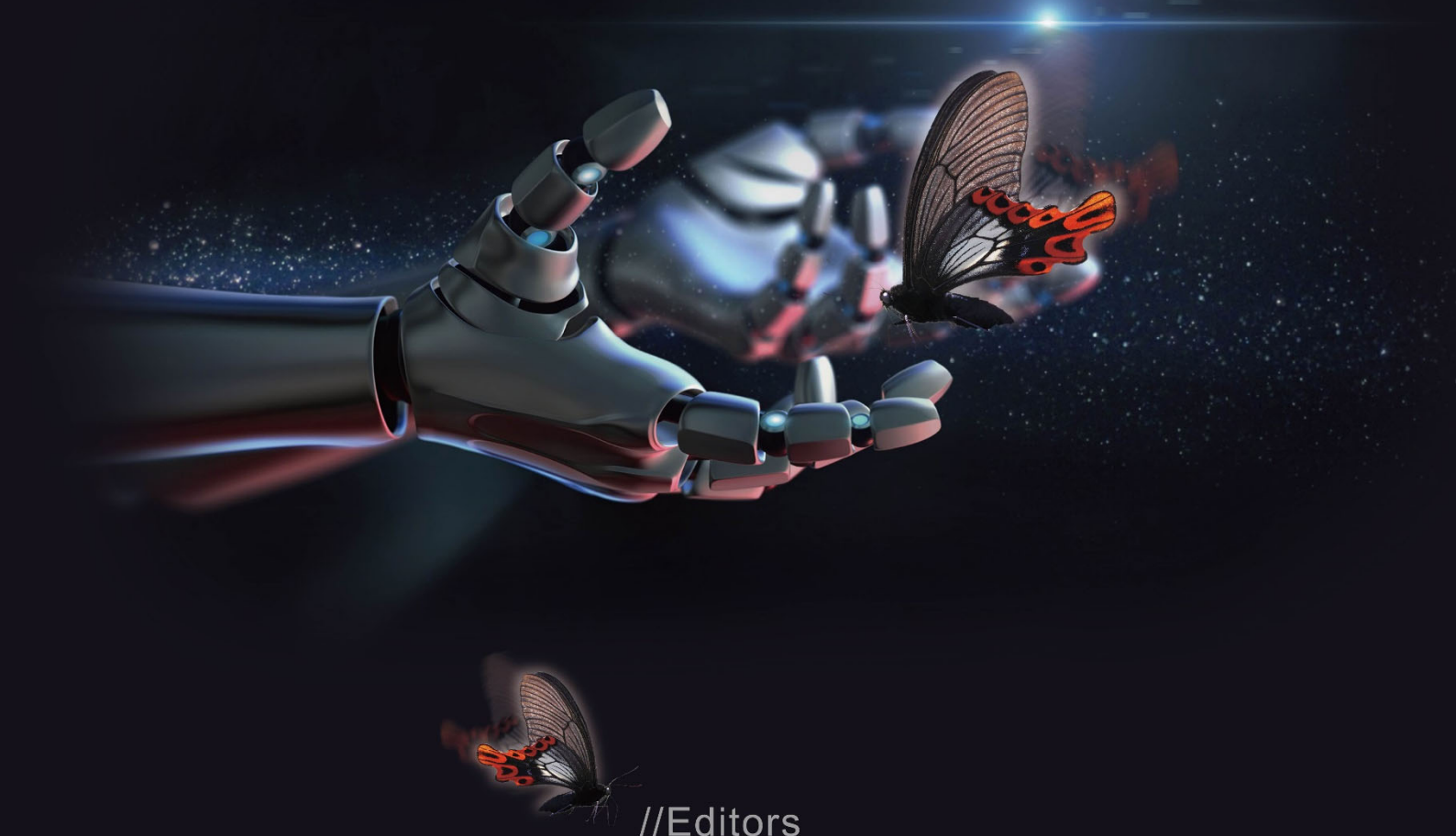
ICRES 2020

5th International Conference
on Robot Ethics
and Standards

Taipei, Taiwan,
28-29 Sept. 2020

Smart Living and Quality Health with Robots



//Editors

Mohammad O. Tokhi
Maria Isabel A. Ferreira
Naveen S. Govindarajulu
Manuel F. Silva
Endre E. Kadar
Jen-Chieh Wang
Aman P. Kaur

SMART LIVING AND QUALITY HEALTH WITH ROBOTS

SMART LIVING AND QUALITY HEALTH WITH ROBOTS

ICRES 2020 Proceedings,
Taipei, Taiwan, 28-29 September 2020

Editors

Mohammad Osman Tokhi
London South Bank University, UK

Maria Isabel Aldinhas Ferreira
University of Lisbon, Portugal

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute, NY, USA

Manuel F. Silva
Porto Polytechnic, Portugal

Endre E. Kadar
University of Portsmouth, UK

Jen-Chieh Wang
Industrial Technology Research Institute, Taipei, Taiwan

Aman P. Kaur
London South Bank Innovation Centre, Cambridge, UK

Published by

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

Smart Living and Quality Health with Robots
Proceedings of the Fifth International Conference on Robot Ethics and Standards

PREFACE

ICRES 2020 is the fifth edition of International Conference series on Robot Ethics and Standards. The conference is organized by CLAWAR Association in collaboration with the Industrial Technology Research Institute (ITRI), and held in Taipei, Taiwan on a virtual platform during 28 – 29 September 2020.

ICRES 2020 brings new developments and new research findings in robot ethics and ethical issues of robotic and associated technologies. The topics covered include fundamentals and principles of robot ethics, social impact of robots, human factors, regulatory and safety issues.

The ICRES 2020 conference includes a total of four plenary lectures, and 26 regular and invited presentations. A special discussion panel session covering the impact of artificial intelligence in context of covid-19 is also organised.

The editors would like to thank members of the International Scientific Committee and Local Organising Committee for their efforts in reviewing the submitted articles, and the authors in addressing the comments and suggestions of the reviewers in their final submissions. It is believed that the ICRES 2020 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies.

M. O. Tokhi, M. I. A. Ferreira, N. S. Govindarajulu, M. F. Silva, E. E. Kadar, J.-C. Wang and A. P. Kaur

CONFERENCE ORGANISERS

CLAWAR Association
www.clawar.org

ITRI
Industrial Technology
Research Institute

CONFERENCE SPONSORS AND SUPPORTERS

CONFERENCE COMMITTEES AND CHAIRS

Conference Chairs and Managers

Tzeng-Yow Lin (General Co-Chair)	– Industrial Technology Research Institute, Taiwan
Mohammad Osman Tokhi (General Co-Chair)	– London South Bank University, UK
Maria Isabel Aldinhas Ferreira (General Co-Chair)	– University of Lisbon, Portugal
Jen-Chieh Wang (Secretariat)	– Industrial Technology Research Institute, Taiwan
Dimitris Chrysostomou (Web-site)	– Aalborg University, Denmark

International Advisory Committee

Gurvinder S. Virk (Chair)	– CLAWAR Association, UK
Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Raja Chatila	– ISIR/UPMC-CNRS, France
Kerstin Dautenhahn	– University of Waterloo, Canada
Edson Prestes	– Federal University of Rio Grande do Sul, Brazil
Alan Winfield	– University of the West of England, UK

International Scientific Committee

Naveen S. Govindarajulu (Co-Chair)	– Rensselaer Polytechnic Institute, USA
Manuel F. Silva (Co-Chair)	– ISEP & INESCTEC, Portugal
Ronald Arkin	– Georgia Institute of Technology, USA
Dimitris Chrysostomou	– Aalborg University, Denmark
Nigel Crook	– Oxford Brookes University, UK
Roeland de Bruin	– Utrecht University, The Netherlands
Sarah Fletcher	– Cranfield University, UK
Khaled Goher	– University of Lincoln, UK
Joseph Johnson	– Rensselaer Polytechnic Institute, USA
Endre E. Kadar	– University of Portsmouth, UK
Aman Kaur	– London South Bank University, UK
Philip Lance	– PA Consulting, UK
João Sequeira	– University of Lisbon, Portugal
Murray Sinclair	– Loughborough University, UK
Anouk van Maris	– University of the West of England, UK
David Vernon	– Carnegie Mellon University Africa, Rwanda
Sean Welsh	– University of Canterbury, New Zealand
Karolina Zawieska	– Industrial Research Institute for Automation and Measurement, Poland

Local Organising Committee

Jia-Jin Chen (Chair)	– Standard Committee of Assistive Technology and Robot for Healthcare Uses, Taiwan
Tzu-Wei Li	– Industrial Technology Research Institute, Taiwan
Tzuan-Ren Jeng	– Industrial Technology Research Institute, Taiwan
Jen-Chieh Wang	– Industrial Technology Research Institute, Taiwan

TABLE OF CONTENTS

Section–1: Plenary presentations

Section–2: Regular presentations

Section–3: Special session

Section–4: Invited presentations

ICRES 2020: 5th International Conference on Robot
Ethics and Standards, Taipei, Taiwan, 28-29 September 2020.

ETHICAL REASONING FOR AUTONOMOUS AGENTS UNDER UNCERTAINTY

MICHAEL GIANCOLA[⋆] and SELMER BRINGSJORD[†] and NAVEEN SUNDAR GOVINDARAJULU[⋆] and CARLOS VARELA[‡]

Rensselaer AI & Reasoning (RAIR) Lab[⋆†]; Worldwide Computing Lab (WCL)[‡]
Department of Computer Science[⋆†‡]; Department of Cognitive Science[†]
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
www.rpi.edu
E-mail: { mike.j.giancola, selmer.bringsjord, naveen.sundar.g } @gmail.com, cvarela@cs.rpi.edu

Autonomous (and partially autonomous) agents are beginning to play significant roles in safety-critical and privacy-critical domains, such as driving and healthcare. When humans operate in these spaces, not only are there regulations and laws dictating proper behavior, but crucially, neurobiologically normal humans can be expected to comprehend how to reason with certain principles to ensure that their actions are legally/ethically/prudentially correct (whether or not these humans choose to abide by the principles in question). It seems reasonable that we should hold autonomous agents to, minimally, the same standard we hold humans to. In this paper, we present a framework for autonomous aircraft piloting agents to reason about ethical problems in the context of emergency landings. In particular, we are concerned with ethical problems in which every option is equally unethical with regard to the ethical principles the options violate; and the only distinguishing factor is the *likelihood* that a plan will violate an ethical principle. We conclude by discussing why, in general, we find an inference-theoretic approach to ethical reasoning to be superior to the model-theoretic approach of prior work.

Keywords: Ethical reasoning; Reasoning under uncertainty; Modal logic

1. Introduction

Autonomous (and partially autonomous) agents are beginning to play significant roles in safety-critical and privacy-critical domains, such as driving and healthcare. When humans operate in these spaces, not only are there regulations and laws dictating proper behavior, but crucially, neurobiologically normal humans can be expected to comprehend how to reason with regulations to ensure that their actions remain within the law (regardless of whether or not they choose to abide by the law). It seems reasonable that we should hold autonomous agents to, minimally, the same standard we hold humans to. That is, autonomous agents must be aware of relevant ethical constraints (those actions which are e.g. *obligatory*, *permissible*, *forbidden*, *supererogatory*, etc.; see [1] for a full discussion of these concepts in deontic logic), and be able to reason with them to determine actions which will satisfy those constraints. In addition, we require of our autonomous agents to not only verify their behavior to be ethical, but to output an argument[a] which can be inspected by a human. By our lights, this requires an inference-theoretic approach to ethical reasoning.

[a]While too large a topic to cover in full formal glory due to space constraints, we briefly address the distinction we draw between a formal 'argument' and a formal 'proof'. In our work, an argument is invariably a *formal* argument, in the sense that it purportedly includes reference, at each inferential link, to an inference schema that sanctions that inference. Hence, all formal proofs are arguments for us. However, any chain of reasoning that makes use of uncertainty measures (e.g., probability values in the real interval $[0, 1]$, or our strength factors $\sigma_i, 1 \leq i \leq 4$ (introduced, explained, and employed below)), even though its inferences are backed by appeal to inference schemata, cannot be classified as a proof. A proof must make no use of uncertainty measures.

We will discuss in greater detail why we find an inference-theoretic approach to be superior to a model-theoretic approach in §4.

1.1. *The Inspiring Prior Work*

This work was inspired by previous work of Dennis et al. [2]. In this interesting 2016 paper, they presented a framework intended to ensure that autonomous systems[b] make certifiably ethically correct decisions. In particular, when no completely ethical decision is available (i.e., each possible decision will violate at least one ethical principle), they formally verified that agents using their framework will always pick the "least unethical" choice available. They achieve this verification using exhaustive model checking over the configuration of the world state as well as the ethical considerations in play.

However, the approach of Dennis et al. has several flaws that our work attempts to remedy. We next quickly summarize these flaws (and simultaneously, the desiderata for our framework) in order of increasing significance.

First, the exhaustive model checking process is very slow. In their first scenario, the process took four days of computation time to verify. We desire agents which can verify ethical behavior on the order of seconds, in order to enable real-time usage.

Second, the formalism used to model their agents, a Belief-Desire-Intention (BDI) language, is too inexpressive. The beliefs in their system are "ground first order formulae" [2]. Hence, nested belief is impossible; e.g. "Alice believes that Bob believes that Alice believes s", where s is itself a declarative sentence, cannot be expressed. We desire a highly expressive framework, which can express not only nested belief, but arbitrarily deep nested statements containing other modalities, such as perception, knowledge, and obligation (e.g. "John *believes* that he is *obligated* to perform action α.").

Third, Dennis et al. have no conception of *uncertainty* in their framework. Specifically, there is no way to specify that one plan is more likely to violate an ethical principle than another plan. We envision a framework that can formally model and reason about uncertainty of ethical violations, in order to select between two plans which have the potential to violate the same ethical principles, but perhaps at different levels of likelihood.

The fourth and final defect plaguing the approach of Dennis et al. is that formal verification based on model checking, and hence by definition on some model theory, which is the approach they take, is infeasible, for various reasons that are detailed below in a separate section (§4). Moreover, the approach to formally verifiying the ethical (or for that matter legal or prudential) correctness of an artificial agent we have invented and follow, which dates back now approaching two decades (e.g. see [5]) is applicable not just to artificial agents, but to computer programs in general; this we explain in the relevant section.

We next discuss a use case, the full solution to which demands that each of these four flaws be addressed; the case is hence one that the system of Dennis et al. [2] cannot handle.

1.2. *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

[b]We use the term 'system' in reference to the work of Dennis et al. as this is the term they use in their own work. However, in keeping with the terminology of standard textbooks in AI [3,4], we use the term 'agent' in our work.

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]."[c] 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 [7].

We are interested in constructing an AI agent which could reason in a way similar to how Sullenberger did. Three properties of our work (which, as mentioned in the prior section, are absent from Dennis et al. [2]), allow our agent to make the right decision. First, it needs to be able to determine and verify the correct decision in a matter of seconds. Second, it needs to be able to express and reason with nested beliefs (in order to express the captain's beliefs about the air traffic controller's beliefs). Finally, it needs to be able to differentiate between options which have the potential to violate the same ethical principles, but with different levels of likelihood.

In particular, consider the following table, which ranks the major potential ethical violations of Captain Sullenberger's landing options in a similar fashion to the presentation of scenarios in Dennis et al. [2]:

Table 1: Ethical concerns violated by each plan and their corresponding rank (i.e. relative significance).

Rank	Land in Hudson	Land at LGA	Land at TEB
3	Do not harm passengers	Do not harm passengers	Do not harm passengers
2	Do not collide with boats on the water	Do not collide with airport infrastructure	Do not collide with airport infrastructure
1	Do not damage own aircraft	Do not damage own aircraft	Do not damage own aircraft

Each plan has identical ethical violations (that is, they each cause one level-3, one level-2, and one level-1 violation); hence, their system would not be able to select one, as under their framework, all three options are equally unethical. Our framework, focusing on minimizing the likelihood of harm of passengers[d] (or equivalently, as we will pose it in §3, maximizing the likelihood of a safe landing), is able to provably verify — that is, produce a formal argument — that landing in the Hudson is the least unethical option.[e]

1.3. *Contributions*

The chief contributions of this paper are two-fold: (1) we show that our agents can solve the problems presented in Dennis et al. [2] (§2.1.1), and (2) show that it can solve a new class of problems not reachable by the work of Dennis et al. (§3.2). We next present the technical preliminaries which enable these contributions.

[c]This quote, recorded by the in-flight cockpit voice recorder (CVR), was retrieved from the NTSB Accident Report [6].

[d]As was likely the chief concern of Captain Sullenberger.

[e]It's important to note that our approach can handle *any* number of principles that are relevant to and active in the decision-making of an artificial agent in an ethically or legally charged situation. A full description of the range of the nature of such principles, taken from ethical/legal theories, and ethical/legal codes that are based upon these theories, is out of scope here; the reader is directed to [8] for details, as well as for a defense and description of how and why these principles must be designed for and engineered into the operating-system level of an artificial agent if this agent is to be ethically/legally correct when deployed.

2. Preliminaries

2.1. *The Formal System*

Our[f] approach to formally capturing ethics so as to install it in an artificial agent has long been grounded in the use of *cognitive calculi* (used e.g. in [9,10]). In short, a cognitive calculus is a multi-operator intensional logic built to capture all propositional attitudes in human cognition. (For information about such attitudes, see [11]; for a wonderful catalogue of all the major categories of human cognition, from *perceiving* to *fearing* to *remembering* to *saying* and beyond, see [12].) While this short paper does not allow for a full discussion of precisely what a cognitive calculus is, the interested reader is pointed to Appendix A in Bringsjord et al. [13].

For the purposes of this paper, it's specifically important to note that a cognitive calculus consists of *essentially* two components: (1) multi-sorted n-order logic with modal operators for modeling cognitive attitudes (e.g. knowledge $\mathbf{K}$, belief $\mathbf{B}$, and obligation $\mathbf{O}$[g]) and (2) inference schemata that — in the tradition of proof-theoretic semantics — express the semantics of the modal operators. In particular, we will utilize the Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) in the work described herein. The inference schemata of $\mathcal{DCEC}$ are shown in the box titled Inference Schemata.[h] The signature (including the types and grammar of the formal language) is in Appendix A.

Inference Schemata

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

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

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

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

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

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

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

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

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

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

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

[f] This collective refers to the RAIR Lab, of which the first three authors are members.

[g] See Appendix A for the rest of the modal operator descriptors.

[h] This is as good a place as any to inform the reader that as a matter of fact there are different dialects of the cognitive calculus shown in the box here, and we are using a particular dialect herein. What we are showing is really a particular cognitive calculus in a sub-family of deontic cognitive event calculi. Minor variations in signatures and collections of inference schemata give rise to different members of the sub-family, but details regarding these variations are not important in the present paper.

2.1.1. *A Proof of the "Fuel low" Scenario of Dennis et al.*

Using the $\mathcal{DCEC}$, we can construct a proof of the "Fuel low" scenario from Dennis et al. [2]. The scenario states that an unmanned aircraft (ua)[i] has run out of fuel and must make an emergency landing. The agent is given three options: land on an empty public road (*road*), a field with overhead power lines (*power*), or a field with people (*people*). Associated with each plan is a multiset of ethical principle violations, including "do not collide with people" (level 5) and "do not damage own aircraft" (level 1).

Due to space limitations, we exclude our implementation of their Ethical Plan Order — which determines the ethical ordering of plans based on their violations — and include the rest of the proof (given the orderings as axioms using the LessUnethical (LU) relation) in Appendix B. A theorem prover for $\mathcal{DCEC}$ — ShadowProver [14] — was able to generate a proof of the goal in 4.5 seconds.

2.2. *The Uncertainty System*

$\mathcal{DCEC}$ is purely deductive and has no formalisms for quantifying uncertainty. Therefore we will employ "strength factors", a nascent framework for formalizing uncertainty in quantified modal logics, first presented in Govindarajulu & Bringsjord [15]. Strength factors can be viewed as a formalization of Chisholm's epistemology [16], 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 $\phi \succ_t^a \psi$, and is read "ϕ is more reasonable than ψ to agent a at time t". 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.

$$(\phi \succ_t^a \psi_1) \text{ and } (\phi \succ_t^a \psi_2) \Rightarrow (\phi \succ_t^a \psi_1 \wedge \psi_2) \qquad [\mathbf{C}_{\wedge_1}]$$

Govindarajulu & Bringsjord [15] also provide a three clause definition for the 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.

In the next section, we provide another definition of the relation, custom designed for deciding between potential options during an emergency landing.

3. The Uncertainty System, Applied to Emergency Landing Scenarios

Our new definition of the reasonableness relation is cognitively 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 [7].

Domain-Specific Reasonableness

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

[i]The parenthesized labels in this section reference the $\mathcal{DCEC}$ proof in Appendix B.

This definition states that it is more reasonable for agent (or pilot) a to land at ϕ at time t[j] 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. [7], 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.

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:

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

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:[k]

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

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

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

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:

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

3.1. *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.

An Ethical Principle for Scenarios with Uncertain Ethical Outcomes

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

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

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

[k] In the United States, the right to disregard Air Traffic Control in an emergency is set out in §91.123 of the Code of Federal Regulations (see https://www.law.cornell.edu/cfr/text/14/91.123).

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[l] at ϕ.

3.2. *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.[m] 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.[n]

Sub-Argument 1

$$\mathbf{S}(atc, capt, t_1, \text{Land}(capt, t_1, lga_{13}))$$
$$\therefore \mathbf{B}\big(capt, t_1, \mathbf{B}(atc, t_1, \text{Land}(capt, t_1, lga_{13}))\big) \qquad [I_{12}]^{o} \checkmark$$
$$\therefore \mathbf{B}^1(capt, t_1, \text{Land}(capt, t_1, lga_{13})) \qquad [\mathbf{B}^1\text{-def}] \checkmark$$

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

$$\mathbf{P}(capt, t_2, \text{Reachable}(capt, t_2, teb) \wedge \neg\text{Reachable}(capt, t_2, lga_{13}))$$
$$\mathbf{P}(capt, t_2, \text{Reachable}(capt, t_2, hud) \wedge \text{safety}(capt, t_2, teb) > \text{safety}(capt, t_2, hud))$$
$$\therefore \text{Land}(capt, t_2, teb) \succ_{t_2}^{capt} \text{Land}(capt, t_2, lga_{13}) \qquad [\succ_t^a\text{-def}] \checkmark$$
$$\therefore \mathbf{B}^2(capt, t_2, \text{Land}(capt, t_2, teb)) \qquad [\mathbf{B}^2\text{-def}] \checkmark$$

Finally[p], 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

$$\mathbf{S}(atc, capt, t_3, \text{Land}(capt, t_3, teb_1))$$
$$\therefore \mathbf{B}\big(capt, t_3, \mathbf{B}(atc, t_3, \text{Land}(capt, t_3, teb_1))\big) \qquad [I_{12}] \checkmark$$
$$\therefore \mathbf{B}^1(capt, t_3, \text{Land}(capt, t_3, teb_1)) \qquad [\mathbf{B}^1\text{-def}] \checkmark$$

[l] Note the lowercase "*land*" used here is an ActionType, as opposed to the capitalized "Land", which is a predicate. For the full list of types, see Appendix A.

[m] Information regarding the actions of the Captain and Air Traffic Control during the event was retrieved from the NTSB Accident Report [6].

[n] The ATC also later suggested runway 4 at LaGuardia. We leave this detail out due to space limitations. The omission does not impact the main thread of reasoning.

[o] Defined in the box titled Inference Schemata in §2.1.

[p] 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.

$$\mathbf{P}(capt, t_3, \neg\text{Reachable}(capt, t_3, \{lga_{13}, teb_1\}))$$
$$\mathbf{P}(capt, t_3, \text{Reachable}(capt, t_3, hud))$$
$$\therefore \text{Land}(capt, t_3, hud) \succ_{t_3}^{capt} \text{Land}(capt, t_3, teb_1) \quad [\succ_t^a \text{-def}] \checkmark$$
$$\therefore \mathbf{B}^2(capt, t_3, \text{Land}(capt, t_3, hud)) \quad [\mathbf{B}^2\text{-def}] \checkmark$$

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

$\mathbf{K}(capt, t_3, \mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4)))$	$[I_{EP}] \checkmark$
$\mathbf{P}(capt, t_3, emergency)$	[Given]
$\mathbf{K}(capt, t_3, emergency)$	$[I_1] \checkmark$
$\mathbf{B}(capt, t_3, emergency)$	$[I_2] \checkmark$
$\mathbf{B}(capt, t_3, \mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4)))$	$[I_2] \checkmark$
$\mathbf{O}(capt, t_3, emergency, happens(action(capt, land(hud)), t_4)))$	$[I_4] \checkmark$
$\mathbf{K}(capt, t_3, \mathbf{I}(capt, t_3, happens(action(capt, land(hud)), t_4)))$	$[I_{14}] \checkmark$
$\mathbf{I}(capt, t_3, happens(action(capt, land(hud)), t_4)))$	$[I_4] \checkmark$
$\mathbf{P}(capt, t_3, happens(action(capt, land(hud)), t_4)))$	$[I_{13}] \checkmark$
$\mathbf{K}(capt, t_3, happens(action(capt, land(hud)), t_4)))$	$[I_1] \checkmark$
$happens(action(capt, land(hud)), t_4)))$	$[I_4] \checkmark$

A nascent automated reasoner called ShadowAdjudicator[q] (building on top of the aforementioned ShadowProver) — 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.

4. A Superior Way to Formally Verify Ethical Reasoning

As noted above, Dennis et al. [2] adopt, and indeed promote, a model-theoretic approach to the verification of ethical decisions (or, as they say, "choices") on the part of autonomous agents. For them, formal verification is quite literally *identified* with model checking; for they write on p. 2, italics theirs: "... *formal verification*, more precisely *model checking*, ..." While we commend Dennis et al. for their pursuit of formal verification, for numerous reasons a superior approach appears to be an inference-theoretic (which builds upon proof-theoretic semantics) one that offers a comprehensive paradigm not only for the verification of autonomous artificial agents that ethically reason and decide, but *for program verification in general*. There is thus more at stake here than simply a quarrel regarding model-based versus proof-based methodology, or whether verification happens offline or online. Before briefly explaining our approach to formal verification and its chief virtues, we first unpack a bit two drawbacks (from among others that are out of scope) of the model-theoretic approach that we merely adumbrated earlier (in §1.1). Our assessment is based upon the

[q] Source code available at: https://github.com/RAIRLab/ShadowAdjudicator.

assumption, confessedly non-negotiable for us, that autonomous artificial agents in play in the current paper are *logicist* in nature; i.e., they compute functions from percepts to actions (including decision-making) by reasoning that is specifically formalized in and with inference schemata in formal logics (not model-theoretic machinery).[r] This means that the very nature of artificial agents is bound up inextricably with step-by-step reasoning, where steps are sanctioned by inference schemata. Here, now, are the aforementioned drawbacks afflicting the model-theoretic approach, unpacked at least to a degree:

- *The Expressivity of Cognitive Calculi Outstrip Models/Semantics.* It has long been appreciated that a given formal language $\mathbf{L}$ for a formal logic and a set of inference schemata (the formulae in which are from $\mathbf{L}$) can far outstrip any available model-theoretic framework. For example, consider a single formula ϕ in some $\mathbf{L}$ that expresses an English sentence such as $s^\star :=$ "The pilot Alice knows now that Bob in ATC believes it is ethically forbidden for Alice to later say s when she has perceived that s'." Here, both s and s' are themselves declarative sentences (but each led by a minuscule English letter). The model-theoretic side of e.g. any BDI logic is insufficiently expressive to represent $s^\star$ and its components. This single sentence, which in the real world is true time and time again (with suitable reassignments to the constants 'Alice' and 'Bob,' and instantiations of the variables s and s') demands elements provided by: modal logics for possibility and necessity, epistemic logics, temporal/tense logics, logics for communication, logics for perception, and so on. There simply is no class of models for formulae that draw from all these families of logics at once; hence model checking is impossible now and in the foreseeable future as a way to verify that formulae representing the likes of $s^\star$ hold in certain contexts.[s] An option that will occur to some readers is to simply throw in the towel on finding some single model-theoretic framework that covers essentially all the families of intensional/modal logics that need to be tapped in order to express the likes of $s^\star$, in favor of using only some extensional logic (e.g. first-order logic). Unfortunately, this route quickly produces inconsistency, for reasons detailed by the proofs provided in [19].

- *Model Theory and Models Unstoppably Reduce to Reasoning Verifiable Only via Inference Schemata.* It is easy enough to see why those in the proof-theoretic-semantics tradition that rejects model theory and models do so because of the reducibility of model-theoretic structures to proof-theoretic ones. Here is a quick particularization of this reducibility argument.
  Consider a simple biconditional

  $$\beta := (p_i \wedge p_j) \leftrightarrow (p_j \wedge p_i)$$

  in the propositional calculus ($\mathcal{L}_{\text{PC}}$). Let ν be any customary truth-value assignment (t.v.a.) of TRUE or FALSE to every relevant propositional atom p_k. We say that any formula in the propositional calculus that's true on every t.v.a. is a *validity*. We can now ask whether β is true on a given t.v.a., and we can also ask if β is a validity. Take the second of these two queries. What is the answer? Of course, the correct answer is an affirmative one. But how does an agent know this? An agent,

[r] That artificial agents compute such functions is the orthodox conception affirmed in the major textbooks for and overview of AI (e.g. [3,4,17]),

[s] The source of the problem, as explained in [18], wherein the first cognitive calculus was presented and implemented, began when possible-world semantics was extended from a reasonable way to make sense formally of 'possibly' and 'necessarily,' to a way to try to make sense formally also of 'knows' and 'believes'.

including us, knows this because it can use the relevant machinery of the formal truth-functional semantics of $\mathcal{L}_{\text{PC}}$ to prove

$$\nu \models \beta \tag{1}$$

and a key part of this machinery is this familiar clause:

$$\nu \models \phi \rightarrow \psi \text{ iff if } \nu \models \phi \text{ then } \nu \models \psi \tag{2}$$

Notice the occurrence in (2) of 'iff' and 'if' and 'then.' These terms are part of what we have called the relevant "machinery". *They are nowhere defined truth-functionally or model-theoretically at all*; they are meta-logical connectives. Now, label this machinery '$\mathcal{M}$.' $\mathcal{M}$ includes (2), and also the background-logic proof theory $\Pi_{\mathcal{M}}$ for how 'iff' and 'if' and 'then' are to be reasoned with deductively (e.g. we have on hand *modus ponens*). So the reducibility in question has happened in front of our eyes. To make it even clearer, abbreviate the assertion that the combination of (1) and $\mathcal{M}$ can be used to prove (2) as

$$(\mathcal{M} \ + (2)) \ \vdash_{\Pi_{\mathcal{M}}} (1) \tag{3}$$

What we have just seen is that in $\mathcal{L}_{\text{PC}}$ getting to (1) reduces to (3). But this generalizes to every single truth-semantic target in $\mathcal{L}_{\text{PC}}$, for every cognizer coming to know that this target holds. In fact, since first-order logic $\mathcal{L}_1$ only augments $\mathcal{L}_{\text{PC}}$ with additional machinery for quantification in the same style, establishing model theoretic assertions of truth in $\mathcal{L}_1$, given what we have just seen, reduces to proof-theoretic semantics.[t]

In stark contrast, here's how easy and generalizable the situation is when our inference-theoretic approach is taken. Before the reader moves to the next paragraph, please look back and observe our repeated use of ✓above for each formal inference that took place.

Let $\mathfrak{a}$ be some artificial agent that makes some ethical decision (or choice) at time t. We define "an agent $\mathfrak{a}$'s making an ethical decision at time t" as believing some proposition of the form $\Omega\phi$ at t, where ϕ is some formula in some formal language of some cognitive calculus ($\approx$ some formal intensional logic), and Ω is specifically a deontic modal operator. This means that we can write that some $\mathfrak{a}$ has made an ethical decision at t this way:

$$\exists \phi, \Omega[\mathbf{B}(\mathfrak{a}, t, \Omega(\phi))].$$

An instantiation of this gives us a particular ethical decision; e.g.

$$\mathbf{B}(\mathfrak{a}, t, \mathbf{O}(\phi^{\star}))$$

says that the agent has made an ethical decision that $\phi^{\star}$ is obligated to be the case. If we wish to leave open what deontic operator is involved in a particular ethical decision, we can obviously write:

$$(+) \quad \mathbf{B}(\mathfrak{a}, t, \Omega(\phi^{\star}))$$

We shall let d be a variable ranging over formulae in any of these forms.

Now, how do we achieve formal verification of an ethical decision d? Doing so is effortless, as long as every ethical decision for an artificial agent is made because it is inferred as the final proposition in a "natural" argument α or proof π that is in accordance with the inference

[t] In considering rich intensional logics (as opposed to $\mathcal{L}_{\text{PC}}$ and $\mathcal{L}_1$, both extensional) and the meaning of their formulae, we likewise look for the conditions under which these formulae are provable.

schemata in play, which is exactly how we proceeded above (and are still proceeding and long have been for machine/robot/AI ethics; see e.g. [20]), for then we can simply follow an approach to program verification presented and advocated e.g. in [21].[u] Let $\Sigma \rightsquigarrow_{\alpha} \phi$ say that there is an argument α from undischarged suppositions (formulae) Σ to ϕ, and $\Sigma \rightsquigarrow_{\pi} \phi$ be the counterpart for a proof. Let $\mathbb{C}$ be a checker for arguments/proofs that receives one of these and returns either VALID = ✓ or INVALID; and suppose that this checker has been classically verified (which is easy: two pages of code, at most). This approach is generalizable, for note that uncertainty as we have systematically used it above creates no problems whatsoever, the simple reason being that inferences must still accord with inference schemata, and if they do all the way through, the overall conclusion/decision is certified.

We are thus done: An ethical decision d is verified (relative to given suppositions/premises) iff we have

$$\Sigma \rightsquigarrow_{\alpha/\pi} d \text{ and } \mathbb{C}(\alpha/\pi) = \text{VALID}.$$

Note, finally, that nothing really changes if we move from decisions to actions: We can focus in that case on what our agent intends (using the **I** modal operator in a cognitive calculus such as $\mathcal{DCEC}$) to do, and the case where the agent in fact does do what it/he/she intends to do.

4.1. *Limitations*

We certainly do not claim that our reasoning agent (or any AI currently in development today in our respective laboratories) is sufficiently capable to take the place of a human pilot. Nonetheless, what we hope to have conveyed is that in order to develop AI which could one day make autonomous, human-level, life-and-death decisions within aircraft piloting systems, the AI in question must have the ability to formally reason, by certifiably correct inferences, about declarative content represented in a cognitive calculus. In particular, it is *crucial* that this artificial agent be able to reason in a manner that dynamically takes explicit account of levels of uncertainty in ethically charged circumstances.

5. Conclusion & Future Work

We highlighted four flaws in the impressive, inspiring prior work of Dennis et al. [2] on autonomous systems which have to make decisions where each decision violates at least one ethical principle. We then presented our framework, which is able to solve the problems presented in the prior work, as well as a new class of problems unreachable by the prior work; specifically, problems in which all options are equally unethical, but have different likelihoods of violating those principles. Further work in this area includes modeling other types of uncertainty, such as the likelihood that completing a plan will actually achieve the goal.

Acknowledgements

The authors are grateful to AFOSR for their support of our development of cutting-edge reasoning systems for high levels of computational intelligence. We also thank ONR, both for their support of our current work on belief adjudication (as part of our vision to surmount Arrow's Impossibility Theorem and related theorems), and for their past support of our r&d in robot/machine ethics (primarily under a MURI to advance the science and engineering of moral competence in robots; PI M. Scheutz, Co-PI B. Malle, Co-PI S. Bringsjord).

[u] See also [22–24]. The first of these gives the first published roots of this novel approach to program verification, in work led by Arkoudas on proof checking [22].

References

1. S. Bringsjord, A 21st-Century Ethical Hierarchy for Humans and Robots: $\mathcal{EH}$, in *A World With Robots: International Conference on Robot Ethics (ICRE 2015)*, eds. I. Ferreira, J. Sequeira, M. Tokhi, E. Kadar and G. Virk (Springer, Berlin, Germany, 2015) pp. 47–61. This paper was published in the compilation of ICRE 2015 papers, distributed at the location of ICRE 2015, where the paper was presented: Lisbon, Portugal. The URL given here goes to the preprint of the paper, which is shorter than the full Springer version. http://kryten.mm.rpi.edu/SBringsjord_ethical_hierarchy_0909152200NY.pdf.
2. L. Dennis, M. Fisher, M. Slavkovik and M. Webster, Formal Verification of Ethical Choices in Autonomous Systems, *Robotics and Autonomous Systems* **77**, 1 (2016) http://dx.doi.org/10.1016/j.robot.2015.11.012.
3. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, New York, NY, 2020). Fourth edition.
4. G. Luger, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (6th Edition)* (Pearson, London, UK, 2008).
5. S. Bringsjord, K. Arkoudas and P. Bello, Toward a General Logicist Methodology for Engineering Ethically Correct Robots, *IEEE Intelligent Systems* **21**, 38 (2006) http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf.
6. D. A. Hersman, C. A. Hart and R. L. Sumwalt, *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) (2010).
7. S. Paul, F. Hole, A. Zytek and C. A. Varela, Flight trajectory planning for fixed wing aircraft in loss of thrust emergencies, in *Dynamic Data-Driven Application Systems (DDDAS 2017)*, (Cambridge, MA, 2017).
8. N. Govindarajulu, S. Bringsjord, A. Sen, J. Paquin and K. O'Neill, Ethical Operating Systems, in *Reflections on Programming Systems*, eds. L. De Mol and G. Primiero, Philosophical Studies, Vol. 133 (Springer, 2018) pp. 235–260 http://kryten.mm.rpi.edu/EthicalOperatingSystems_preprint.pdf.
9. N. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, ed. C. Sierra (International Joint Conferences on Artificial Intelligence, 2017).
10. S. Bringsjord, N. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014). IEEE Catalog Number: CFP14ETI-POD. Papers from the *Proceedings* can be downloaded from IEEE at URL provided here.
11. M. Nelson, Propositional Attitude Reports, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2015) https://plato.stanford.edu/entries/prop-attitude-reports.
12. M. Ashcraft and G. Radvansky, *Cognition* (Pearson, London, UK, 2013). This is the 6th edition.
13. S. Bringsjord, N. S. Govindarajulu, J. Licato and M. Giancola, Learning *Ex Nihilo*, in *GCAI 2020. 6th Global Conference on Artificial Intelligence (GCAI 2020)*, eds. G. Danoy, J. Pang and G. Sutcliffe, EPiC Series in Computing, Vol. 72 (EasyChair, 2020).
14. N. Govindarajulu, S. Bringsjord and M. Peveler, On Quantified Modal Theorem Proving for Modeling Ethics, in *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements (ARCADE 2019)*, eds. M. Suda and S. Winkler, Electronic Proceedings in Theoretical Computer Science, Vol. 311 (Open Publishing Association, Waterloo, Australia, 2019) pp. 43–49. The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/. http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf.
15. N. S. Govindarajulu and S. Bringsjord, Strength Factors: An Uncertainty System for Quantified Modal Logic, in *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning" (LFU-2017)*, eds. V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade and G. Qi (Melbourne, Australia, 2017).
16. R. Chisholm, *Theory of Knowledge 3rd ed* (Prentice-Hall, Englewood Cliffs, NJ, 1987).
17. S. Bringsjord and N. S. Govindarajulu, Artificial Intelligence, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2018) https://plato.stanford.edu/entries/artificial-intelligence.
18. K. Arkoudas and S. Bringsjord, Propositional Attitudes and Causation, *International Journal of Software and Informatics* **3**, 47 (2009) http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_

041709.pdf.
19. S. Bringsjord and N. S. Govindarajulu, Given the Web, What is Intelligence, Really?, *Metaphilosophy* **43**, 361 (2012), This URL is to a preprint of the paper http://kryten.mm.rpi.edu/SB\NSG\Real\Intelligence\040912.pdf.
20. K. Arkoudas, S. Bringsjord and P. Bello, Toward Ethical Robots via Mechanized Deontic Logic, in *Machine Ethics: Papers from the AAAI Fall Symposium; FS–05–06*, (American Association for Artificial Intelligence, Menlo Park, CA, 2005) pp. 17–23 http://www.aaai.org/Library/Symposia/Fall/fs05-06.php.
21. S. Bringsjord, A Vindication of Program Verification, *History and Philosophy of Logic* **36**, 262 (2015), This url goes to a preprint. http://kryten.mm.rpi.edu/SB_progver_selfref_driver_final2_060215.pdf.
22. K. Arkoudas and S. Bringsjord, Computers, Justification, and Mathematical Knowledge, *Minds and Machines* **17**, 185 (2007) http://kryten.mm.rpi.edu/ka_sb_proofs_offprint.pdf.
23. S. Bringsjord, Logicist Remarks on Rapaport on Philosophy of Computer Science$^+$, *Newsletter on Philosophy and Computers* **18**, 28 (2018), The URL here is to a preprint only. http://kryten.mm.rpi.edu/SBonBR.pdf.
24. S. Bringsjord, Computer Science as Immaterial Formal Logic, *Philosophy & Technology* (August 2019), DOI: https://doi.org/10.1007/s13347-019-00366-7 http://kryten.mm.rpi.edu/CompSciAsImmaterialFormalLogicPreprint.pdf.

Appendix A. The Deontic Cognitive Event Calculus Signature

Signature

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

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

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

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

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

Assumptions:

$\mathbf{K}\big(ua, now, \mathrm{LU}(road, power)\big)$ [1]

$\mathbf{K}\big(ua, now, \mathrm{LU}(road, people)\big)$ [2]

$\mathbf{K}\Big(ua, now, \big(\mathrm{LU}(road, power) \wedge \mathrm{LU}(road, people)\big) \rightarrow \mathrm{BestOption}(road)\Big)$ [best-option-axiom]

$\mathbf{K}\Big(ua, now, \forall x\ \mathrm{BestOption}(x) \rightarrow \mathbf{O}\big(ua, now, crisis, happens\big(action(ua, land(x)), next\big)\big)\Big)$ [best-implies-obligated]

$\mathbf{B}(ua, now, crisis)$ [agent-awareness]

$now < next$ [order-of-time]

Goal: $happens\big(action(ua, land(road)), next\big)$

Sub-Proof 1: $\mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)$

$\mathbf{K}\big(ua, now, \mathrm{LU}(road, power)\big)$ [1]

$\Rightarrow \mathbf{K}\big(ua, now, \mathrm{LU}(road, power)\big) \wedge \mathbf{K}\big(ua, now, \mathrm{LU}(road, people)\big)$ [$\wedge$-Intro]

$\Rightarrow \mathrm{LU}(road, power) \wedge \mathrm{LU}(road, people)$ [I_4]

$\Rightarrow \mathrm{BestOption}(road)$ [best-option-axiom with I_4]

$\Rightarrow \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)$ [best-implies-obligated with I_4]

Sub-Proof 2: $\mathbf{B}\Big(ua, now, \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)\Big)$

Chain 1:

$\mathbf{K}\big(ua, now, \mathrm{LU}(road, power)\big)$ [1]

$\Rightarrow \mathbf{K}\big(ua, now, \mathrm{LU}(road, power)\big) \wedge \mathbf{K}\big(ua, now, (\mathrm{LU}\ road, people)\big)$ [$\wedge$-Intro]

$\Rightarrow \mathbf{B}\big(ua, now, \mathrm{LU}(road, power)\big) \wedge \mathbf{B}\big(ua, now, (\mathrm{LU}\ road, people)\big)$ [I_2]

$\Rightarrow \mathbf{B}\big(ua, now, \mathrm{LU}(road, power) \wedge (\mathrm{LU}\ road, people)\big)$ [I_{11b}]

Chain 2:

$\mathbf{K}\Big(ua, now, \big(\mathrm{LU}(road, power) \wedge \mathrm{LU}(road, people)\big) \rightarrow \mathrm{BestOption}(road)\Big)$ [best-option-axiom]

$\Rightarrow \mathbf{B}\Big(ua, now, \big(\mathrm{LU}(road, power) \wedge \mathrm{LU}(road, people)\big) \rightarrow \mathrm{BestOption}(road)\Big)$ [I_2]

$\Rightarrow \mathbf{B}\big(ua, now, \mathrm{BestOption}(road)\big)$ [I_{11a} with Chain 1]

Chain 3:

$\mathbf{K}\Big(ua, now, \forall x\ \mathrm{BestOption}(x) \rightarrow \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(x)), next\big)\Big)\Big)$ [best-implies-obligated]

$\Rightarrow \mathbf{B}\Big(ua, now, \forall x\ \mathrm{BestOption}(x) \rightarrow \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(x)), next\big)\Big)\Big)$ [I_2]

$\Rightarrow \mathbf{B}\Big(ua, now, \mathrm{BestOption}(road) \rightarrow \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)\Big)$ [$\forall$-Elim with $road$]

$\Rightarrow \mathbf{B}\Big(ua, now, \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)\Big)$ [I_{11a} with Chain 2]

Proof of Goal: $happens\big(action(ua, land(road)), next\big)$

$\mathbf{B}(ua, now, crisis)$ [agent-awareness]

$\mathbf{B}(ua, now, crisis) \wedge \mathbf{B}\Big(ua, now, \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)\Big)$ [$\wedge$-Intro]

$\wedge \mathbf{O}\Big(ua, now, crisis, happens\big(action(ua, land(road)), next\big)\Big)$

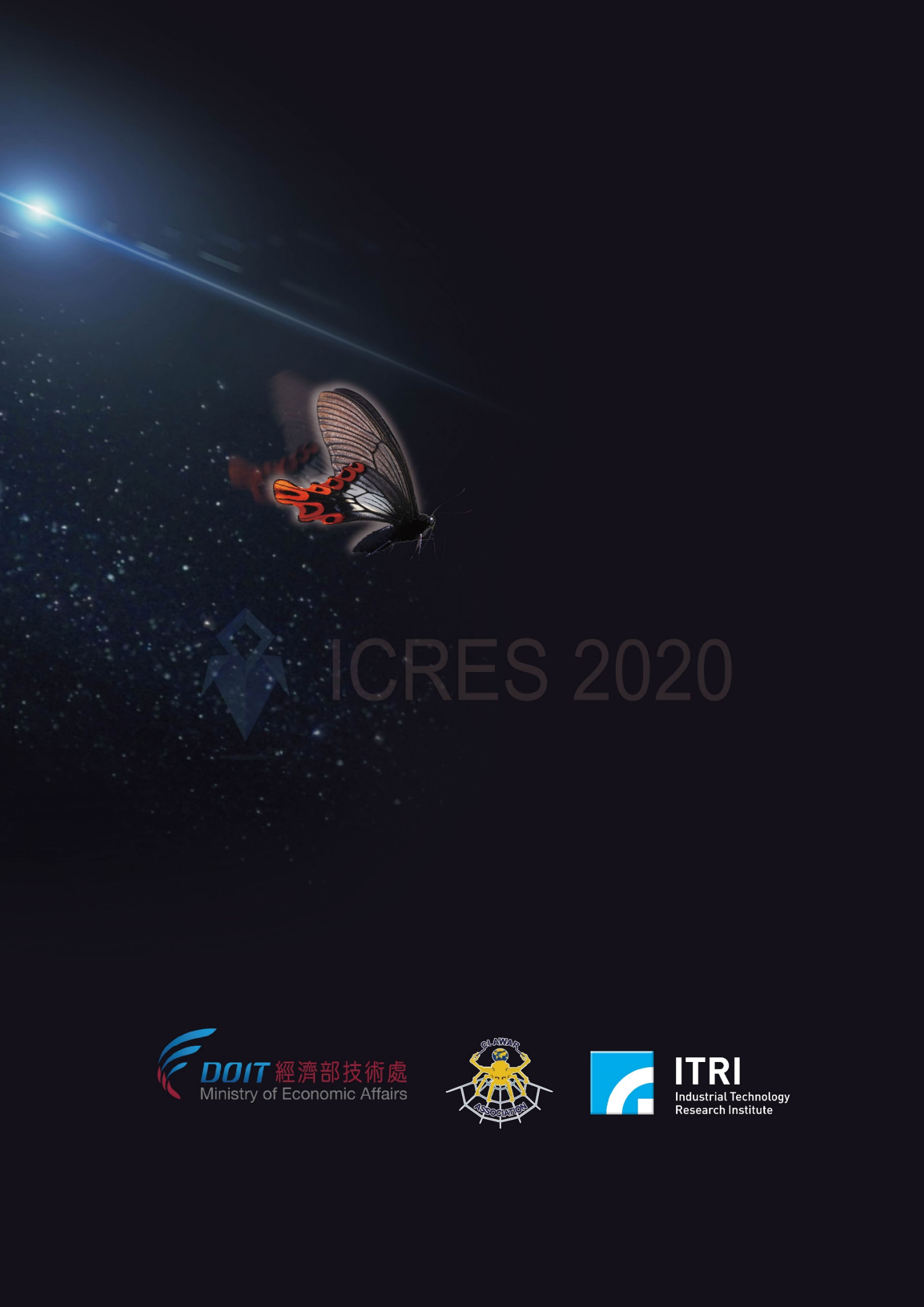
$\mathbf{K}\Big(ua, now, \mathbf{I}\Big(ua, now, happens\big(action(ua, land(road)), next\big)\Big)\Big)$ [I_{14}]

$\mathbf{I}\Big(ua, now, happens\big(action(ua, land(road)), next\big)\Big)$ [I_4]

$\mathbf{P}\Big(ua, now, happens\big(action(ua, land(road)), next\big)\Big)$ [I_{13}]

$\mathbf{K}\Big(ua, now, happens\big(action(ua, land(road)), next\big)\Big)$ [I_1]

$happens\big(action(ua, land(road)), next\big)$ [I_4]


ICRES 2020
DOIT 經濟部技術處
Ministry of Economic Affairs
CLAWAR
ASSOCIATION
ITRI
Industrial Technology
Research Institute