ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

A SOLUTION TO AN ETHICAL SUPER DILEMMA VIA A RELAXATION OF THE DOCTRINE OF TRIPLE EFFECT

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

Rensselaer AI & Reasoning Lab[⋆†‡]
Department of Computer Science[⋆†]*; Department of Cognitive Science*[†]
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
E-mail: { *mike.j.giancola, selmer.bringsjord, naveen.sundar.g* } *@gmail.com*

We denote *ethical super dilemmas* as those ethical dilemmas which cannot be solved via any currently existing ethical principles or automated-reasoning technology. In particular, we analyze an ethical dilemma attributed to Bernard Williams, which we refer to as "Jim's Dilemma". After making clear that neither the Doctrine of Double Effect nor the more permissive Doctrine of Triple Effect enable one to sanction action in Jim's Dilemma, we present a novel relaxation of the Doctrine of Triple Effect, by which Jim's Dilemma can be solved. Moreover, we argue that Jim's Dilemma motivates further R&D on morally creative agents.

Keywords: Ethical reasoning; moral creativity.

1. Introduction

Human being inevitably encounter situations in which a decision is to be made and there is no single best decision. Specifically, in ethically-charged situations, we call these scenarios *ethical dilemmas*. In this paper, we define a trichotomy of ethical dilemmas, ranked by their relative difficulty. We then present two solutions to a problem in the most challenging category, which we call *ethical super dilemmas*.

The rest of the paper is as follows. Section 2 provides a review of several topics which lay the groundwork for the work herein. In section 3, we present our trichotomy of ethical dilemmas and an example dilemma in each partition. We then introduce a modification of the Doctrine of Triple Effect (§4) by which we solve an ethical super dilemma (§5). We then discuss future work and conclude.

2. Preliminaries

What follows are brief reviews of various topics necessary for understanding the main content of the paper. Readers may wish to selectively read only those subsections for which they do not have prior knowledge.

2.1. *Solving Ethical Problems*

What is required of a solution to an ethical problem, in our conception? Essentially, two components: first, a decision, and second, a formal proof (or argument) which can be mechanically verified. In particular, such a proof typically employs one or more ethical principles, and proves that some action α can be sanctioned by the principle(s).

In our approach, this is done by formalizing both the principle(s) and the dilemma in the language of a cognitive calculus, then using an automated reasoner to find a proof which shows that the action satisfies the constraints of the principle(s). In §2.6 and §2.7, we discuss two such principles which we have used in prior work [2,9] and which are relevant to the present paper.

2.2. *Cognitive Calculi*

Our approach to formally capturing ethics so as to install it in an artificial agent has long been grounded in the use of cognitive calculi [1–3]. In short, a cognitive calculus is a multi-operator quantified intensional logic built to capture all propositional attitudes in human cognition.[a] While a longer discussion of precisely what a cognitive calculus is is out of scope, the interested reader is pointed to Appendix A in Bringsjord et al. [5].

For 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 $\mathbf{O}$) 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 Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$) in the work described herein. We next review a predecessor of $\mathcal{IDCEC}$, the (deductive) Deontic Cognitive Event Calculus ($\mathcal{DCEC}$).

2.3. *Deontic Cognitive Event Calculus*

The Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) consists of a signature and a set of inference schemata. The signature includes the calculus' sorts, function signatures, and grammar. Most significantly, grammatical forms for modal operators (e.g. knowledge $\mathbf{K}$, belief $\mathbf{B}$) are specified. Also, an automated reasoner for $\mathcal{DCEC}$ — ShadowProver [6] — has been created, is available, and is under active development. For a more in-depth discussion of $\mathcal{DCEC}$, including the full signature and set of inference schemata, see Appendix B in Bringsjord et al. [5].

2.4. *Inductive Deontic Cognitive Event Calculus*

$\mathcal{DCEC}$ employs no uncertainty system (e.g., probability measures, *strength factors*, or likelihood measures) and hence is purely deductive. Therefore, as we wish to enable our agents to reason about situations involving uncertainty, we must ultimately utilize the *Inductive* $\mathcal{DCEC}$: $\mathcal{IDCEC}$.

In general, to go from a deductive to an inductive cognitive calculus, we require two components: (1) an uncertainty system, and (2) inference schemata that delineate the methods by which inferences linking formulae and other information can be used to build formally valid arguments. The uncertainty system we employ herein is *cognitive likelihood*, which we discuss in §2.5. As this paper will work at the level of proof/argument sketches, we do not present the inference schemata here. The interested reader can find a nascent set of inference schemata for $\mathcal{IDCEC}$ in [7].

2.5. *Cognitive Likelihood*

Our approach to quantifying the uncertainty of beliefs within cognitive calculi eschews traditional probability values in favor of *likelihood* values. The 11 likelihood values are shown in Table 1.

Likelihood values can be obtained in either of two ways; both ways immediately reveal that we take likelihood to be *subjective*. The first way is to take as primitive a cognitive binary relation on formulae from the perspective of a rational agent (e.g., ϕ is *more reasonable than* ψ), and then build up formally to the partial or total order in question. This approach is first formalized in [8] and is deployed in e.g. [1]. Another approach, the one taken here, is to independently justify each likelihood value by appeal to rational human-level cognition.

[a] E.g. perceiving, fearing, remembering, saying [4].

Table 1: The 11 Cognitive Likelihood Values

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

For example, that which is CERTAIN applies to propositions that a perfectly rational human-level cognizer would affirm as such — that 2+2=4 (Base-10), that $0 \neq 1$, and so on for any theorem that has been certifiably deduced from what is itself CERTAIN. Propositions are EVIDENT typically when they are given by immediate perception in the absence of conditions known to frequently cause illusory perception. For example, currently the lead author perceives his laptop's screen in front of him, and hence that there is such a screen in front of him is EVIDENT. For a longer discussion of Cognitive Likelihood, see [7].

2.6. *Doctrine of Double Effect*

The Doctrine of Double Effect ($\mathcal{DDE}$) is an ethical principle which sanctions some actions which have both positive and negative effects. Bringsjord & Govindarajulu previously formalized $\mathcal{DDE}$ in a cognitive calculus and used it to solve two variants of the Trolley Problem [2]. Informally, they specify that an action is $\mathcal{DDE}$-compliant iff:[b]

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [10], and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;
- $\mathbf{C}_4$ the bad effects are not used as a means to obtain the good effects.

2.7. *Doctrine of Triple Effect*

The Doctrine of Triple Effect ($\mathcal{DTE}$) relaxes some restrictions of $\mathcal{DDE}$, allowing it to sanction some actions which cannot be sanctioned by $\mathcal{DDE}$[c]. To do this, $\mathcal{DTE}$ employs the concepts of *primary* and *secondary* intentions. Peveler et al. [9] used Bratman's test for intentions [11] to define an intention as primary iff[d] the following conditions hold:

[b] If and only if.
[c] The astute reader will likely notice that a further relaxation of this kind is exactly what we intend to do herein to enable the solution of increasingly challenging ethical dilemmas.
[d] That is, an intention is *secondary* if any of the conditions do not hold.

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

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

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [10], and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action **primarily** intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not **primarily** intend any of the bad effects, **but may secondarily intend some of them**;
- $\mathbf{C}_4$ no **primarily** intended bad effects are used as a means to obtain the good effects, **but secondarily intended bad effects may be**.

3. A Trichotomy of Ethical Dilemmas

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

(1) *Simple ethical dilemmas* are those which can be solved using state-of-the-art automated reasoning/planning.
(2) *Standard ethical dilemmas* are those which require sophisticated ethical principles and automated reasoning to solve.
(3) *Ethical super dilemmas* are those which *cannot* be solved via any currently existing ethical principles or automated reasoning technology.

To illustrate this trichotomy, we give an example problem and solution in each partition.

3.1. *Simple Ethical Dilemmas*

Consider the Heinz Dilemma, as presented by Lawrence Kohlberg [12]:

The Heinz Dilemma

> In Europe, a woman was near death from a very bad disease, a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium for which a druggist was charging ten times what the drug cost him to make. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about half of what it cost. He told the druggist that his wife was dying, and asked him to sell it cheaper or let him pay later. But the druggist said, "No, I discovered the drug and I'm going to make money from it." So Heinz got desperate and broke into the man's store to steal the drug for his wife.

Should Heinz have stolen the drug? Or should he have not, and allowed his wife to die? While there is no single, universally correct answer, one can quite easily arrive at a solution once they have determined the relative priority of their ethical obligations. That is, if one values the principle that people deserve adequate health care over the principle that one

should not steal, then Heinz was right to steal the drug. If not, Heinz should not have stolen the drug. Both possible solutions (as well as potentially others) can be generated, along with verifiable proofs, by state-of-the-art automated planners.

3.2. *Standard Ethical Dilemmas*

Perhaps the most widely-discussed ethical dilemma, the Trolley Problem is a member of our second partition:

The Trolley Problem

In the classic scenario, illustrated in Figure 1, a trolley is going down a track towards two people. The trolley's brakes are not functioning, so if no action is taken, the trolley will kill the two people. There is a switch which would allow the trolley to switch to a branching track and avoid the two people, but it would cause the train to kill a single person stuck on the branch.

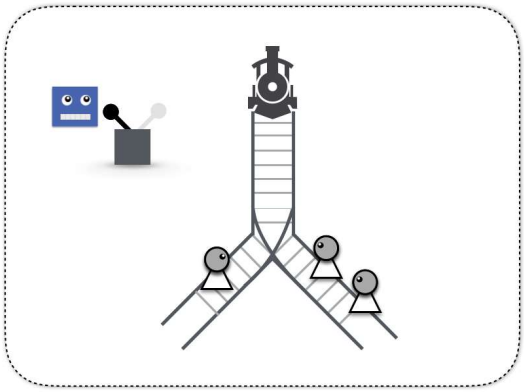

Fig. 1. The "Classic" Trolley Problem

There are several variants of the Trolley Problem. In the "Push Case", there is no switch or branching track, but there is a large person who, if pushed onto the track, will stop the train and prevent it from killing the two stuck on the track. In the "Loop Case", there is a switch which will send the trolley onto a track which will loop around and go back onto the main track. However, there is a large person on the loop who will be killed and stop the train before it loops back to the main track.

The classic Trolley problem, as well as these two variants, are all *Standard Ethical Dilemmas*. The classic and "Push Case" were solved[e] by utilizing the Doctrine of Double Effect [2], and the "Loop Case" was solved[f] via the Doctrine of Triple Effect [9].

3.3. *Ethical Super Dilemmas*

The following example, which will be the focal point of the rest of the present paper, is attributed to Bernard Williams [13]:

[e]Specifically, flipping the switch in the classic Trolley Problem is ethically permissible, whereas pushing the person onto the track in the "Push Case" is not.
[f]Flipping the switch in the "Loop Case" was shown to be ethically permissible.

Jim's Dilemma

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

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

However, as is required by our third partition, the authors know of no ethical principles which could sanction either decision (shoot or abstain) given the original constraints. In particular, Bedau [14] gives a detailed analysis showing that the decision to shoot cannot be sanctioned by the Doctrine of Double Effect. Put briefly, the murder of an innocent is a forbidden action, hence Jim shooting a hostage would violate the first clause of $\mathcal{DDE}$. Also, as this same clause is present in the Doctrine of Triple Effect, it too cannot sanction the shooting.

4. A Relaxation of the Doctrine of Triple Effect

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

- $\mathbf{C}_1^*$ if the action is forbidden, then the agent must believe it is *overwhelmingly likely* that:
  - $\mathbf{C}_{1.1}^*$ no possible action can achieve a higher utility;
  - $\mathbf{C}_{1.2}^*$ inaction has lower utility.

Let μ denote a utility function ranging over the set of possible actions. Then for an agent

a, we can formalize the notion that action α^* satisfies clause $\mathbf{C}_1^*$ using the $\mathcal{IDCEC}$ formula:[g]

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

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

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

We will first show that Jim shooting a hostage – should he choose to do so – is a secondary intention, as defined in §2.7. Recall that three clauses must hold in order for an intention to be *primary*. We shall show that one of these clauses – $\mathbf{D}_2$ – does not hold in this case.

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

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

- $\mathbf{C}_2$ the utility is positive (more hostages will be saved than slain);
- $\mathbf{C}_{3a}$ Jim only primarily intends to save the remaining 19 hostages;
- $\mathbf{C}_{3b}$ Jim secondarily intends to shoot one hostage;
- $\mathbf{C}_4$ Only a secondarily intended bad effect – shooting a hostage – is used as a means to obtain a good effect – saving the remaining 19 hostages.

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

5.1. *Two Possible Solutions*

First, consider the most pure realization of the dilemma[i]. Jim has three possible actions: (1) accept the captain's offer and shoot a hostage, (2) reject the captain's offer, or (3) attempt to defeat the captain and his guards. Based on a pure interpretation of the situation, we can assume that Jim believes it is *overwhelmingly likely* (= belief level 3) that (1) if Jim shoots a hostage, the other 19 will be set free, (2) if Jim does not shoot a hostage, all 20 will be killed, and (3) if Jim attempts to defeat the captain and his guards, Jim, along with all 20 hostages, will be killed.
We can formalize this in $\mathcal{IDCEC}$ using the following set of formulae:

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain\})$$
$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19)$$
$$\mathbf{B}^3(jim, \mu(abstain) = -20)$$
$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21)$$
$$Forbidden(shoot\_hostage)$$

[g] $\mathbf{B}^3(a, \dots)$ can be read as "Agent a believes it is *overwhelmingly likely* that ...".
[h] For reference, see §2.7.
[i] That is, we will only consider the options given in the original text of the dilemma, without extrapolating alternate possibilities.

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

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

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

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

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain, negotiate\})$$
$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19)$$
$$\mathbf{B}^3(jim, \mu(abstain) = -20)$$
$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21)$$
$$\mathbf{B}^2(jim, \mu(negotiate) > 0)$$
$$Forbidden(shoot\_hostage)$$

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

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

6. Future Work

Assuming Jim asserts the assumptions by which $\mathcal{DTE}_R$ sanctions his killing a hostage, he still has no ethically-grounded mechanism to select which one. Bedau [14] discusses the option of selecting at random. But by which ethical principle is this allowed? Bedau also discusses the possibility that a hostage might sacrifice themselves. If one did not, Jim could request a sacrifice. Would any of these options be ethical? What ethical principle could sanction them?

Also, we would obviously prefer an autonomous agent which could identify and pursue the option to negotiate rather than shooting a hostage (even if that is ethically permissible under the circumstances). An agent of this kind would need to be *morally creative*. The authors know of no agent framework enabling such a level of moral creativity, but see it as a pressing area of future R&D.

7. Conclusion

> We do not have an algorithm that yields a definite answer when all and only the relevant reasons are specified, or a morality machine into which we can type in the information about a given problem case, such as Jim's, then press a sequence of keys, and get a printout with the morally correct verdict. (pg. 95 of [14])

We still don't have a universal "morality machine", but what we have created is a mechanizable ethical principle by which Jim's Dilemma can be solved. We have also motivated further R&D into morally creative agents which can find "escape hatches" in ethically challenging scenarios.

References

1. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Ethical Reasoning for Autonomous Agents Under Uncertainty, in *Smart Living and Quality Health with Robots • Proceedings of ICRES 2020*, eds. M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang and A. Kaur (CLAWAR, London, UK, September 2020). Paper vailable at the URL given above. The ShadowAdjudicator system can be obtained here: `https://github.com/RAIRLab/ShadowAdjudicator`.
2. 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).
3. 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.
4. M. Ashcraft and G. Radvansky, *Cognition* (Pearson, London, UK, 2013). This is the 6th edition.
5. 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).
6. 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`.
7. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Making Maximally Ethical Decisions via Cognitive Likelihood & Formal Planning, in *Towards Trustworthy Artificial Intelligent Systems*, ed. M. Ferreira (Springer, 2021) .
8. 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).
9. M. Peveler, N. S. Govindarajulu and S. Bringsjord, *Toward Automating the Doctrine of Triple Effect*, in *Hybrid Worlds: Societal and Ethical Challenges; Proceedings of the International Conference on Robot Ethics and Standards (ICRES) 2018*, eds. S. Bringsjord, M. Osman Tokhi, M. Isabel Aldinhas Ferreira and N. S. Govindarajulu (CLAWAR, 2018), pp. 82–88. Available (within full e-book) at `http://kryten.mm.rpi.edu/HybridWorlds.pdf`.
10. 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`.

11. M. Bratman *et al.*, *Intention, plans, and practical reason* (Harvard University Press Cambridge, MA, 1987).
12. L. Kohlberg, The Claim to Moral Adequacy of a Highest Stage of Moral Judgment, *Journal of Philosophy* **70**, 630 (1973) .
13. B. Williams and J. Smart, *Utilitarianism: For and Against* (Cambridge University Press, Cambridge, UK, 1973).
14. H. A. Bedau, *Making Mortal Choices: Three Exercises in Moral Casuistry* (Oxford University Press, Incorporated, 1997).

ICRES 2021


CLAWAR
ASSOCIATION


RENSSELAER POLYTECHNIC INSTITUTE
1824