must* have internal inference-to-inference structure, each step of which is sanctioned by an inference schema.

To illustrate the significance of the internal structure of arguments, we turn again to the Nixon Diamond, which we introduced earlier in §8.3. The Nixon Diamond can be represented in a Dung-style argumentation framework where $A = \{a, b\}$, where a is the first argument and b the second. Therefore the attack relation is $R = \{(a, b), (b, a)\}$. However, this is essentially the end of the analysis. Nothing can be definitively concluded about Nixon based on this argumentation framework. This is in accordance with a skeptical approach to handling irresolvable conflicts. However, as we discussed earlier (§8.3), this is clearly unacceptable, and our cognitive-calculus-based approach can produce a rational resolution which aligns with how most human cognizers reason about the problem [51].

Dung [38] has inspired decades of subsequent work on abstract argumentation, all of which fails to meet our desiderata for (at minimum) the same reasons as Dung's original framework [8, 14, 32].

8.4.2 ASPIC$^+$

Modgil and Prakken [82] presents a general framework for structured argumentation based on the Dungian conception of abstract argumentation frameworks. (Hence from the start d_4 and d_5 cannot be satisfied; see §8.4.1.) This framework, ASPIC$^+$, is also 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" (pg. 31, [82]). 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 $\frac{\phi \to \psi, \phi}{\psi}$.

8.4.3 Argument Interchange Format

Rahwan and Reed [99] describes the Argument Interchange Format (AIF). As its name suggests, it is not an argumentation framework itself, but rather a format which an argumentation framework can use to represent arguments — including, unlike Dungian arguments, their internal structure (d_4). The research thrust which developed AIF had two main goals, the first of which was "to facilitate the development of (closed or open) multi-agent systems capable of argumentation-based reasoning and interaction using a shared formalism." (pg. 384, [99])

The format is fairly unrestrictive. There are two types of nodes: information nodes, which contain declarative statements, and scheme nodes, which link information nodes together via inference rules. Or more accurately, a set of information nodes (the premises) infer another set of information nodes (the conclusions) via a scheme node. Rahwan and Reed [99] mentions that non-deductive inference schemata in particular are supported (d_5). Also, while not mentioned explicitly, it seems that nothing in AIF would preclude the use of an expressive formal language which captures intensional operators; likewise for inference schemata which enable automated reasoning over those operators (d_6). Therefore it seems that $\mathcal{IDCEC}_1$-based arguments could be represented in AIF. However this is out of scope of the present dissertation.

8.4.4 Cognitive Argumentation

Cognitive Argumentation formalizes methods of reasoning used by humans (which may or may not be logically sound) as *cognitive principles* [105]. For example, their "Maxim of Quality" expresses that we (humans) typically assume statements we are told are true if we don't have a reason to believe otherwise (e.g. that the speaker may be lying or incompetent). $\mathcal{IDCEC}_1$ schema $[I_5^\ell]$ can be thought of as a formalization of this maxim.

Cognitive Argumentation is mainly a descriptive framework, whereas the focus of Cognitive Likelihood / $\mathcal{IDCEC}_1$ is on normative reasoning. That is, Cognitive Argumentation

describes how humans typically reason; Cognitive Likelihood captures how humans (or any sentient beings) *ought to* reason.[75]

Of all the related work discussed in the present chapter, this approach is, in the author's opinion, likely the most compatible with our own. Intuitively, it seems their cognitive principles could be formalized as inference schemata, and vice versa, inference schemata could be transformed into cognitive principles. However, as the framework is relatively new, it is unclear to what degree it could tackle problems which involve intensionality (d_6) or meta-logical queries about provability (d_7). Recently, Koumi et al. [69] gave an early look at COGNICA, a system for creating and visualizing arguments of Cognitive Argumentation. While it is an impressive tool for creating and adjudicating arguments, intensional operators and meta-logical queries, as of now, are not supported.

8.5 Other Related Work

Finally we mention a few related works which don't fit into one of the three fields of research we've discussed so far.

8.5.1 Bayesian Approaches

There are many methods of uncertainty quantification which we group under the umbrella of *Bayesian Approaches*. These include Bayesian { Epistemology, Networks, Inference, Probability }. We will discuss the first two herein.

Perhaps the most closely related to our work, *Bayesian Epistemology* has its roots in Bayes [11]. It is interested in quantifying the strength of beliefs, using what are called *credences* in the literature. Credences are non-negative real numbers which, like the strength values of OSCAR (§8.3.3), are not probabilities. Bayesian Epistemology is fundamentally based on two norms: *Probabilism* and the *Principle of Conditionalization* [74]. A simple example presented in Lin [74] eloquently displays these norms in action.

Consider an arbitrary hypothesis H which is supported by arbitrary evidence E. There are three possible, mutually exclusive cases: (1) both H and E are true, (2) H is false but E is true, and (3) both H and E are false. Probabilism requires that the credences of these

[75] While cognitive calculi are typically used to capture normative reasoning, they are perfectly capable of modeling descriptive reasoning as well. See Bringsjord et al. [20] for a case study in which both normative and descriptive reasoning are modeled within a cognitive calculus.

three cases follow some of the basic properties of probabilities:[76] they must be non-negative and sum to one. We arbitrarily set $P(H) = \frac{1}{2}$, $P(E \wedge \neg H) = \frac{1}{4}$, and $P(\neg E \wedge \neg H) = \frac{1}{4}$, which is depicted in the bar graph on the left side of Figure 8.2.

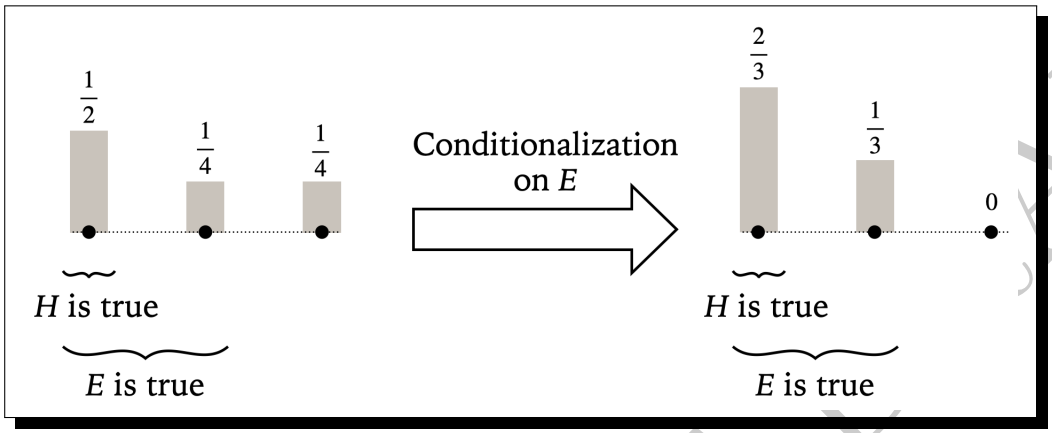

Figure 8.2: The Principle of Conditionalization, in Action. Reproduced (permission not needed) from: Hanti Lin. 2022. Bayesian epistemology. In *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

Next, if we determine that E is true, we can update our credences accordingly using the Principle of Conditionalization, which, like the first norm, operates on credences like probabilities (even though they are not required to be probabilities). Naturally, we set $P(\neg E) = 0$. Then, we normalize the remaining non-zero credences, which is shown in the bar graph on the right side of Figure 8.2.

While not explicitly inference-theoretic one could implement the two norms as inference schemata (d_5). However, Bayesian Epistemology — as well as the rest of the Bayesian Approaches — cannot reason over mental states in order to infer uncertain beliefs from mental states (d_6).

Bayesian Networks operate towards a similar goal: that is, updating the strength of belief in connected propositions based on new evidence. They utilize a directed, acyclic graph (DAG) to represent all of the links between various propositions in an economical fashion [24]. They essentially utilize Bayes' Theorem as the sole inference rule, which updates all of the conditional probabilities in the graph in light of new evidence. While in a loose sense Bayesian networks have inference schemata (d_5), they are certainly not argument-based (d_4), and as mentioned above, have no capacity for reasoning over mental states (d_6).

[76] Even though, as mentioned earlier, credences need not be probabilities.

8.5.2 Modal Probability Logic

There is much prior work on appending modal logics with probability [34]; we will discuss one particular work here. Fagin & Halpern [41, 42] created an epistemic logic which enables agents to express their perceived probability of some proposition. That is, one can express that "according to agent i, formula ϕ holds with probability at least b." (pg. 345, [42])

Their work, while impressive, is limited to a single modality: knowledge (d_6). Our work in §7 enables our cognitive calculus — which of course captures many modalities — to incorporate both probability and likelihood in the reasoning of agents. Furthermore, they utilize possible worlds semantics which are incompatible with the purely inference-theoretic semantics of cognitive calculi (d_5).

8.6 Discussion

Reflecting on all of the work discussed in the present chapter, we note that many approaches are compatible, to varying degrees, with our approach. A cognitive calculus could plausibly be engineered, for example, with inference schemata for default reasoning (§8.3.2) or to capture the methods of belief revision of AGM Theory (§8.2.2). However, as we have detailed throughout this chapter, none of these prior approaches alone could satisfy our list of desiderata $\mathcal{D}$ given in Figure 1.1. The only framework fully satisfying $\mathcal{D}$, to the author's knowledge, is $\mathcal{IDCEC}_1$.

CHAPTER 9
CONCLUSION

> "But what in life is absolutely certain? As Bertrand Russell once wrote, 'All human knowledge is uncertain, inexact, and partial.' Yet somehow we humans manage. When machines can finally do the same, representing and reasoning about that sort of knowledge — uncertain, inexact, and partial — with the fluidity of human beings, the age of flexible and powerful, broad AI will finally be in sight."
>
> —Gary Marcus & Ernest Davis, Rebooting AI

Recall the major contributions of the present dissertation outlined in §1.3, each of which has now been met. First, we formalized a cognitive calculus — $\mathcal{IDCEC}_1$ — which satisfied $\mathcal{D}$ (more on that in §9.1) in §4. Next, we implemented $\mathcal{IDCEC}_1$ in an automated reasoner called *ShadowAdjudicator* (§5) and described the novel algorithm which enables it to find $\mathcal{IDCEC}_1$ proofs (§5.2). Finally, we utilized ShadowAdjudicator to solve reasoning problems in several autonomous driving scenarios (§6) and to adjudicate competing arguments regarding the Monty Hall Problem (§7).

9.1 Desiderata, Met

Recall next the set of desiderata given in Figure 1.1, which $\mathcal{IDCEC}_1$ satisfies. To make this explicit, we will briefly discuss how the various components presented herein met each:

1. Inference schema $[I^{\ell}_{DROP}]$ (defined in Figure 4.2) enables agents to retract prior reasoning when new, contradictory evidence is available.

2. The signature of $\mathcal{IDCEC}_1$ (given in Figure 4.1) enables the representation of beliefs which may contain cognitive inconsistencies. Likewise, the inference schemata of $\mathcal{IDCEC}_1$ enable the resolution of these cognitive inconsistencies.

3. Cognitive Likelihood permits values beyond TRUE, FALSE, and UNKNOWN. Namely, beliefs at eleven likelihood values.

4. $\mathcal{IDCEC}_1$ naturally produces arguments due to its inference-theoretic semantics.

5. The inference schemata of $\mathcal{IDCEC}_1$ are given in Figure 4.2.

6. $\mathcal{IDCEC}_1$ includes modal operators and inference schemata which capture belief with likelihood, as well as many other mental states e.g., knowledge, perception, obligation (the full list of modal operators is given in Table 2.2).

7. Several $\mathcal{IDCEC}_1$ inference schemata require reasoning about Turing-unsolvable problems. For example, $[I^{\ell}_{WLP}]$ (defined in Figure 4.2) requires multiple queries about provability to sanction inference.

9.2 Objections & Rebuttals

Next, we anticipate a trio of objections to our work and present our rebuttals.

9.2.1 The Knowledge Transduction Problem

The first objection can be stated as follows:

> What about the knowledge transduction problem? That is, in order to integrate $\mathcal{IDCEC}$/ShadowAdjudicator into a fully end-to-end AI agent, you would need to generate declarative content for your formalisms from raw input (e.g., video/audio). How would you get such content into your formalisms?

Before addressing the objection directly, it's important to mention why we believe that the existence of this problem shouldn't (and doesn't) preclude our work on reasoning-based AI. To wit, we firmly believe, as discussed in reference to d_4 and d_5 in §1.2, that formal reasoning is *crucial* to the pursuit of artificially-intelligent agents. Specifically, general AI agents will need to be able to explain and verify their decision making in order to be genuinely trustworthy. Hence, we need not wait around for the knowledge transduction problem to be solved before we develop reasoning-based AI technology.

With that said, we admit that this problem has existed since the dawn of AI[77] and has proven to be a tenacious challenge to the integration of reasoning-based technology into

[77]Herbert Simon reported, in reference to his and Newell's creation of Logic Theorist [87], that they "invented a thinking machine." (pg. 206, [112]) However he later explains why they diverted from their original goal — producing a chess-playing machine — to producing an automated theorem prover: "we found this aspect of human mental process (the perceptual) the most difficult to simulate. Hence, we turned to a problem-solving field that is less 'visual' in its content." (pg. 206, [112])

larger AI architectures. While it doesn't appear that a full, robust solution to the problem is on the horizon, progress *has* been made.

In §3.3.1.7, we identified a vision-based landing-site detection system, presented in Shen et al. [111]. We argue their system could plausibly supply our formal-reasoning machinery with the percepts required for the task discussed therein. More generally, significant progress has been made in the fields of computer vision and speech recognition. The YOLO object detection system [101] is a prototypical example of the former, which has steadly improved since its conception in 2016. Examples abound for the latter, including systems which support virtual assistants such as Amazon's Alexa and Apple's Siri. While none of these systems transduce raw input into formal representations, they do synthesize content which could feasibly be transduced into formal representations. We therefore admit that this problem is a roadblock for the integration of our contributions into larger AI architectures; but, we also expect that this problem is surmountable.

For an earlier discussion of this objection and another corresponding rebuttal, the interested reader is directed to §4.3 of Bringsjord et al. [27].

9.2.2 Desideratum d_7

Now, the next objection:

> Desideratum d_7 doesn't seem necessary; couldn't you just enumerate all of the proof rules in place of using a provability query?

Again, before addressing the objection directly, we first acknowledge that there are certainly rational reasons to exclude d_7 from some automated reasoning systems, depending on the goals of the work. Generally speaking, in knowledge representation and reasoning (KR&R), there is a tradeoff between expressivity and tractability. [72] In some sub-areas of KR&R, tractability and decidability are highly desirable; for example, research on description logic historically focused on logics which are tractable and decidable.[78]

[78] An excerpt from *The Description Logic Handbook* explains the significance of decidability to description logic research:

> Because Description Logics are a KR formalism, and since in KR one usually assumes that a KR system should always answer the queries of a user in reasonable time, ... unlike, e.g., first-order theorem provers, these procedures should always terminate, both for positive and for negative answers. Since the guarantee of an answer in finite time need not imply that the answer is given in reasonable time, investigating the computational complexity of a given DL

In our work, tractability is not crucial, but expressivity is. We desire the ability to model agents who can reason about provability queries, and express cognitive attitudes toward them. Moreover, we argue that provability queries are necessary to enable agents to reason about the content of their beliefs.

One could argue against this claim; that instead, one could simply enumerate the inference schemata within the belief operator. So for example, instead of having a single schema which enables the inference of a belief in anything which is provable from the agent's belief set,[79] one could have a schema which expresses "If an agent believes ϕ and believes ψ, then the agent can infer a belief in $\phi \wedge \psi$." Another schema would be needed for e.g., "If an agent believes ϕ, then the agent can infer a belief in $\phi \vee \psi$." and so on for every other inference schema.

We argue that, at best, this approach is cumbersome, as it requires a set of inference schemata where our approach only requires one. Unfortunately, this is only the case when the level of nesting in beliefs is restricted to some finite number,[80] which is not the case in $\mathcal{IDCEC}_1$. If beliefs can be nested arbitrarily deep, it is impossible to enumerate the aforementioned set of inference schemata as the set would have infinite cardinality.

Binas and Ioerger [17] proposes an alternative approach to reasoning about beliefs in first-order logic, which uses decomposition lemmas to express inference rules about arbitrarily nested beliefs.[81] While this approach sidesteps the need for provability queries, it comes at a significant cost. Their representation can only capture conjunctive rules; they mention in §4.2 several formulae to which their decomposition lemmas cannot be applied. For example, they cannot reason with $\mathbf{B}(a, \mathbf{B}(b, p \vee q))$ or $\mathbf{B}(a, \neg\mathbf{B}(b, \mathbf{B}(c, \phi)))$. Therefore, to the authors' knowledge, there is no approach to reasoning with arbitrarily nested beliefs which can handle the entirety of even first-order logic without utilizing provability queries.

9.2.3 Handling Inconsistent Belief Sets in the ShadowAdjudicator Algorithm

The final objection is as follows:

> In the ShadowAdjudicator Algorithm, when you discover that removing likeli-

with decidable inference problems is an important issue. [9]

[79] Schema $[I^{\ell}_{WLP}]$ captures this in $\mathcal{IDCEC}_1$.

[80] A clarifying example: if belief nesting is restricted to $n = 2$ levels, then the formula $\mathbf{B}(a, t, \mathbf{B}(b, t, \phi))$ is valid but $\mathbf{B}(a, t, \mathbf{B}(b, t, \mathbf{B}(c, t, \phi)))$ is not.

[81] See §4.1 of Binas and Ioerger [17].

hood annotations has created an inconsistent belief set, you simply remove all beliefs with likelihood before calling ShadowProver. Isn't this in conflict with desideratum d_3, which requires that cognitive inconsistencies be handled "in a manner which fully permits reasoning to continue"?

Our rebuttal consists of two major arguments. First, we argue that d_3 is in fact satisfied: this component of the algorithm does not stop reasoning from continuing. In the case where we remove all beliefs with likelihood before shadowing, we can still reason deductively about the remaining non-annotated formulae. Furthermore, we can also still apply inductive modal inference schemata to the entire declarative base.

Second, we do admit that utilizing more nuanced techniques when we encounter inconsistent belief sets could enable the generation of proofs that aren't possible using the current algorithm. One potential technique would be to determine the maximally consistent subsets of formulae and shadow each set individually in order to both maintain consistency at the deductive level and also enable reasoning with all formulae. As an example, consider the following declarative base:

$$\Gamma = \begin{cases} \mathbf{B}^1(a,t,\phi) \\ \mathbf{B}^2(a,t,\neg\phi) \\ \mathbf{B}^3(a,t,\psi) \\ \mathbf{B}^4(a,t,\neg\psi) \end{cases} \tag{9.1}$$

The maximally consistent subsets of formulae for Γ are:

$$\Gamma'_1 = \begin{cases} \mathbf{B}(a,t,\phi) \\ \mathbf{B}(a,t,\psi) \end{cases} \quad \Gamma'_2 = \begin{cases} \mathbf{B}(a,t,\neg\phi) \\ \mathbf{B}(a,t,\psi) \end{cases} \quad \Gamma'_3 = \begin{cases} \mathbf{B}(a,t,\phi) \\ \mathbf{B}(a,t,\neg\psi) \end{cases} \quad \Gamma'_4 = \begin{cases} \mathbf{B}(a,t,\neg\phi) \\ \mathbf{B}(a,t,\neg\psi) \end{cases} \tag{9.2}$$

It should be clear that this approach comes at significant costs. Determining the maximally consistent subsets for a large declarative base is hard, and would likely considerably slow the algorithm down. Then, ShadowProver would need to be called for each subset. These calls could be performed in parallel, so the time cost would be negligible but the memory and computational costs would increase.

Finally, we point the interested reader to Fagin and Halpern [40], which discussed a logic which could handle inconsistent beliefs using a notion of local belief. They state that "Our key observation is that one reason that people hold inconsistent beliefs is that beliefs tend to come in non-interacting clusters." [40] Consequently, they allow agents to hold inconsistent beliefs, so long as those beliefs are held in different "frames of mind" which are locally consistent. However, the logics of Fagin and Halpern [40] are based on possible-worlds semantics, which are directly at odds with the purely inference-theoretic semantics of cognitive calculi.

9.3 Future Work

As was mentioned previously, this dissertation is only the first step in the research direction of inductive cognitive calculi with Cognitive-Likelihood-based inference-theoretic semantics (hence the subscript in $\mathcal{IDCEC}_1$). In this final section, we mention a few interesting directions for future work.

9.3.1 Other Graded Modalities

One natural next step of this research direction is to extend the concept of graded modality more broadly to other modalities. For instance, the desire modality **D** could have the gradations given in Table 9.1, which could enable the modeling of agents who must adjudicate competing desires.

Another promising example to which graded modality could be applied is perception, which could indicate the quality of perception. Perception could be compromised by e.g. environmental factors (e.g. fog), faulty hardware, intoxication, etc. [22]. Whereas in this work we assumed perception was perfect and thereby warranted an EVIDENT belief, graded perception would allow for the inferred belief to be assigned a likelihood commensurate with the quality of the percept.

9.3.2 A Cognitive Calculus Dispatcher

Next, we propose that a larger AI architecture (capable of e.g., reasoning, but also perception, action, etc.) should contain a cognitive calculus dispatcher. That is, agents should wield multiple automated reasoners — corresponding to multiple cognitive calculi — and be able to deploy each of them when appropriate. For example, in situations requiring purely

Table 9.1: A 11-Value Spectrum of Desire

Numerical	Linguistic
5	Unstoppable (e.g. severe addiction)
4	Required for Life (e.g. eat when starving, sleep when exhausted)
3	High Priority
2	Medium Priority
1	Low Priority
0	Neutral
-1	Low Priority to Avoid
-2	Medium Priority to Avoid
-3	High Priority to Avoid
-4	Must be Avoided for Life (e.g. being caught by a predator)
-5	Incapable of Being Avoided

deductive reasoning, ShadowProver — and not ShadowAdjudicator — should be dispatched to perform the requisite reasoning. Likewise, if the situation only calls for reasoning at the level of first-order logic, a first-order theorem prover should be deployed. If done well, such a dispatcher could essentially subvert the expressivity vs. tractability tradeoff mentioned in §9.2.2.

This concept of a dispatcher, while of course not fully fleshed out, seems to align well with Kahneman's theories involving System 1 and System 2. [66] System 1, the fast, intuitive component, can oftentimes (though not always) identify when it is insufficient, and must call upon System 2 to perform slow, deliberate reasoning. In this way, we envision that a cognitive calculus dispatcher would call upon a particular cognitive calculus when necessary to perform reasoning tasks.

9.3.3 Abductive Cognitive Calculi

Finally, Cognitive Likelihood could serve as a useful stepping stone toward an *abductive* cognitive calculus. Abductive reasoning is colloquially known as "Inference to the Best Explanation." [36] That is, from a set of believed declarative statements, what hypothesis best explains the statements? Such reasoning is beyond the standard paradigms of deduction and even induction. We believe Cognitive Likelihood could be a useful element of an abductive

cognitive calculus; specifically, to the adjudication of potential abductive hypothesis. That is, using Cognitive Likelihood, one could assign likelihood values to each hypotheses, and subsequently select the hypothesis with the highest likelihood.

REFERENCES

[1] Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: partial meet contraction and revision functions. *The J. of Symbolic Logic* 50, 2 (June 1985), 510–530.

[2] Konstantine Arkoudas and Selmer Bringsjord. 2008. Toward formalizing common-sense psychology: an analysis of the false-belief task. In *PRICAI 2008: Trends in Artificial Intelligence*, Tu-Bao Ho and Zhi-Hua Zhou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 17–29.

[3] Konstantine Arkoudas and Selmer Bringsjord. 2009. Propositional attitudes and causation. *International J. of Software and Informatics* 3, 1 (2009), 47–65.

[4] Konstantine Arkoudas and David Musser. 2017. *Fundamental Proof Methods in Computer Science: A Computer-Based Approach*. MIT Press, Cambridge, MA.

[5] Mark Ashcraft. 1994. *Human Memory and Cognition*. HarperCollins, New York, NY.

[6] Ella Atkins. 2010. Emergency Landing Automation Aids: An Evaluation Inspired by US Airways Flight 1549. In *AIAA Infotech@Aerospace 2010*. American Institute of Aeronautics and Astronautics, Atlanta, Georgia. `https://doi.org/10.2514/6.2010-3381`

[7] Bráulio Coelho Ávila, Jair Minoro Abe, and José Pacheco de Almeida Prado. 1997. ParaLog_e: a paraconsistent evidential logic programming language. In *Proceedings of the Seventeenth International Conference of the Chilean Computer Science Society*. IEEE, 2–8.

[8] Edmond Awad, Richard Booth, Fernando Tohmé, and Iyad Rahwan. 2017. Judgment aggregation in multi-agent argumentation. *J. of Logic and Computation* 27, 1 (2017), 227–259.

[9] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK.

[10] Jonas Bäckstrand and Nicolas Seger. 2016. *Final Report RL 2016:11e*. Technical Report. Swedish Accident Investigation Authority, Stockholm, Sweden.

[11] Thomas Bayes. 1958. An essay towards solving a problem in the doctrine of chances. *Biometrika* 45, 3-4 (December 1958), 296–315. https://doi.org/10.1093/biomet/45.3-4.296

[12] Ensar Becic. 2019. *Vehicle Automation Report HWY18MH010*. Technical Report. National Transportation Safety Board, Washington, D.C.

[13] Hugo Adam Bedau. 1997. *Making Mortal Choices: Three Exercises in Moral Casuistry*. Oxford University Press, Incorporated, New York, NY.

[14] Trevor Bench-Capon. 2003. Persuasion in practical argument using value based argumentation frameworks. *J. of Logic and Computation* 13, 3 (2003), 429–428.

[15] Christoph Benzmüller and Dale Miller. 2014. Automation of Higher-Order Logic. In *Handbook of the History of Logic; Volume 9: Logic and Computation*. North Holland, Amsterdam, The Netherlands.

[16] Christoph Benzmüller and Bruno Woltzenlogel Paleo. 2016. The inconsistency in gödel's ontological argument: a success story for ai in metaphysics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, Subbarao Kambhampati (Ed.). AAAI Press, 936–942.

[17] Arnold Binas and Thomas R. Ioerger. 2004. *Multi-agent belief reasoning in a first-order logic backchainer*. Technical Report TSSTI-TR-10-04. Training System Science and Technology Initiative, Texas A&M University, College Station, TX.

[18] Michael E. Bratman. 1987. *Intention, Plans, and Practical Reason*. Vol. 100. Harvard University Press, Cambridge, MA.

[19] Selmer Bringsjord. 2015. A 21st-Century Ethical Hierarchy for Humans and Robots: $\mathcal{EH}$. In *A World With Robots: International Conference on Robot Ethics (ICRE 2015)*, Isabel Ferreira, Joã Sequeira, M. Tokhi, E. Kadar, and G. Virk (Eds.). Springer, Berlin, Germany, 47–61.

[20] Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2020. Culturally aware social robots that carry humans inside them, protected by defeasible argumentation systems. In *Culturally Sustainable Social Robotics (Proceedings of Robophi-*

losophy 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 335), Marco Nørskov, Johanna Seibt, and Oliver Santiago Quick (Eds.). IOS Press, 440–456. https://doi.org/10.3233/FAIA200941

[21] Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. 2023. Logic-Based Modeling of Cognition. In *The Cambridge Handbook of Computational Cognitive Sciences*, Ron Sun (Ed.). Cambridge University Press, Cambridge, UK. Forthcoming.

[22] Selmer Bringsjord, Michael Giancola, Naveen Sundar Govindarajulu, John Slowik, James Oswald, Paul Bello, and Micah Clark. 2023. Argument-based inductive logics for reasoning under compromised perception. Under review.

[23] Selmer Bringsjord and Naveen Sundar Govindarajulu. 2012. Given the web, what is intelligence, really? *Metaphilosophy* 43, 4 (July 2012), 464–479.

[24] Selmer Bringsjord and Naveen Sundar Govindarajulu. 2022. Bayesian Nets (Supplement to Artificial Intelligence). In *The Stanford Encyclopedia of Philosophy* (fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[25] Selmer Bringsjord and Naveen Sundar Govindarajulu. 2022. The OSCAR Project (Supplement to Artificial Intelligence). In *The Stanford Encyclopedia of Philosophy* (fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[26] Selmer Bringsjord, Naveen Sundar Govindarajulu, and Elmore Christina. 2019. Logicist computational cognitive modeling of infinitary false-belief tasks. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, A. Goel, C. Seifert, and C. Freksa (Eds.). Cognitive Science Society, Montreal, QB, 43–45.

[27] Selmer Bringsjord, Naveen Sundar Govindarajulu, and Michael Giancola. 2021. Automated argument adjudication to solve ethical problems in multi-agent environments. *Paladyn, J. of Behavioral Robotics* 12 (July 2021), 310–335. https://doi.org/10.1515/pjbr-2021-0009

[28] Selmer Bringsjord, John Licato, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Atriya Sen. 2015. Real robots that pass tests of self-consciousness. In *Proccedings of the*

24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015). IEEE, New York, NY, 498–504.

[29] Selmer Bringsjord and Atriya Sen. 2016. On creative self-driving cars: hire the computational logicians, fast. *Applied Artificial Intelligence* 30 (2016), 758–786. Issue 8.

[30] Ruth Byrne. 1989. Suppressing valid inferences with conditionals. *J. of Memory and Language* 31 (1989), 61–83.

[31] Roderick Chisholm. 1987. *Theory of Knowledge* (3 ed.). Prentice-Hall, Englewood Cliffs, NJ.

[32] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. 2007. On the merging of dung's argumentation systems. *Artificial Intelligence* 171 (2007), 730–752.

[33] Newton Carneiro Affonso da Costa, José Pacheco de Almeida Prado, Jair Minoro Abe, Bráulio Coelho Ávila, and Márcio Rillo. 1995. ParaLog: um prolog paraconsistente baseado em lógica anotada, (ParaLog: a annotated logic-based paraconsistent prolog). *Institute for Advanced Studies, University of São Paulo, São Paulo, Brazil* (April 1995).

[34] Lorenz Demey, Barteld Kooi, and Joshua Sack. 2019. Logic and Probability. In *The Stanford Encyclopedia of Philosophy* (summer 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[35] Huimin Dong and Yı N Wáng. 2021. A modal logic of defeasible reasoning. In *Proceedings of the First International Workshop on Logics for New-Generation AI*, Beishui Liao, Jieting Luo, and Leendert van der Torre (Eds.), Vol. 1. College Publications, London, UK, 68–80.

[36] Igor Douven. 2021. Abduction. In *The Stanford Encyclopedia of Philosophy* (summer 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[37] Jon Doyle. 1979. A truth maintenance system. *Artificial Intelligence* 12, 3 (November 1979), 231–272.

[38] Phan Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77 (1995), 321–357.

[39] Heinz-Dieter Ebbinghaus, Jörg Flum, and Wolfgang Thomas. 1994. *Mathematical Logic* (2 ed.). Springer-Verlag, New York, NY.

[40] Ronald Fagin and Joseph Y. Halpern. 1987. Belief, awareness, and limited reasoning. *Artificial intelligence* 34, 1 (December 1987), 39–76. https://doi.org/10.1016/0004-3702(87)90003-8

[41] Ronald Fagin and Joseph Y. Halpern. 1988. Reasoning about knowledge and probability. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge (TARK)*, M. Y. Vardi (Ed.). Morgan Kaufmann, Pacific Grove, CA, 277–293.

[42] Ronald Fagin and Joseph Y. Halpern. 1994. Reasoning about knowledge and probability. *J. of the ACM (JACM)* 41, 2 (March 1994), 340–367.

[43] Fred Feldman. 1978. *Introductory Ethics*. Prentice-Hall, Englewood Cliffs, NJ.

[44] Richard E Fikes and Nils J Nilsson. 1971. STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2, 3-4 (September 1971), 189–208.

[45] Melvin Fitting. 2015. Intensional Logic. In *The Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[46] Kenneth D. Forbus. 2019. *Qualitative Representations: How People Reason and Learn About the Continuous World*. MIT Press, Cambridge, MA.

[47] Martin Gardner. 1982. *Aha! Gotcha: Paradoxes to Puzzle and Delight*. W.H. Freeman and Company, New York, NY.

[48] Michael Genesereth and Nils Nilsson. 1987. *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann, Los Altos, CA.

[49] Gerhard Gentzen. 1935. Investigations into Logical Deduction. In *The Collected Papers of Gerhard Gentzen*, M. E. Szabo (Ed.). North-Holland, Amsterdam, The Netherlands, 68–131.

[50] Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2021. A solution to an ethical super dilemma via a relaxation of the doctrine of triple effect. In *Life-world for Artificial and Natural Systems, Proceedings of the Sixth International Conference on Robot Ethics and Standards (ICRES 2021)*, S. Bringsjord, M.O. Tokhi, M.I.A. Ferreira, N.S. Govindarajulu, and M.F. Silva (Eds.). CLAWAR, London, UK, 23–32.

[51] Michael Giancola, Selmer Bringsjord, and Naveen Sundar Govindarajulu. 2022. Novel intensional defeasible reasoning for AI: is it cognitively adequate?. In *Proceedings of the IJCAI Workshop on "Cognitive Aspects of Knowledge Representation" (CAKR 2022)*, Jesse Heyninck, Thomas Meyer, Marco Ragni, Matthias Thimm, and Gabriele Kern-Isberner (Eds.), Vol. 3251. CEUR-WS, Vienna, Austria.

[52] Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and John Licato. 2020. Adjudication of symbolic & connectionist arguments in autonomous driving AI. In *Proceedings of the Sixth Global Conference on Artificial Intelligence (GCAI 2020) (EPiC Series in Computing, Vol. 72)*, Gregoire Danoy, Jun Pang, and Geoff Sutcliffe (Eds.). EasyChair, Manchester, UK, 28–33. https://doi.org/10.29007/k647

[53] Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and Carlos Varela. 2020. Ethical reasoning for autonomous agents under uncertainty. In *Smart Living and Quality Health with Robots, Proceedings of the Fifth International Conference on Robot Ethics and Standards (ICRES 2020)*, M.O. Tokhi, M.I.A. Ferreira, N.S. Govindarajulu, M.F. Silva, E.E. Kadar, J. Wang, and A.P. Kaur (Eds.). CLAWAR, London, UK, 26–41.

[54] Michael Giancola, Selmer Bringsjord, Naveen Sundar Govindarajulu, and Carlos Varela. 2022. Making maximally ethical decisions via cognitive likelihood and formal planning. In *Towards Trustworthy Artificial Intelligent Systems*, Maria Isabel Aldinhas Ferreira and Mohammad Osman Tokhi (Eds.). Springer International Publishing, Cham, Switzerland, 127–142. https://doi.org/10.1007/978-3-031-09823-9_10

[55] E. Bruce Goldstein. 2008. *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (5 ed.). Cengage Learning, Boston, MA.

[56] Valentin Goranko and Antje Rumberg. 2022. Temporal Logic. In *The Stanford Encyclopedia of Philosophy* (summer 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[57] Naveen Sundar Govindarajulu. 2017. *Spectra*. https://naveensundarg.github.io/Spectra/ (Last Accessed January 18, 2023).

[58] Naveen Sundar Govindarajulu and Selmer Bringsjord. 2017. On automating the doctrine of double effect. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, C. Sierra (Ed.). International Joint Conferences on Artificial Intelligence, 4722–4730. https://doi.org/10.24963/ijcai.2017/658

[59] Naveen Sundar Govindarajulu and Selmer Bringsjord. 2017. 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)*, V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade, and G. Qi (Eds.). Melbourne, Australia, 34–40.

[60] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, Vincent Conitzer, Gillian Hadfield, and Shannon Vallor (Eds.). ACM, New York, NY, 29–35.

[61] Naveen Sundar Govindarajulu, Selmer Bringsjord, and John Licato. 2013. On deep computational formalization of natural language. In *Proceedings of the Workshop: "Formalizing Mechanisms for Artificial General Intelligence and Cognition" (Formal MAGiC) at Artificial General Intelligence 2013*, Ahmed M.H. Abdel-Fattah & Kai-Uwe Kühnberger (Ed.).

[62] Naveen Sundar Govindarajulu, Selmer Bringsjord, and Matthew Peveler. 2019. 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)*, Martin Suda and Sarah Winkler (Eds.). Electronic Proceedings in Theoretical Computer Science, Vol. 311. Open Publishing Association, Waterloo, Australia, 43–49.

[63] Sven Ove Hansson. 2022. Logic of Belief Revision. In *The Stanford Encyclopedia of Philosophy* (spring 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[64] Deborah A Hersman, Christopehr A Hart, and Robert L Sumwalt. 2010. *Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River*. Accident Report NTSB/AAR-10/03. National Transportation Safety Board (NTSB), Washington, D.C.

[65] Gregory Johnson. 2016. *Argument & Inference: An Introduction to Inductive Logic*. MIT Press, Cambridge, MA.

[66] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY.

[67] Lawrence Kohlberg. 1973. The claim to moral adequacy of a highest stage of moral judgment. *J. of Philosophy* 70, 18 (1973), 630–646.

[68] Andrey Nikolaevich Kolmogorov, Nathan Morrison, and Albert T Bharucha-Reid. 2018. *Foundations of the Theory of Probability* (2 ed.). Dover Publications, Mineola, NY.

[69] Adamos Koumi, Antonis Kakas, and Emmanuelle Dietz. 2022. COGNICA: cognitive argumentation. In *Proceedings of the Ninth International Conference on Computational Models of Argument (COMMA 2022) (Frontiers in Artificial Intelligence and Applications, Vol. 353)*, Francesca Toni, Sylwia Polberg, Richard Booth, Martin Caminada, and Hiroyuki Kido (Eds.). IOS Press, 361–362. https://doi.org/10.3233/FAIA220173

[70] Robert Kowalski and Marek Sergot. 1986. A logic-based calculus of events. *New Generation Computing* 4, 1 (March 1986), 67–95.

[71] Legal Information Institute. 1995. Compliance with ATC Clearances and Instructions. In *Electronic Code of Federal Regulations*. Ithaca, NY.

[72] Hector J. Levesque and Ronald J. Brachman. 1985. A Fundamental Tradeoff in Knowledge Representation and Reasoning (Revised Version). In *Readings in Knowledge Representation*. Morgan Kaufmann, Los Altos, CA, 41–70.

[73] John Licato. 2015. Formalizing deceptive reasoning in breaking bad: default reasoning in a doxastic logic. In *2015 AAAI Fall Symposium Series*.

[74] Hanti Lin. 2022. Bayesian Epistemology. In *The Stanford Encyclopedia of Philosophy* (fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[75] George Luger. 2008. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving* (6 ed.). Pearson, London, UK.

[76] John McCarthy. 1980. Circumscription—A Form of Non-Monotonic Reasoning. *Artificial Intelligence* 13 (1980), 27–39.

[77] Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. 1998. *PDDL – the planning domain definition language*. Technical Report CVC TR-98-003. Yale Center for Computational Vision and Control.

[78] Alison McIntyre. 2014. The Doctrine of Double Effect. In *The Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[79] McMahon Associates, Inc. 2018. *Road Safety Audit: Kelley Square*. Technical Report. Boston, MA.

[80] Paul McNamara. 2010. Deontic Logic. In *The Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[81] Albert Mills, Gabrielle Durepos, and Elden Wiebe. 2010. *Encyclopedia of Case Study Research*. SAGE Publications, Inc., Thousand Oaks, California, Chapter Multiple Sources of Evidence, 585–586. `https://doi.org/10.4135/9781412957397`

[82] Sanjaya Modgil and Henry Prakken. 2014. The ASPIC$^+$ framework for structured argumentation: a tutorial. *Argument & Computation* 5, 1 (2014), 31–62.

[83] Erik Mueller. 2014. *Commonsense Reasoning: An Event Calculus Based Approach*. Morgan Kaufmann, San Francisco, CA.

[84] Franz Carl Müller-Lyer. 1889. Optische urteilstäuschungen. *Archiv für Anatomie und Physiologie, Physiologische Abteilung* 2 (1889), 263–270.

[85] National Transportation Safety Board (NTSB). 2009. Transcript - Public Hearing Day 1 (06/09/09).

[86] Michael Nelson. 2015. Propositional Attitude Reports. In *The Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[87] Allen Newell and Herbert A. Simon. 1956. The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory* 2, 3 (September 1956), 61–79. https://doi.org/10.1109/TIT.1956.1056797

[88] Jeffrey Paris and Alena Vencovská. 2015. *Pure Inductive Logic*. Cambridge University Press, Cambridge, UK.

[89] Saswata Paul, Frederick Hole, Alexandra Zytek, and Carlos A. Varela. 2017. Flight trajectory planning for fixed wing aircraft in loss of thrust emergencies. In *Dynamic Data-Driven Application Systems (DDDAS 2017)*. Cambridge, MA.

[90] Judea Pearl. 1988. On probability intervals. *International J. of Approximate Reasoning* 2, 3 (July 1988), 211–216.

[91] Matthew Peveler, Naveen Sundar Govindarajulu, and Selmer Bringsjord. 2018. Toward Automating the Doctrine of Triple Effect. In *Hybrid Worlds: Societal and Ethical Challenges, Proceedings of the Third International Conference on Robot Ethics and Standards (ICRES 2018)*, Selmer Bringsjord, Mohammad Osman Tokhi, Maria Isabel Aldinhas Ferreira, and Naveen Sundar Govindarajulu (Eds.). CLAWAR, London, UK, 82–88.

[92] Steven Pinker. 2021. *Rationality: What it is, Why it Seems Scarce, Why it Matters*. Penguin Random House, New York, NY.

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

[94] John Pollock. 1992. How to reason defeasibly. *Artificial Intelligence* 57, 1 (September 1992), 1–42.

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

[96] John Pollock. 2001. Defasible reasoning with variable degrees of justification. *Artificial Intelligence* 133 (December 2001), 233–282.

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

[98] Graham Priest, Koji Tanaka, and Zach Weber. 2018. Paraconsistent Logic. In *The Stanford Encyclopedia of Philosophy* (summer 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[99] Iyad Rahwan and Chris Reed. 2009. The Argument Interchange Format. In *Argumentation in Artificial Intelligence*, Guillermo Simari and Iyad Rahwan (Eds.). Springer, Boston, MA, 383–402. https://doi.org/10.1007/978-0-387-98197-0_19

[100] Anand S. Rao and Michael P. Georgeff. 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of Knowledge Representation and Reasoning (KR&R-91)*, R. Fikes and E. Sandewall (Eds.). Morgan Kaufmann, San Mateo, CA, 473–484.

[101] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

[102] Raymond Reiter. 1980. A logic for default reasoning. *Artificial Intelligence* 13 (1980), 81–132.

[103] Raymond Reiter and Giovanni Criscuolo. 1981. On interacting defaults. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI'81)*. 270–276.

[104] Stewart Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach* (4 ed.). Pearson, New York, NY.

[105] Emmanuelle-Anna Dietz Saldanha and Antonis Kakas. 2020. Cognitive argumentation and the suppression task. arXiv:2002.10149. Retrieved from https://arxiv.org/abs/2002.10149 (Last Accessed February 14, 2023).

[106] Peter Schroeder-Heister. 2018. Proof-Theoretic Semantics. In *The Stanford Encyclopedia of Philosophy*, Edward Zalta (Ed.).

[107] Atriya Sen, Paul Mayol, Biplav Srivastava, Kartik Talamadupula, Naveen Sundar Govindarajulu, and Selmer Bringsjord. 2018. For AIs, is it ethically/legally permitted that ethical obligations override legal ones?. In *Hybrid Worlds: Societal and Ethical Challenges, Proceedings of the Third International Conference on Robot Ethics and Standards (ICRES 2018)*, Selmer Bringsjord, Mohammad Osman Tokhi, Maria Isabel Aldinhas Ferreira, and Naveen Sundar Govindarajulu (Eds.). CLAWAR, 26–32.

[108] Glenn Shafer. 1986. Probability Judgment in Artificial Intelligence. In *Uncertainty in Artificial Intelligence*, Laveen N. Kanal and John F. Lemmer (Eds.). Machine Intelligence and Pattern Recognition, Vol. 4. Elsevier, Amsterdam, The Netherlands, 127–135. ht tps://doi.org/10.1016/B978-0-444-70058-2.50014-0

[109] Glenn Shafer. 1987. Probability judgment in artificial intelligence and expert systems. *Statist. Sci.* 2, 1 (February 1987), 3–16. https://doi.org/10.1214/ss/1177013426

[110] Murray Shanahan. 1999. The Event Calculus Explained. In *Artificial Intelligence Today (LNAI 1600)*, M. Wooldridge and M. Veloso (Eds.). Springer, New York, NY, 409–430.

[111] Yu-Fei Shen, Zia-Ur Rahman, Dean Krusienski, and Jiang Li. 2013. A vision-based automatic safe landing-site detection system. *IEEE Trans. Aerospace Electron. Systems* 49, 1 (January 2013), 294–311.

[112] Herbert A. Simon. 1996. *Models of My Life*. MIT Press, Cambridge, MA, Chapter Climbing the Mountain: Artificial Intelligence Achieved, 198–214.

[113] Beth Skwarecki. 2023. Here's what it actually means when someone flashes their headlights at you. *LifeHacker* (February 2023).

[114] J.J.C. Smart and Bernard Williams. 1973. *Utilitarianism: For and Against*. Cambridge University Press, Cambridge, UK.

[115] Christian Strasser and G. Aldo Antonelli. 2019. Non-Monotonic Logic. In *The Stanford Encyclopedia of Philosophy* (summer 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[116] Carlos A. Varela. 2019. Too many airplace systems rely on too few sensors. *The Conversation* (April 2019).

[117] Andrew Vazsonyi. 1999. Which door has the cadillac? *Decision Line* (December/January 1999), 17–19.

[118] Jørgen Villadsen and Anders Schlichtkrull. 2017. Formalizing a Paraconsistent Logic in the Isabelle Proof Assistant. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXIV*, Abdelkader Hameurlain, Josef Küng, Roland Wagner, and Hendrik Decker (Eds.). Springer, Berlin, Heidelberg, 92–122. `https://doi.org/10.1007/978-3-662-55947-5_5`

[119] Marilyn vos Savant. 1990. Game show problem. *Parade* 16 (September 1990).

APPENDIX A
ShadowAdjudicator Output

In this appendix we show the ShadowAdjudicator output corresponding to the arguments presented in §4, 6 and 7.[82] Note that the arguments generated by ShadowAdjudicator may not correspond exactly to those presented in the text. In many cases there are several valid proofs; ShadowAdjudicator searches for an argument and returns the first one it finds.

A.1 Arguments of Chapter 4

```
(base) root@97e884a1add6:/base# python diss_examples/intro_schema_usage_examples.py
PROOF OF: (Believes!0 a t0 (Raining tomorrow))
Applied '[I^\ell_0]' to: []

PROOF OF: (Believes!1 a t2 (Raining tomorrow))
Applied '[I^\ell_1]' to: (Says! m1 a t1 (Raining tomorrow))
  PROOF OF: (Says! m1 a t1 (Raining tomorrow))
  GIVEN

PROOF OF: (Believes!2 a t3 (Raining tomorrow))
Applied '[I^\ell_2]' to: (Says! m1 a t1 (Raining tomorrow)); (Says! m2 a t2 (Raining tomorrow))
  PROOF OF: (Says! m1 a t1 (Raining tomorrow))
  GIVEN
  PROOF OF: (Says! m2 a t2 (Raining tomorrow))
  GIVEN

PROOF OF: (Believes!3 a t5 (Raining now))
Applied '[I^\ell_3]' to: (Perceives! a t4 (Raining now)); (not (Perceives! a t5 (Raining now)))
  PROOF OF: (Perceives! a t4 (Raining now))
  GIVEN
  PROOF OF: (not (Perceives! a t5 (Raining now)))
  GIVEN

PROOF OF: (Believes!4 a t4 (Raining t4))
Applied '[I^\ell_4]' to: (Perceives! a t4 (Raining t4))
  PROOF OF: (Perceives! a t4 (Raining t4))
  GIVEN

PROOF OF: (Believes!5 a t7 (Raining next_scene))
Applied '[I^\ell_5]' to: (Says! d a t6 (Raining next_scene)); (Knows! a t6 (Authoritative d))
  PROOF OF: (Says! d a t6 (Raining next_scene))
  GIVEN
  PROOF OF: (Knows! a t6 (Authoritative d))
  GIVEN
```

Figure A.1: Schema Usage Examples: Introduction Schemata

[82] The following link points to a specific commit of the ShadowAdjudicator codebase. Therefore if you download and run the examples using that version of the code, you should get the same output as is shown here. `https://github.com/RAIRLab/ShadowAdjudicator/tree/06b5a2c3170b5a7a644418b548640d572fb23e24` (Last Accessed February 10, 2023)

```
(base) root@97e884a1add6:/base# python diss_examples/defeasible_belief_generation_examples.py
Calling ShadowProver (goal=(and (Raining tomorrow) (Windy tomorrow)))...
ShadowProver Done.
Calling ShadowProver (goal=(and (Raining tomorrow) (Windy tomorrow)))...
ShadowProver Done.
Calling ShadowProver (goal=(and (Raining tomorrow) (Windy tomorrow)))...
ShadowProver Done.
PROOF OF: (Believes!1 a t5 (and (Raining tomorrow) (Windy tomorrow)))
Applied '[I^\ell_{WLP}]' to: (Believes!1 a t5 (Windy tomorrow)); (Believes!1 a t5 (Raining tomorrow))
  PROOF OF: (Believes!1 a t5 (Windy tomorrow))
  Applied '[I^\ell_{PROP}]' to: (Believes!1 a t4 (Windy tomorrow))
    PROOF OF: (Believes!1 a t4 (Windy tomorrow))
    Applied '[I^\ell_1]' to: (Says! m2 a t3 (Windy tomorrow))
      PROOF OF: (Says! m2 a t3 (Windy tomorrow))
      GIVEN
  PROOF OF: (Believes!1 a t5 (Raining tomorrow))
  Applied '[I^\ell_{PROP}]' to: (Believes!1 a t4 (Raining tomorrow))
    PROOF OF: (Believes!1 a t4 (Raining tomorrow))
    Applied '[I^\ell_{PROP}]' to: (Believes!1 a t3 (Raining tomorrow))
      PROOF OF: (Believes!1 a t3 (Raining tomorrow))
      Applied '[I^\ell_{PROP}]' to: (Believes!1 a t2 (Raining tomorrow))
        PROOF OF: (Believes!1 a t2 (Raining tomorrow))
        Applied '[I^\ell_1]' to: (Says! m1 a t1 (Raining tomorrow))
          PROOF OF: (Says! m1 a t1 (Raining tomorrow))
          GIVEN

PROOF OF: (Believes!1 a t6 (and (Raining tomorrow) (Windy tomorrow)))
Applied '[I^\ell_{PROP}]' to: (Believes!1 a t5 (and (Raining tomorrow) (Windy tomorrow)))
  PROOF OF: (Believes!1 a t5 (and (Raining tomorrow) (Windy tomorrow)))
  GIVEN
Note: The next proof should fail. It serves to verify that the belief propagation schema is correctly
blocking the propagation of beliefs whose negation can be proved.
Calling ShadowProver (goal=phi)...
ShadowProver Done.
Calling ShadowProver (goal=phi)...
ShadowProver Done.
FAILED

PROOF OF: (not (Believes!1 a t1 (Raining t1)))
Applied '[I^\ell_{DROP}]' to: (Believes!1 a t0 (Raining t1)); (Believes!4 a t1 (not (Raining t1)))
  PROOF OF: (Believes!1 a t0 (Raining t1))
  GIVEN
  PROOF OF: (Believes!4 a t1 (not (Raining t1)))
  Applied '[I^\ell_4]' to: (Perceives! a t1 (not (Raining t1)))
    PROOF OF: (Perceives! a t1 (not (Raining t1)))
    GIVEN
```

Figure A.2: Schema Usage Examples: Defeasible Belief Generation

A.2 Arguments of Chapter 6

```
(base) root@97e884a1add6:/base# python diss_examples/driving_examples.py
Calling ShadowProver (inconsistency check)...
ShadowProver Done.
Calling ShadowProver (goal=(Intends! a t0 (not (happens (action a cross_intersection) t1))))...
ShadowProver Done.
PROOF OF: (Intends! a t0 (not (happens (action a cross_intersection) t1)))
Proved via ShadowProver [and a sub-argument of (Perceives! a t0 (Light red)); (Knows! a t0 (Ought! a t0 (Light red) (not (happens (action a cross_int
ersection) t1))))]:
-->(AssumptionsNowContainsGoal
--> Givens:
-->((Intends! a t0 (not (happens (action  a cross_intersection) t1)))
-->
-->((Perceives! a t0 (Light red))
-->    (GIVEN[]))
-->
-->((Believes! a t0 (Light red))
-->    (GIVEN[]))
-->
-->((Knows! a t0 (Light red))
-->    (Perception to knowledge (Perceives! a t0 (Light red))
--> [(GIVEN[])]))
-->
-->((Knows! a t0 (Ought! a t0 (Light red) (not (happens (action  a cross_intersection) t1))))
-->    (GIVEN[]))
-->(Ought! a t0 (Light red) (not (happens (action  a cross_intersection) t1)))
-->
-->((Believes! a t0 (Ought! a t0 (Light red) (not (happens (action  a cross_intersection) t1))))
-->    InferenceJustification{base=[(Knows! a t0 (Ought! a t0 (Light red) (not (happens (action  a cross_intersection) t1))))], message='DR5'})
-->(not (happens (action  a cross_intersection) t1)))
--> Goals:
-->(Intends! a t0 (not (happens (action  a cross_intersection) t1))))
  PROOF OF: (Perceives! a t0 (Light red))
  GIVEN
  PROOF OF: (Knows! a t0 (Ought! a t0 (Light red) (not (happens (action a cross_intersection) t1))))
  GIVEN

Calling ShadowProver (inconsistency check)...
ShadowProver Done.
Calling ShadowProver (goal=(Intends! a t2 (happens (action a cross_intersection) t2)))...
ShadowProver Done.
PROOF OF: (Intends! a t2 (happens (action a cross_intersection) t2))
Proved via ShadowProver [and a sub-argument of (Knows! a t1 (Authoritative a*)); (Knows! a t2 (Ought! a t2 (PullOver t2) (happens (action a cross_int
ersection) t2)))]:
-->(AssumptionsNowContainsGoal
--> Givens:
-->((happens (action  a cross_intersection) t2)
-->
-->((Knows! a t1 (Authoritative a*))
-->    (GIVEN[]))
-->
-->((Authoritative a*)
-->    (GIVEN[]))
-->
-->((Believes! a t1 (Authoritative a*))
-->    InferenceJustification{base=[(Knows! a t1 (Authoritative a*))], message='DR5'})
-->(Intends! a t2 (happens (action  a cross_intersection) t2))
-->
-->((Knows! a t2 (Ought! a t2 (PullOver t2) (happens (action  a cross_intersection) t2)))
-->    (GIVEN[]))
-->(Ought! a t2 (PullOver t2) (happens (action  a cross_intersection) t2))
-->
-->((Believes! a t2 (Ought! a t2 (PullOver t2) (happens (action  a cross_intersection) t2)))
-->    InferenceJustification{base=[(Knows! a t2 (Ought! a t2 (PullOver t2) (happens (action  a cross_intersection) t2)))], message='DR5'})
-->
-->((Says! a* t1 (PullOver t2))
-->    (GIVEN[]))
-->
-->((Believes! a t2 (PullOver t2))
```

Figure A.3: Driving Scenarios (Part 1 of 2)

```
-->    (GIVEN[]))
-->
-->((Believes! a* t1 (PullOver t2))
-->    (Says to belief
--> [(GIVEN[])])))
--> Goals:
-->(Intends! a t2 (happens (action  a cross_intersection) t2)))
  PROOF OF: (Knows! a t1 (Authoritative a*))
  GIVEN
  PROOF OF: (Knows! a t2 (Ought! a t2 (PullOver t2) (happens (action a cross_intersection) t2)))
  GIVEN

Calling ShadowProver (inconsistency check)...
ShadowProver Done.
Calling ShadowProver (goal=(Perceives! b t1 (Stop sign)))...
ShadowProver Done.
PROOF OF: (Believes!3 a t1 (Perceives! b t1 (Stop sign)))
Applied '[I^\ell_{WLP}]' to: (Believes!3 a t1 (implies (and (Perceives! b t1 sign) (Slow sign)) (Perceives! b t1 (Stop sign)))); (Believes!4 a t1 (Pe
rceives! b t1 sign)); (Believes!4 a t1 (Slow sign))
  PROOF OF: (Believes!3 a t1 (implies (and (Perceives! b t1 sign) (Slow sign)) (Perceives! b t1 (Stop sign))))
  GIVEN
  PROOF OF: (Believes!4 a t1 (Perceives! b t1 sign))
  Applied '[I^\ell_4]' to: (Perceives! a t1 (Perceives! b t1 sign))
    PROOF OF: (Perceives! a t1 (Perceives! b t1 sign))
    GIVEN
  PROOF OF: (Believes!4 a t1 (Slow sign))
  Applied '[I^\ell_{PROP}]' to: (Believes!4 a t0 (Slow sign))
    PROOF OF: (Believes!4 a t0 (Slow sign))
    Applied '[I^\ell_4]' to: (Perceives! a t0 (Slow sign))
      PROOF OF: (Perceives! a t0 (Slow sign))
      GIVEN

PROOF OF: (Believes!3 a t0 (Safe (happens (action a enter_lane) t1)))
Applied '[I^\ell_{WLP}]' to: (Believes!3 a t0 (Perceives! b t0 (Stop sign))); (Believes!3 a t0 (implies (Believes!4 b t0 (Stop sign)) (Intends! b t0
(not (happens (action b enter_lane) t1))))); (Believes!3 a t0 (implies (Intends! b t0 (not (happens (action b enter_lane) t1))) (Safe (happens (actio
n a enter_lane) t1))))
  PROOF OF: (Believes!3 a t0 (Perceives! b t0 (Stop sign)))
  GIVEN
  PROOF OF: (Believes!3 a t0 (implies (Believes!4 b t0 (Stop sign)) (Intends! b t0 (not (happens (action b enter_lane) t1)))))
  GIVEN
  PROOF OF: (Believes!3 a t0 (implies (Intends! b t0 (not (happens (action b enter_lane) t1))) (Safe (happens (action a enter_lane) t1))))
  GIVEN

PROOF OF: (Believes!3 a t0 (Intends! b t0 (GoFirst b)))
Applied '[I^\ell_{WLP}]' to: (Believes!3 a t0 (implies (InMotion b t0) (Intends! b t0 (GoFirst b)))); (Believes!4 a t0 (InMotion b t0))
  PROOF OF: (Believes!3 a t0 (implies (InMotion b t0) (Intends! b t0 (GoFirst b))))
  GIVEN
  PROOF OF: (Believes!4 a t0 (InMotion b t0))
  Applied '[I^\ell_4]' to: (Perceives! a t0 (InMotion b t0))
    PROOF OF: (Perceives! a t0 (InMotion b t0))
    GIVEN

PROOF OF: (Believes!3 a t0 (Intends! b t0 (GoFirst a)))
Applied '[I^\ell_{WLP}]' to: (Believes!3 a t0 (implies (FlashesLights b t0) (Intends! b t0 (GoFirst a)))); (Believes!4 a t0 (FlashesLights b t0))
  PROOF OF: (Believes!3 a t0 (implies (FlashesLights b t0) (Intends! b t0 (GoFirst a))))
  GIVEN
  PROOF OF: (Believes!4 a t0 (FlashesLights b t0))
  Applied '[I^\ell_4]' to: (Perceives! a t0 (FlashesLights b t0))
    PROOF OF: (Perceives! a t0 (FlashesLights b t0))
    GIVEN
```

Figure A.4: Driving Scenarios (Part 2 of 2)

A.3 Arguments of Chapter 7

```
(base) root@97e884a1add6:/base# python diss_examples/monty_hall_problem.py
Modeling Valid Reasoning in MHP...
PROOF OF: (Believes![l=4,p=2/3] c1 t3 (CB d2))
Applied 'Probabilistic Disjunctive Syllogism' to: (Believes![l=5,p=2/3] c1 t3 (or (CB d2) (CB d3))); (Believes!4 c1 t3 (not (CB d3)))
  PROOF OF: (Believes![l=5,p=2/3] c1 t3 (or (CB d2) (CB d3)))
  Applied '[I^\ell_{PROP}]' to: (Believes![l=5,p=2/3] c1 t2 (or (CB d2) (CB d3)))
    PROOF OF: (Believes![l=5,p=2/3] c1 t2 (or (CB d2) (CB d3)))
    Applied 'Additive Law of Probability' to: (Believes![l=5,p=1/3] c1 t2 (CB d2)); (Believes![l=5,p=1/3] c1 t2 (CB d3))
      PROOF OF: (Believes![l=5,p=1/3] c1 t2 (CB d2))
      Applied 'Probabilistic Belief Intro' to: (Believes!5 c1 t2 (Odds! (CB d2) (POS (CB d2)) (NEG (CB d1) (CB d3))))
        PROOF OF: (Believes!5 c1 t2 (Odds! (CB d2) (POS (CB d2)) (NEG (CB d1) (CB d3))))
        Applied '[I^\ell_5]' to: (Says! h c1 t1 (Odds! (CB d2) (POS (CB d2)) (NEG (CB d1) (CB d3)))); (Knows! c1 t1 (Authoritative h))
          PROOF OF: (Says! h c1 t1 (Odds! (CB d2) (POS (CB d2)) (NEG (CB d1) (CB d3))))
          GIVEN
          PROOF OF: (Knows! c1 t1 (Authoritative h))
          GIVEN
      PROOF OF: (Believes![l=5,p=1/3] c1 t2 (CB d3))
      Applied 'Probabilistic Belief Intro' to: (Believes!5 c1 t2 (Odds! (CB d3) (POS (CB d3)) (NEG (CB d1) (CB d2))))
        PROOF OF: (Believes!5 c1 t2 (Odds! (CB d3) (POS (CB d3)) (NEG (CB d1) (CB d2))))
        Applied '[I^\ell_5]' to: (Says! h c1 t1 (Odds! (CB d3) (POS (CB d3)) (NEG (CB d1) (CB d2)))); (Knows! c1 t1 (Authoritative h))
          PROOF OF: (Says! h c1 t1 (Odds! (CB d3) (POS (CB d3)) (NEG (CB d1) (CB d2))))
          GIVEN
  PROOF OF: (Believes!4 c1 t3 (not (CB d3)))
  Applied '[I^\ell_4]' to: (Perceives! c1 t3 (not (CB d3)))
    PROOF OF: (Perceives! c1 t3 (not (CB d3)))
    GIVEN
Modeling Invalid Reasoning in MHP...
PROOF OF: (and (Believes![l=2,p=1/2] c2 t2 (CarBehind d1)) (Believes![l=2,p=1/2] c2 t2 (CarBehind d2)))
Applied 'Modus Ponens' to: (implies (Believes!4 c2 t2 (not (CarBehind d3))) (and (Believes![l=2,p=1/2] c2 t2 (CarBehind d1)) (Believes![l=2,p=1/2] c2
 t2 (CarBehind d2)))); (Believes!4 c2 t2 (not (CarBehind d3)))
  PROOF OF: (implies (Believes!4 c2 t2 (not (CarBehind d3))) (and (Believes![l=2,p=1/2] c2 t2 (CarBehind d1)) (Believes![l=2,p=1/2] c2 t2 (CarBehind
d2))))
  GIVEN
  PROOF OF: (Believes!4 c2 t2 (not (CarBehind d3)))
  Applied '[I^\ell_4]' to: (Perceives! c2 t2 (not (CarBehind d3)))
    PROOF OF: (Perceives! c2 t2 (not (CarBehind d3)))
    GIVEN
```

Figure A.5: The Monty Hall Problem

APPENDIX B
Authoritative Context

In this appendix we return to the discussion of authoritative agents and authoritative contexts begun in §4.2. To review, inference schema $[I_5^\ell]$ enables an agent a to infer a CERTAIN belief in ϕ if an authoritative agent says ϕ. This concept of an authoritative agent requires more space to flesh out than was possible in that section.

What we mean by *authoritative* is that the agent can be expected *in the given context* to know, without a shadow of doubt, that ϕ is true, generally because they are in control of the situation in some way. It *doesn't* mean that the agent is an authority figure, or that they have authority over the other agent because of their relationship. While either of these *could* be the case, they are not required for an agent to be authoritative.

What is required is what we call *authoritative context*. That is, it is the *context* which gives the agent some authority, and nothing else. If a police officer waives you through a red light, it is clear that they have the authority to do so, since they have the power to direct traffic when needed. But if a police officer in a grocery store tells you to steal a loaf of bread, clearly the officer is not authoritative in this context.

As was said previously, the agent need not be an authority figure. Consider if I show you the Müller-Lyer Illusion [84], a version of which is shown in Figure B.1. Your perception will indicate that the lines are of unequal length. But I can tell you, authoritatively, that they're the same length, since I know about the illusion.

One can even serve introspectively as their own authoritative agent. If you hadn't seen the illusion before, use a ruler or other straightedge to confirm that the endpoints are aligned. Once you remove the straightedge, the lines will once again look uneven. Therefore your perception still indicates that you should believe it is EVIDENT that the lines are unequal; but you know they are not! So, you can essentially tell yourself that, due to your knowledge of the illusion, you should believe that they are the same length, despite what your direct perception indicates.

The notion of authoritative contexts is incredibly important to autonomous driving. Human drivers can easily determine the appropriate contexts in which they should trust others to direct them. It is obvious that a police officer is authoritative when directing traffic. But what about when parking in a field (e.g. at a sporting event or concert) and a regular-

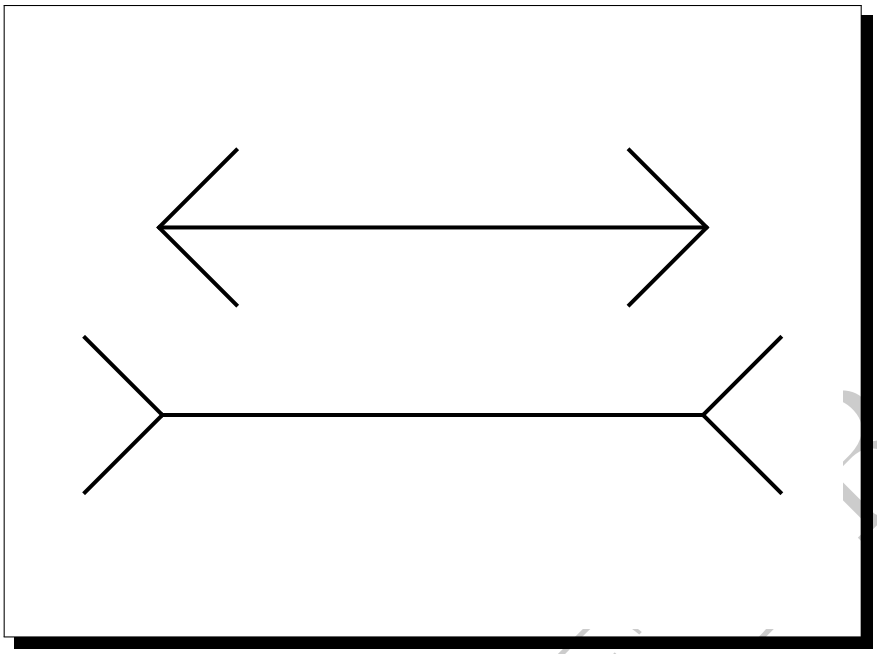

Figure B.1: The Müller-Lyer Illusion

looking person is directing you where to go? Without an understanding of authoritative context, an autonomous car would have to either never trust these people — and hence not truly be autonomous — or trust everyone blindly — and then any person can direct it wherever they please.

APPENDIX C
Supplemental Files

C.1 Permissions for Springer Nature Content

This file contains license details and terms and conditions for the reproduction of material from Giancola et al. [54] for use in this dissertation.

File name: Giancola_2022_Springer_Permissions.pdf

File type: PDF

File size: 87 KB

Required application software: PDF viewer

Special hardware requirements: None