Culturally Aware Social Robots that Carry Humans Inside Them, Protected by Defeasible Argumentation Systems

Selmer BRINGSJORD [a], Michael GIANCOLA [b,1] and
Naveen Sundar GOVINDARAJULU [c]
[a] *Rensselaer AI & Reasoning (RAIR) Laboratory, Department of Computer Science, Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI), Troy, NY, USA*
[b] *RAIR Lab, Dept. of Computer Science, RPI, Troy, NY, USA*
[c] *RAIR Lab, RPI, Troy, NY, USA*

Abstract. After taking note of the conceptual fact that robots may well carry humans inside them, and more specifically that modern AI-infused cars, jets, spaceships, etc. can be viewed as such robots, we present a case study in which inconsistent attitude measurements resulted in the tragic crash in Sweden of such a jet and the death of both pilots. After setting out desiderata for an automated defeasible inductive reasoner able to suitably prevent such tragedies, we formalize the scenario in a first-order defeasible reasoner—OSCAR—and find that it can quickly generate a *partial* solution to the dilemma the pilots couldn't conquer. But we then note and address the shortcomings of OSCAR relative to the desiderata, and adumbrate a solution supplied by a more expressive reasoner based on an inductive defeasible multi-operator cognitive calculus ($\mathcal{IDCEC}$) that is inspired by a merely deductive (monotonic) precursor ($\mathcal{DCEC}$). Our solution in this calculus exploits both the social and cultural aspects of of the jet/robot we suggest be engineered in the future. After describing our solution, some remarks about related prior work follow, we present and rebut two objections, and then wrap up with a brief conclusion.

Keywords. Inductive logic, defeasible reasoning, argumentation

1. Introduction

1.1. Getting Inside Robots—Even Social Ones

The term 'robot' usually conjures up in humans a mental picture of a physical artifact that is physically separate from humans. In the fantastical worlds of Asimov, for instance, 'robot' matches this picture. The world of entertainment, in particular television and film, in no small part has long promoted this mental picture as well.[2] For philosophers,

[1] Corresponding author: Michael Giancola, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA; e-mail: mike.j.giancola@gmail.com

[2] E.g. those of the vintage of the lead author, at least in the United States, well remember the roughly humanoid robot Class M3 Model B9 General Utility Non-Theorizing Environmental Control Robot in *Lost in Space*. (Ironically, this robot had an actor inside it in order to given viewers the impression that it was in many ways a social robot.)

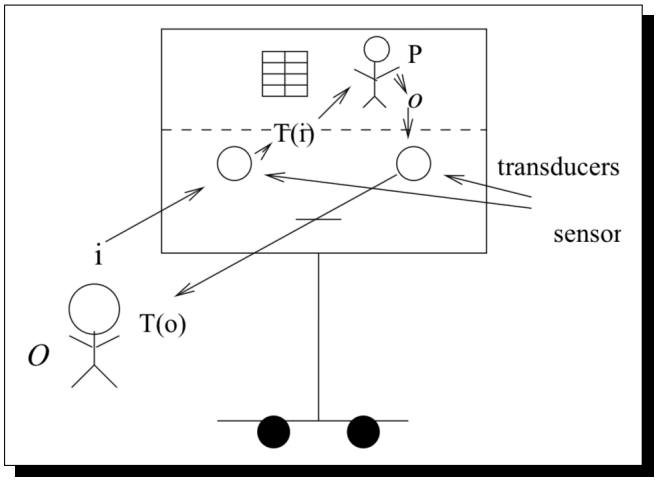

Figure 1. Searle Inside a Robot. (See Searle's "Robot Reply" in [1]. See also variants in [2]. In the figure here, the robot's transducers provide interchange between input i and the output o sent from Searle, which is transduced to $T(o)$, and is then in turn given to those outside the robot.)

especially philosophers of mind, and certainly specifically for philosophers of AI/robots, this picture is certainly not the exclusive one, since for instance Searle [1], in a famous rebuttal to objections to his well-known "Chinese Room Argument," gives the "Robot Reply," in which he is inside the head of a linguistically adept robot; see Figure 1. We believe that increasingly human beings will interact with social robots while they are inside such robots; in fact, human-inside robots may for a long time be the dominant type of robot on Earth. The robots we have in mind include smart, autonomous cars, and advanced aircraft and spaceships—but this is probably only a short, initial enumeration of a list that will grow to include many vehicles. Of course, this would be a list of *mobile* robots. But at least in the abstract, it seems to us inevitable that as stationary spaces become increasingly infused not only with non-physical AI, but with intelligent sensors, manipulators, and displays, the humans-inside-social-robots conception and framework will apply to such spaces as well. For example, our own homes, at least in technologized countries, are increasingly becoming "immobots" that encase us. In short, then, human beings will increasingly spend significant parts of their lives *inside* robots.

1.2. The Hope of Fast, Safe, and Argument-Centric Human-Containing Robots

The performance of modern processors seems to indicate that their descendants will be orders of magnitude faster than the human brain, at least at some substantive tasks. This entails that robots of the future, since they will inevitably be based on such processors, will in turn be capable of reasoning and decision-making at remarkable speeds. Even if such a future never arrives, it surely seems reasonable to hold that eventually a robot equipped with sufficient perceptual and reasoning mechanisms should be able to solve problems (or, as in the case study presented below, avert disasters) that would be impossible for humans to solve, simply because the time allowed for humans to save the day is markedly less than what such a robot requires. Therefore, it would seem prudent to try to engineer robots that can step in when human reasoning is insufficiently fast. If

such robots are in addition safe, then when these robots have us inside them, we will be in circumstances made all the more agreeable. This is the kind of future we seek. This future will be one in which such robots have *fast-acting* social, and specifically cultural, capability—that is, the sort of capability we describe below for future aircraft.

But we believe it will also be necessary for the robots in question to specifically understand arguments, to generate and share arguments, and to judge arguments that "compete" with one another. Bad things, it seems to us, will happen if such capability is absent in robots that we get inside, especially if harm is possible when things go awry. Specifically, as we explain below, robots will need to be able to have the ability to *adjudicate* arguments whose internal elements are cognitive/social, and cultural.

1.3. Plan for the Remainder

The plan for the remainder of the present paper is as follows: We next discuss a case study in which the inadequacy of on-board systems in a jet thrust two (human) pilots therein into a dire position primarily because the speed of their reasoning was insufficient to prevent disaster. This discussion allows us to cite the first two key arguments that were in play during the disaster (which we later analyze). Next, after a brief review of defeasible inductive reasoning, including specifically a setting out of six desiderata (= $\mathcal{D}$) that we believe such reasoning in automated form should satisfy, we discuss how a robot, powered by philosopher John Pollock's artificial rational agent OSCAR [3], which satisfies some of $\mathcal{D}$, could have perhaps averted the disaster. After briefly reflecting on the shortcomings of OSCAR relative to $\mathcal{D}$, we proceed to report part of our formalization of the case study in a defeasible multi-operator cognitive calculus ($\mathcal{IDCEC}$) (which for now the reader can regard to essentially be a highly expressive formal logic able to express cognitive attitudes such as *knows*, *believes*, *intends*, *communicates*, etc), the automated defeasible inductive reasoner that surmounts these shortcomings. After we sketch this success, we consider some relevant prior work, then anticipate and rebut two objections, and end with a few final remarks.

2. A Tragic Case Study

Bäckstrand and Seger [4] 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.

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, in what follows,

due to space constraints, we are forced to for the most part 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 logic/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 [4]. 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.

3. Why?

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 [4]. However, several other factors contributed to the accident; as will be seen, they are, from the standpoint of formal logic and logic-based AI, social/cultural factors.

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 [5]; 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 my detail this missing, life-saving AI.

4. Potential Routes to a Solution

One candidate simple solution, which could be implemented immediately, is to simply prevent the declutter function from hiding any alerts or messages from the comparator function. However, even if the pilots had this information, it is still unlikely they would have been able to prevent the accident, for several reasons, among which are: First, the PIC acted totally out of instinct, without thinking, as one often must in emergency situations with insufficient time for deliberate decision-making. Hence any potential solutions which still require the pilots to resolve the problem without help from a computational reasoner will likely lead to the PIC throwing the plane down in situations that parallel our case study.

The second reason that blocking declutter is no panacea runs as follows: After the initial reaction by the PIC, the pilots were experiencing extreme G-forces, and were overwhelmed by visual warnings and auditory chimes. On page 70, the accident report states that these factors "contributed to cognitive tunnel vision and [led each pilot to] focus on each on-side attitude indicator" [4]. Another way to put this is that any socio-cognitive relationship between the two pilots was obliterated. Because of this, even if the declutter function were modified and a sophisticated communication protocol for pilots in emergency situations was implemented (something in fact recommended by SHK in their report), it is unlikely the pilots would have been able to rationally reason about what was happening, let alone do so within a timespan short enough to avert disaster. Given this, we seek AI/robots that *themselves* sustain a socio-cognitive relationship with the humans within their purview.

4.1. Desiderata for Argument-Centric AI/Robots

Our proposed solution is different. We suggest that an automated reasoner be implemented with the capability to detect and resolve these kinds of inconsistencies autonomously, requiring no intervention from the pilots. This is only one aspect of the kind

of AI capability we seek. We denote the seven-fold desiderata for this capability by '$\mathcal{D}$,' and assert that an automated reasoner of the kind we seek must:

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 inconsistencies when appropriate, and tolerate them when necessary in a manner that fully permits reasoning to continue;
- d_3 make use of values beyond standard bivalence and standard trivalence (e.g. beyond the Kleenean TRUE, FALSE, UNKNOWN trio), specifically probabilities *and* strength-factors (the latter case giving rise to 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 the socio-cognitive elements of knowledge, belief, desire, perception etc. of the humans who are to be helped by this AI;
- 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."

5. Defeasible Reasoning

As is well-known, classical first-order logic is *monotonic*: new information cannot change the result of previous inferences. Defeasible reasoning is *non*-monotonic. It has long been known in AI that such reasoning is desirable when formalizing much real-world reasoning; e.g. see the early, classic default logics of Reiter [6], in which epistemic possibilities hold in default of information to the contrary. In general, it is desirable to be able to reason based on beliefs which could potentially be false, and to be able to retract such beliefs when new, countervailing information arrives. As has been seen, however, our desiderata $\mathcal{D}$ call for more than this. Default logic, despite having many virtues, doesn't satisfy $\mathcal{D}$, a fact space constraints preclude fully discussing here. (Default logic has no provision for operators corresponding to socially and culturally relevant propositional attitudes like *believes*, etc.) A parallel diagnosis must be rendered with respect to circumscription, an impressive nonmonotonic form of reasoning introduced long ago

by John McCarthy [7]. Specifically, circumscription provides no machinery for modal operators to capture mental attitudes, and doesn't include the kind of human-digestible arguments we require. There have been defeasible reasoning models and systems that do include arguments that compete against each other. For an excellent survey of defeasible reasoning systems that are at least to some degree argument-based, see [8].[3] We turn now as promised to a specific automated reasoner that is argument-based, and as such *partially*[4] satisfies desideratum d_4.

6. OSCAR

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

OSCAR employs standard first-order inference schemata as well as Pollock's methods for defeasible reasoning. 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: i.e., the formula which OSCAR will try to establish from the givens. The strength of formulae 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. 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.[5]

7. A Partial Solution in OSCAR

We formalized salient aspects of the crash scenario in OSCAR; this is provided below in lightly modified form to improve readability and save space. First, OSCAR is provided the following premises:

[3] For an efficient overview of defeasible reasoning in general, the interested reader for whom defeasible/nonmonotonic reasoning is new is directed first to [9].

[4] Note that, in addition to requiring that reasoning be argument-based, d_4 necessitates that arguments have internal inference-to-inference structure. This is a desidertatum on which OSCAR falls short, as discussed later in our assessment of OSCAR.

[5] Pollock's original code for OSCAR was written in Macintosh Common Lisp. Kevin O'Neill, while a researcher in the RAIR Lab, resurrected OSCAR after Pollock's passing, in a more modern version of Lisp, Steel Bank Common Lisp (SBCL). This is the version we used to run the simulation discussed herein; however, the results are equivalent to what could be expected from the original code, aside from potentially improved speed.

$$\neg ReadsNormal(iru1)\ \mathrm{J} = 1.0 \tag{P1}$$

$$ReadsNormal(iru2)\ \mathrm{J} = 1.0 \tag{P2}$$

$$MatchesBackup(iru2)\ \mathrm{J} = 1.0 \tag{P3}$$

$$\forall i_1 \forall i_2 \left[\begin{array}{l} \bigwedge \begin{pmatrix} \neg ReadsNormal(i_1), \\ ReadsNormal(i_2), \\ MatchesBackup(i_2) \end{pmatrix} \\ \rightarrow NormalAttitude \end{array} \right] \mathrm{J} = 0.9 \tag{P4}$$

OSCAR is also given the following *prima facie*[6] reason:

$$\begin{pmatrix} \neg ReadsNormal(iru1) \\ \rightarrow \neg NormalAttitude \end{pmatrix} \text{ strength} = 0.6 \tag{R1}$$

Finally, OSCAR was given the goal of proving

$$NormalAttitude$$

This formalization is consistent with (but doesn't capture) the scenario. We are confident that the PIC's IRU was giving an abnormal reading, the co-pilot's IRU was giving a normal reading, and that the co-pilot's IRU matched the reading from the backup instruments. OSCAR's goal is to establish (or refute) that the plane was at a normal attitude when the faulty IRU reading materialized.

The particular numeric values chosen for the justifications and strengths are irrelevant. The only significance with regard to OSCAR's reasoning is the *relative* strength of two statements. That is, the values 0.9 and 0.6 above are not intrinsically significant beyond that, when there is a discrepancy between the two IRUs, the reading which more closely matches the backup instruments should have higher—to used Pollock's terminology—warrant. Therefore, although OSCAR doesn't treat the justifications/strengths as probability values, OSCAR's argument is uncertain and hence inductive, satisfying (at least in part) desideratum d_3.

OSCAR finds a proof of *NormalAttitude* in 0.22 seconds, as well as a defeated proof for the negation of (P4); i.e., OSCAR found a proof of *NormalAttitude* which critically utilized (P4), but also found a weaker proof that would invalidate the proof for *NormalAttitude* if it were stronger. In effect, OSCAR proved that the plane should have stayed on its course with the primary IRU left aside, assuming we trust the conjunct of the co-pilot's IRU and the backup instruments over the PIC's IRU.

8. Assessing OSCAR

While OSCAR is able to satisfactorily model some aspects of the case study, it falls short of meeting the list $\mathcal{D}$ of desiderata we prescribed earlier. First, while OSCAR includes a set of deductive inference schemata for first-order logic, it has no inference schemata whatsoever for its *inductive* arguments. Hence its adjudication of several argu-

[6] This has a technical sense in OSCAR which space constraints compel us to leave aside.

ments makes use of no analysis of the internal structure of individual inference steps that human beings routinely engage in. 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; a classic scheme in this tradition is presented in [10]. OSCAR therefore doesn't satisfy d_4, and doesn't satisfy d_5. Also, as it is limited to first-order logic, OSCAR cannot satisfy d_6 without falling into unsoundness, as shown in [11].

9. Steps Toward a Satisfactory Defeasible Reasoner

We argue that a system which meets desiderata $\mathcal{D}$ will be able to model and reason about a larger set of scenarios, including a richer version of the flight-crash scenario described above. To the authors' knowledge, no such system currently exists. In order to create one, we will need to add several elements to our logicist armamentarium. The first addition is to turn to so-called *cognitive calculi*. We quickly summarize these calculi in the next subsection, and after that turn back to presenting our techniques in connection to the crash scenario discussed previously.

9.1. Cognitive Calculi (Deductive)

Essentially, a deductive cognitive calculus is a quantified multi-operator modal logic such that its: proof/argument theory is specified in "natural deduction" form [traceable back to [12,13]], operators cover all or most of human-level cognition needed for social interaction and culture (e.g., *believing*, *knowing*, *perceiving*, *communicating*, and also *obligations*, etc.), and semantics is exclusively proof-theoretic in nature. Proof-theoretic semantics eschews model-theoretic and possible-worlds semantics in favor of the basic idea that meaning is provided to formulae and their constituents solely by virtue of the nature of proofs in which these things appear. [For more on proof-theoretic semantics see [14,15,12,16].]

In the present work, we specifically utilize elements of the *Deontic Cognitive Event Calculus* ($\mathcal{DCEC}$) to model the perceptions (or lack thereof), beliefs, intentions, and obligations of the pilots. A dialect of $\mathcal{DCEC}$ is specified and used in [17]. $\mathcal{DCEC}$ includes all of the introduction and elimination schemata for first-order logic, plus a host of inference schemata to cover its many modal operators. Soundness proofs for cognitive calculi have been obtained but are out of scope. Also, an automated theorem prover for $\mathcal{DCEC}$-–ShadowProver [18]—has been created, repeatedly deployed, and is still under active development.

9.2. Inductive Cognitive Calculi

$\mathcal{DCEC}$ is purely deductive and employs no uncertainty system, so it fails to satisfy d_1–d_4. Therefore, to meet desiderata d_1 and d_2, we will utilize the *Inductive* $\mathcal{DCEC}$ ($\mathcal{IDCEC}$), which has been modified to handle inductive arguments.[7]

[7] Such arguments have long been taken by philosophers to be the heart of what inductive logic is; see for instance [19].

To meet desideratum d_3, we employ "strength factors" within a nascent system for formalizing uncertainty in quantified modal logics, first presented in [20]. Strength factors can be viewed as a formalization of the philosopher Roderick Chisholm's epistemology [21]. The current version, which this work is based on, has a 13-value spectrum of strength introduced by Bringsjord to extend Chisholm's scheme, with zero being *counterbalanced* (no belief for or against some formula), increasing positive integers indicating stronger belief in favor of some formula, and decreasing negative integers indicating stronger belief against some formula. The strength factors relevant to the work we present herein are *evident* (level 5) and *overwhelmingly likely* (level 4). The highest strength level—*certain* (level 6)—is not used herein.

Finally, to meet d_4, we will need methods for adjudicating conflicting arguments with regard to which action to take in response to the alerting IRU reading. Prior work in this area was presented in [22].

10. 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 a proof as justification for their conclusion. When two (or more) automated reasoners disagree, the adjudicator $\mathfrak{a}^\star$ resolves the conflict.

For the purposes of the central case study of the present paper, denote two AI automated reasoners $\mathfrak{r}_1$ and $\mathfrak{r}_2$, and the adjudicator $\mathfrak{a}^\star$.[8] It should become clear as we walk through the reasoning that it is shot through with socio-cognitive aspects that are part and parcel of what human culture is. Now here the reasoning itself:

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:

$$\mathbf{B}^5(\mathfrak{r}_1, t_0, iru_1) \wedge \mathbf{B}^5(\mathfrak{r}_2, t_0, iru_1)$$

This formula expresses that both automated reasoners believe it is *evident* that they have received a reading from iru_1.[9] 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.

10.1. 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 iru_1 seems

[8] This adjudicator is intended to be an analog to the "dictator" in Arrow's Impossibility Theorem (AIT) [23], except that the adjudicator is capable of using each voter's choice and corresponding argument to make its conclusion in a way that is fair to each voter. In this way we are able to overcome AIT.

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

irregular, $\mathfrak{r}_1$ observes the readings of iru_2 and *backup*, and because this reasoner has direct perceptual access to these sub-systems, belief at level 5 is justified; specifically:

$$\mathbf{B}^5(\mathfrak{r}_1, t_1, iru_2) \wedge \mathbf{B}^5(\mathfrak{r}_1, t_1, backup)$$

$\mathfrak{r}_1$ 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, can:[10]

$$\begin{aligned} &\mathbf{P}(\mathfrak{r}_1, t_1, celestial\_bodies) \\ \mathbf{B}^5(\mathfrak{r}_1, t_1, &\neg\mathbf{P}(p_1, t_1, celestial\_bodies) \\ &\wedge\neg\mathbf{P}(p_2, t_1, celestial\_bodies)) \end{aligned}$$

Knowing this, $\mathfrak{r}_1$ believes it is *evident* (belief level 5) that telling p_1 the reading of iru_1 will lead him to believe that the plane is going to stall:

$$\begin{aligned} &\mathbf{B}^5(\mathfrak{r}_1, t_2, \\ &\quad (\neg\mathbf{P}(p_1, t_1, celestial\_bodies) \wedge \mathbf{S}(\mathfrak{r}_1, p_1, t_2, iru_1)) \\ &\qquad \rightarrow \mathbf{B}^5(p_1, t_3, GoingToStall(plane, t_3))) \end{aligned}$$

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

$$\begin{aligned} \mathbf{B}^5(\mathfrak{r}_1, t_3, &\mathbf{B}^5(p_1, t_3, GoingToStall(plane, t_3)) \rightarrow \\ &\mathbf{I}(p_1, t_4, LowerPitch(plane))) \end{aligned}$$

$\mathfrak{r}_1$ 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:

$$\begin{aligned} &\mathbf{B}^5(\mathfrak{r}_1, t_4, \mathbf{P}(\mathfrak{r}_1, t_4, celestial\_bodies) \\ &\quad \rightarrow NormalAttitude(plane, t_4)) \end{aligned}$$

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:

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

$$\mathbf{B}^5(\mathfrak{r}_1, t_4, (iru_1 \wedge iru_2 \wedge backup) \\ \rightarrow NormalAttitude(plane, t_4))$$

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

$$\Big(\mathbf{B}^5(\mathfrak{r}_1, t_4, NormalAttitude(plane, t_4)) \wedge \\ \mathbf{B}^5(\mathfrak{r}_1, t_4, LowerPitch(p_1, plane, t_4))\Big) \\ \rightarrow Crash(plane, t_5)$$

Finally, $\mathfrak{r}_1$ concludes that, due to an obligation to keep the pilots safe, it is obligated to not transmit the data from iru_1 to p_1:[11]

$$\mathbf{O}(\mathfrak{r}_1, t_2, \neg\mathbf{S}(\mathfrak{r}_1, p_1, t_2, iru_1))$$

10.2. Argument 4 (= α_4)

Reasoner $\mathfrak{r}_2$ disregards the output of iru_2 and *backup*, perhaps to save time:

$$\neg\mathbf{B}^5(\mathfrak{r}_2, t_1, iru_2) \wedge \neg\mathbf{B}^5(\mathfrak{r}_2, t_1, backup)$$

The reasoner instead generates the argument that iru_1 is generally reliable and hence should be trusted. However, $\mathfrak{r}_2$'s belief that iru_1 is reliable can only at highest reach level 4; that is, that it is *overwhelmingly likely* that iru_1 is reliable, as IRU's rarely malfunction. However, it is not *evident* (level 5), as these malfunctions are certainly known to sometimes happen. We can depict a sketch of $\mathfrak{r}_2$'s reasoning as follows:

$$\mathbf{B}^4(\mathfrak{r}_2, t_1, IsReliable(iru_1)) \\ IsReliable(iru_1) \rightarrow \mathbf{S}(\mathfrak{r}_2, p_1, t_2, iru_1) \\ \therefore \mathbf{O}(\mathfrak{r}_2, t_2, \mathbf{S}(\mathfrak{r}_2, p_1, t_2, iru_1))$$

Clearly, $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are at an impasse. In our aspirational version of the scenario, the adjudicator $\mathfrak{a}^\star$ 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 5—while $\mathfrak{r}_2$'s argument hinges on a belief at level 4. Therefore, acting as a defeasible reasoner, and employing the axiomatic principle that no

[11] 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.

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 $IsReliable(iru_1)$ 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.

10.3. On Cultural Aspects of the Solution

We now very briefly indicate the cultural aspects of the scenario and reasoning we have described. These aspects are based on a basic foundation for a formalization of culture given in [24], which is inspired by [25]. According to this foundation, such a formalization must include six elements: three sets of formulae: The Real, The Book, and The Hope; a set of actions that are typical for members of the culture in question (= The Habits); a set of processes by which a member of the relevant culture has learned those things in The Book, The Hope, and The Habits; and finally a set of patterns of reasoning (= The Reasoning), not at all necessarily logically valid, that members of the culture in question follow. Here are a quick set of comments to indicate how each of the six elements are operative in the case of the kind of socially adept, argument-centric AI-infused aircraft we have described above:

- *The Real* A salient member of this set of propositions in the scenario described above is that celestial bodies in the night sky are clear and steady, and not "rising."
- *The Book* Any number of true propositions about piloting, and about the aircraft/robot in question, all known to both humans, are included is this set of formulae. For instance, both pilots know where each of their altitude indicators are located.
- *The Hope* According to [24], every culture has members that believe things that, in the minds of those outside the culture, are, to put it mildly, far from certain. In the case of professional pilots in general, and in particular in the case the pilots in our case study, there clearly are strong candidates for propositions in this category, but we refrain from venturing examples, in the interests of decorum.
- *The Habits* When the pilots took off from Oslo-Gardenmoen, they did so by following a specified protocol. There are in fact innumerable habits that pilots routinely follow.
- *The Inculcation* Obviously, both pilots are highly trained in the "habits" of their profession.
- *The Reasoning* This is here unfortunately a disturbing category in the six-part basis for culture, since it was faulty reasoning that contributed to the tragedy.

11. Related Work

The need for automated systems to detect and resolve issues created by faulty sensor readings was discussed in [26]. However, to our knowledge, no work has been published regarding the attempt to use automated defeasible reasoning (whether purely deductive or—as in our approach—deductive *and* inductive) to manage life-threatening erroneous signals in the operation of aircraft, let along more broadly in human-holding social robots. Gómez et al. [27] and Nakamatsu et al. [28] created models to improve safety in air-traffic control using defeasible reasoning and paraconsistent logic, respectively; however, their logics are not expressive enough to capture intensional attitudes at the heart of cognitive/social cognition (e.g. theory-of-mind attitudes: knowledge, belief, obligation, etc.), and hence fail on desideratum d_6. Also, [29] modeled a complex case of deceptive reasoning and planning from the award-winning television series *Breaking*

Bad using default logic. Their work did use a cognitive calculus (the Cognitive Event Calculus, $\mathcal{CEC}$) to model the beliefs and intentions of various agents, but didn't have a formalism for assigning strengths to beliefs; therefore, while commendable on many fronts, their system does not satisfy d_3.

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

1. Modgil and Prakken [30] present a general framework—ASPIC$^+$—for structured argumentation, and the framework is certainly computational in nature. ASPIC$^+$ is in fact Pollockian in nature, at least in part. More specifically this framework is based upon two fundamental principles, the second of which is that "arguments are built with two kinds of inference rules: strict, or deductive rules, whose premises guarantee their conclusion, and defeasible rules, whose premises only create a presumption in favor of their conclusion" [p. 31 of [30]]. This second principle is directly at odds with desideratum d_5. In our approach, all non-deductive inference schemata are checkable, in exactly the way that deductive inference schemata are. For instance, if some inference is analogical in nature, as long as the schema $\frac{\Phi}{C}$ (Φ for a collection of premises and C for the conclusion) for an analogical inference is correctly followed, the inference is watertight, no different than even *modus ponens*, where of course specifically we have $\frac{\phi \rightarrow \psi, \phi}{\psi}$.[12]
2. Cerutti et al. [32] is an overview of implementations of formal-argumentation systems. However, the overview is highly constrained by two attributes. The first is that their emphasis is on Turing-decidable reasoning problems. As to the second attribute, the authors are careful to say that their work is constrained by the "basic requirement" that "conflicts" between arguments are "solved by selecting subsets of arguments," where "none of the selected arguments attack each other." Both of these attributes are rejected in our approach. In fact, with respect to the first, most of the interesting parts of automated-reasoning science and technology for us only *start* with problems at the level of the *Entscheidungsproblem*; see in this regard desideratum d_7. As to the second attribute, it too is not true of our approach.

12. Two Objections

We anticipate a number of objections. In what follows, we summarize two of these, and in each case offer a brief rebuttal.

12.1. "Be realistic about stochastic environments!"

The objection here runs as follows: "It seems to me that you actually believe that if the right sort of automated-reasoning capability is produced and deployed, then accidents in which the humans inside robot vehicles are injured or killed can be avoided. This is a misleading promise, as AI can only build rational agents, not omniscient beings. In par-

[12] For a discussion of this sort of explicit rigidity in the case of analogical inference, see [31].

tially observable and stochastic environments, which are surely the type of environments your humans-inside robots will operate in, it is impossible for rational agent's actions to always save the day."

Our reply is straightforward. We do indeed think that as humans risk life and limb by going inside robots, their best bet for safety is a socio-cultural intelligence on the part of these robots that is of the sort we have described above. But this in no way entails that we expect these robots to be omniscient. As it happens, there has been some work [33] by authors who have explained that stochastic environments are challenging, but they also point out that after all since such environments still obey the laws of physics, there is no reason why the robots in question cannot anticipate even seemingly exogenous events by rapid reasoning over the combination of such laws and information from sensors. Unlike the robot/jet in the sad scenario of the crash in Scandinavia that we have analyzed above, we hope that in the future the sensors in question will perceive more extensive information about the environment.

12.2. "Can this scale?"

The second objection can be expressed as follows: "In your desiderata $\mathcal{D}$, d_5 requires automated reasoning over the knowledge, beliefs, perceptions, etc. of human pilots. How would this system possibly scale to the level of complexity of human decision making in emergency aircraft situation?"

This objection is tantamount to denying that fast-thinking human-holding social robots are possible. Unfortunately, the critic merely gives a rhetorical question, to which we are within our dialectical rights in responding with: "You supply no reason whatsoever why a social humans-inside robot cannot reason over thousands or even millions of propositions involving mental attitudes and emotions. Besides, if we don't earnestly try in social robotics to push in the direction we depict, one thing's for sure: we'll never get there."

13. Conclusion

Without employment of appropriate automated reasoners, accidents like the one analyzed above will likely continue to happen. This suggests to us that robots of the sort that hold humans must have the kind of capability we have described and, in part, demonstrated. While AI researchers have achieved some impressive automated defeasible reasoners, these systems are unable to satisfy our wide range of relevant desiderata; in particular, the mental attitudes of human pilots (i.e. d_6 in $\mathcal{D}$), which are central to social, and specifically, cultural, cognition and computing. We hope that the maturation of our research and that of others leads to the development of safer aircraft, and safer human-inside robots of other kinds.

Acknowledgments

The authors are indebted to both AFOSR & ONR for support of our r&d in the area of automated defeasible reasoning (in both deductive and inductive modes), of development of the formalisms underlying such reasoning, and of work toward applications made

possible by such reasoning. We are especially grateful to ONR for supporting our attempt to surmount Arrow's Impossibility Theorem via automated argument adjudication. We thank a number of scholars who participated in Robophilosophy 2020 and specifically provided helpful feedback to us, and the same holds at this point for four anonymous referees. We are also grateful to Jonas Bäckstrand and Nicolas Seger, the authors of the accident report upon which we have relied, for their kind responsiveness and assistance in interpreting their very helpful report.

References

[1] Searle J. Minds, Brains and Programs. Behavioral and Brain Sciences. 1980;3:417–424.

[2] Bringsjord S, Noel R. Real Robots and the Missing Thought Experiment in the Chinese Room Dialectic. In: Preston J, Bishop M, editors. Views into the Chinese Room: New Essays on Searle and Artificial Intelligence. Oxford, UK: Oxford University Press; 2002. p. 144–166.

[3] Pollock JL. Cognitive Carpentry: A Blueprint for How to Build a Person. Mit Press; 1995.

[4] Bäckstrand J, Seger N. Final Report RL 2016:11e. Swedish Accident Investigation Authority; 2016. Accessible as of November 13, 2019 here: https://havkom.se/assets/reports/RL-2016%5f11e.pdf.

[5] Varela CA. Too Many Airplace Systems Rely on Too Few Sensors; 2019. Article published in TheConversation.com. Available from: http://theconversation.com/too-many-airplane-systems-rely-on-too-few-sensors-114394.

[6] Reiter R. A Logic for Default Reasoning. Artificial Intelligence. 1980;13:81–132.

[7] McCarthy J. Circumscription—A Form of Non-Monotonic Reasoning. Artificial Intelligence. 1980;13:27–39.

[8] Prakken H, Vreeswijk G. Logics for Defeasible Argumentation. In: Gabbay D, Guenthner F, editors. Handbook of Philosophical Logic. Dordrecht, The Netherlands: Springer; 2001. p. 219–318.

[9] Koons R. Defeasible Reasoning. In: Zalta E, editor. The Stanford Encyclopedia of Philosophy; 2017. Available from: https://plato.stanford.edu/entries/reasoning-defeasible/index.html.

[10] Dung P. On the Acceptability of Arguments and its Fundamental Fole in Nonmonotonic Reasoning, Logic Programming and N-Person Games. Artificial Intelligence. 1995;77:321–357.

[11] Bringsjord S, Govindarajulu NS. Given the Web, What is Intelligence, Really? Metaphilosophy. 2012;43(4):464–479.

[12] Gentzen G. Untersuchungen über das logische Schlieben I. Mathematische Zeitschrift. 1935;39:176–210.

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

[14] Dummett M. Frege. Philosophy of Language (2nd ed). London, UK: Duckworth; 1981.

[15] Dummett M. The Logical Basis of Metaphysics. London, UK: Duckworth; 1991.

[16] Prawitz D. The Philosophical Position of Proof Theory. In: Olson RE, Paul AM, editors. Contemporary Philosophy in Scandinavia. Baltimore, MD: Johns Hopkins Press; 1972. p. 123–134.

[17] Govindarajulu NS, Bringsjord S. On Automating the Doctrine of Double Effect. In: Sierra C, editor. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). International Joint Conferences on Artificial Intelligence; 2017. p. 4722–4730. Available from: https://doi.org/10.24963/ijcai.2017/658.

[18] Govindarajulu NS, Bringsjord S, Peveler M. On Quantified Modal Theorem Proving for Modeling Ethics. In: Suda M, Winkler S, editors. Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements (ARCADE 2019). vol. 311 of Electronic Proceedings in Theoretical Computer Science. Waterloo, Australia: Open Publishing Association; 2019. p. 43–49. The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/. Available from: http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf.

[19] Johnson G. Argument & Inference: An Introduction to Inductive Logic. Cambridge, MA: MIT Press; 2016.

[20] Govindarajulu NS, Bringsjord S. Strength Factors: An Uncertainty System for Quantified Modal Logic. In: Belle V, Cussens J, Finger M, Godo L, Prade H, Qi G, editors. Proceedings of the IJCAI Workshop on

"Logical Foundations for Uncertainty and Machine Learning (LFU-2017). Melbourne, Australia; 2017. p. 34–40. Available from: http://homepages.inf.ed.ac.uk/vbelle/workshops/lfu17/proc.pdf.

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

[22] Giancola M, Bringsjord S, Govindarajulu NS, Licato J. Adjudication of Symbolic & Connectionist Arguments in Autonomous Driving AI. In: Danoy G, Pang J, Sutcliffe G, editors. GCAI 2020. 6th Global Conference on Artificial Intelligence. vol. 72 of EPiC Series in Computing. EasyChair; 2020. p. 28–33. Available from: https://easychair.org/publications/paper/Vtl4.

[23] Morreau M. Arrow's Theorem. In: Zalta E, editor. The Standford Encyclopedia of Philosophy. Winter 2016 ed.; 2014. Available from: https://plato.stanford.edu/entries/arrows-theorem.

[24] Bringsjord S. Toward Formalizing Culture: First Steps. Hypothesis. 2014;1(1):19–27. Available from: http://kryten.mm.rpi.edu/SB_FormalizingCulture_121313.pdf.

[25] March J. Exploration and Exploitation in Organizational Learning. Organization Science. 1991;2(1):71–87.

[26] Imai S, Blasch E, Galli A, Zhu W, Lee F, Varela CA. Airplane Flight Safety Using Error-Tolerant Data Stream Processing. IEEE Aerospace and Electronics Systems Magazine. 2017;32(4):4–17. Available from: http://www.brightcopy.net/allen/aesm/32-4/index.php#/6.

[27] Gómez SA, Goron A, Groza A. Assuring Safety in an Air Traffic Control System with Defeasible Logic Programming. In: XLIII Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO)-XV Argentine Symposium on Artificial Intelligence (ASAI)(Buenos Aires, 2014); 2014. .

[28] Nakamatsu K, Suito H, Abe J, Suzuki A. Paraconsistent Logic Program Based Safety Verification for Air Traffic Control. In: IEEE International Conference on Systems, Man and Cybernetics. vol. 5. IEEE; 2002. .

[29] Licato J. Formalizing Deceptive Reasoning in Breaking Bad: Default Reasoning in a Doxastic Logic. In: 2015 AAAI Fall Symposium Series; 2015. .

[30] Modgil S, Prakken H. The ASPIC$^+$ Framework for Structured Argumentation: A Tutorial. Argument & Computation. 2014;5(1):31–62.

[31] Bringsjord S, Licato J. By *Dis*analogy, Cyberwarfare is Utterly New. Philosophy and Technology. 2015;28(3):339–358. Available from: http://kryten.mm.rpi.edu/SB_JL_cyberwarfare_disanalogy_DRIVER_final.pdf.

[32] Cerutti F, Gaggl SA, Thimm M, Wallner J. Foundations of Implementations for Formal Argumentation. In: Baroni P, Gabbay D, Giacomin M, Van der Torre L, editors. The IfCoLog Journal of Logics and their Applications; Special Issue Formal Argumentation. vol. 4. College Publications; 2017. p. 2623–2705.

[33] Bringsjord S, Sen A. On Creative Self-Driving Cars: Hire the Computational Logicians, Fast. Applied Artificial Intelligence. 2016;30:758–786. The URL here goes only to an uncorrected preprint. Available from: http://kryten.mm.rpi.edu/SB_AS_CreativeSelf-DrivingCars_0323161130NY.pdf.