

## TOWARD AUTOMATING THE DOCTRINE OF TRIPLE EFFECT

M. PEVELER\*, N. S. GOVINDARAJULU, and S. BRINGSJORD

*Rensselaer AI & Reasoning (RAIR) Lab  
Rensselaer Polytechnic Institute (RPI)  
Troy NY 12180 USA*

\*E-mail: [pevelm@rpi.com](mailto:pevelm@rpi.com), [naveensundarg@gmail.com](mailto:naveensundarg@gmail.com), [selmer.bringsjord@gmail.com](mailto:selmer.bringsjord@gmail.com)

The **Doctrine of Double Effect** ( $DDE$ ) is a long-studied ethical principle governing whether taking an action that has both significant positive and negative effects is ethically permissible. Unfortunately, despite its storied history,  $DDE$  does not fully account for the permissibility of actions taken in certain particularly challenging moral dilemmas that have recently arrived on the scene. The **Doctrine of Triple Effect** ( $DTE$ ) can be employed in these dilemmas, to separate the intention to perform an action *because* an effect will occur, versus *in order* for that effect to occur. This distinction allows an agent to permissibly pursue actions that may have foreseeable negative effects resulting from those actions — as long as the negative effect is not the agent’s primary intention. By  $DDE$  such actions are not classified as ethically permissible. We briefly present  $DTE$  and, using a first-order multi-operator modal logic (the **deontic cognitive event calculus**), formalize this doctrine. We then give a proof-sketch of a situation for which  $DTE$  but not  $DDE$  can be used to classify a relevant action as permissible. We end with a look forward to future work.

*Keywords:* doctrine of double effect, doctrine of triple effect, machine ethics, AI

### 1. Introduction

On a daily basis, humans are faced with moral dilemmas, in which all available options have both good and bad consequences. In these situations, humans are forced to weigh the costs of their actions, and are often required to provide some explanation of why their actions justify the potential negative effects. These explanations are even more vital when the negative effects include the death, or possibility of death, of another human. To provide these explanations for a given decision in these dilemmas, much work has been done in the study and development of various ethical principles and doctrines. These works, often couched in hypothetical situations such as the well-known trolley problems, seek to provide a basis for ethical philosophers to create explanations and to provide a basis for various empirical studies. From this work, we see a rise of principles that humans will readily mix and match depending on the situation that they are faced with and their underlying socio-demographic characteristics such as race, religion, etc. Additionally, and more concerning to use of these principles in AI, we see primarily informal definitions for these principles and the conditions in which they apply, which while sufficient for a motivated human reader, cannot be readily used in AI agents that are tasked into similar situations.

As we task AI agents with more of these potentially morally charged dilemmas, it is important that we build up a library of ethical principles that have been given a rigorous and formal definition, such that they can mix and match as necessary for a given situation, as well as explain any decision they make. In pursuit of these objectives, we look to formal reasoning, in the vein of a logic that is deontic in nature to handle various obligations and permissions agents may have and that is able to describe and reason about cognitive states of agents. In our case, we readily turn to the expressive **deontic cognitive event calculus** ( $DCEC$ ), presented and used for example in Ref. 1.

One of the most common and well-studied ethical principles is the Doctrine of Double

Effect (*DDE*). This doctrine states that an action in a dilemma is permissible *iff* (1) it is morally neutral; (2) the net good consequences outweigh the bad consequences by a large amount; and (3) some of the good consequences are intended and none of the bad effects are intended<sup>a</sup>. Additionally, Ref. 3,4 show how the *DDE* has been found to be used by untrained humans for various dilemmas. However, there are certain dilemmas that the *DDE* fails to account for. In some of these situations, humans will violate principle (3) in intending bad effects to accomplish a task. To solve some of these situations, we can turn to Ref. 5’s Doctrine of the Triple Effect (*DTE*), which allows for a differentiation between committing an action *because* an effect will occur and doing it *in order* for the effect to occur.

We provide a brief overview of the rest of the paper. First, in Section 2, we begin with some brief remarks on prior work done around these two doctrines and support for why the *DCEC* is well suited for this task. Next, in Section 3 we describe the *DCEC* in minimal detail as necessary to understand the following sections. In Section 4, we describe three motivating examples of trolley problems which will be used in the following sections. Following this, we then provide an informal definition of the *DTE* in Section 5 and then provide a more rigorous formal definition in Section 6. Finally, we provide a proof sketch of the use of the formal definition in solving our principle example using the *DTE* in Section 7. In Section 8, the paper concludes with a brief conclusion where we identify some promising lines of work.

## 2. Prior Work

The *DDE* has been well-studied in both ethical philosophy and automating it for autonomous agents. However, it is not without its detractors, e.g. Ref. 6. The *DTE* on the other hand, being a newer theory, has not had as much discussion and study around it, but it should be noted that it also is not without its detractors, e.g. Ref. 7. We do not intend to cast a judgement on the validity of these arguments for or against either, but rather just focus on utilization of them within AI agents.

To build our formalization, we start with prior work done by Ref. 1 on formalizing and automating the *DDE*. Additionally, while there does exist a formalization of the *DTE* presented by Ref. 8, it is done using counterfactuals in an extensional propositional system. While impressive, this system is unfortunately not expressive enough for our needs, and that it can also generate inconsistencies when dealing with intentional states such as knowledge, belief, intention, etc. (see appendix of Ref. 1 for further discussion).

## 3. The Calculus

In this section, we present the calculus we will use to formalize the *DTE*, the **deontic cognitive event calculus** (*DCEC*). This logic has been used previously in Ref. 1 to successfully formalize the *DDE* for use in an automated theorem prover. While fully describing the calculus is out of the scope of this paper, we give a brief overview (see appendix A of Ref. 1 for a more thorough treatment). The *DCEC* is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus from Ref. 9, a first-order calculus used for modeling events and their effects. The proof calculus is based on natural deduction<sup>b</sup> and includes all the introduction and elimination rules for first-order logic, as well as an inference schema for the modal operators and related structures. The *DCEC* belongs to the general family of **cognitive event calculi** introduced in Ref. 11 as is an intensional system as opposed to an extensional system. Variations of the dialect have

<sup>a</sup>See Ref. 2 for a fuller treatment on the subject.

<sup>b</sup>We assume our readers are familiar with natural deduction and extensional logics such as FOL, such as described in Ref. 10

been used to formalize and automate intensional reasoning tasks, such as the false-belief task in Ref. 11 and *akrasia* (succumbing to temptation to violate moral principles) in Ref. 12. In the *DC $\mathcal{E}$*  there are modal operators for **B**elief, **K**nowledge, **P**erception, **O**bligation, and **I**ntention. As these are modal operators, as opposed to expressive operators, we allow for agents to have nested structures of these obligations and for them to apply these to combinations of agents, such as for modeling the sentence “Bob believes that Alice knows the *DC $\mathcal{E}$* ”, which is not properly expressible in an extensional system.

#### 4. Scenarios

To analyze these two doctrines, and the need for the *DT $\mathcal{E}$* , we utilize the well-known domain of trolley problems, focusing on three variants taken from Ref. 13 and Ref. 14. In all variants, an out of control trolley is going down a track, *track*<sub>1</sub> towards two people<sup>c</sup>, *P*<sub>1</sub> and *P*<sub>2</sub>, who are next to each other on the track and who will be hit by the trolley if no action is taken. The goal is for an agent to save these two people, and in each case, the agent is faced with an ethical dilemma to figure out. These scenarios are briefly summarized below:

**Scenario 1 - Switch Case** There is a switch that can route the trolley to a second track, *track*<sub>2</sub>. There is a person, *P*<sub>3</sub>, on *track*<sub>2</sub>. If the switch is flipped, *P*<sub>3</sub> will be hit and killed.

**Scenario 2 - Push Case** An agent can push *P*<sub>3</sub> onto the track in front of the trolley. The trolley would hit *P*<sub>3</sub> and kill him, but it would be damaged and come to a stop.

**Scenario 3 - Loop Case** An agent can flip a switch to direct the trolley onto a second track, *track*<sub>2</sub>, which will then loop back onto *track*<sub>1</sub>. However, *P*<sub>3</sub> is on *track*<sub>2</sub>, and if the trolley hits him, it will be damaged and come to a stop.

#### 5. Informal *DT $\mathcal{E}$*

In the above scenarios, the *DDE* allows us to derive that it is ethically permissible to flip the switch in the Switch Case and not permissible to push the man in the Push Case. However, it does not instantiate for the Loop Case, which disagrees with the empirical studies discussed in Ref. 15 and moral philosophers referenced in Ref. 16. This is because in the Loop Case, to flip the switch, an agent is intending that *P*<sub>3</sub> be hit so as to stop the trolley, which goes against principle (3) of the *DDE*. The *DT $\mathcal{E}$*  however gives us a more fine-grained view of intentions and the bad effects that may follow from them. Given an action, an agent can do it *because* the bad effects will happen or *in order* for the bad effects to happen. While the latter remains impermissible, the former is, so long as the good still outweighs the bad. We can use this distinction to classify an agent’s intentions as either being a secondary intention *I*<sub>S</sub> (the former case) or a primary intention *I*<sub>P</sub> (the latter case). While both intentions are used in pursuit of a goal, an agent will only actively pursue and attempt to fully follow through on primary intentions. To determine if something is a primary intention, we turn to Bratman’s test for intentions from Ref. 17. An intention is a primary intention *iff*:

- D**<sub>1</sub> if an agent intends to bring about some effect, then that agent seeks the means to accomplish the ends of bringing it about;
- D**<sub>2</sub> 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);
- D**<sub>3</sub> 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.

Using this test to create a distinction between intention types, we can proceed informally defining the *DT $\mathcal{E}$* . Just as in the case of the *DDE*, we assume we have at hand an ethical

---

<sup>c</sup>For computational purposes, the exact number of persons is not important as long as it is greater than one.

hierarchy of actions in the deontological case (e.g. forbidden, neutral, obligatory), such as presented in Ref. 18. Also, we assume that we have at hand agent-specific utility functions. We build upon the informal definition from Ref. 1 (adding emphasis on our changes) for the  $\mathcal{DTE}$  with our addition of adverbs for classifying intentions from above. For an agent, an action in a situation is said to be  $\mathcal{DTE}$ -compliant *iff*:

- $C_1$  the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord,<sup>18</sup> and require that the action be neutral or above neutral in such a hierarchy);
- $C_2$  the net utility or goodness of the action is greater than some positive amount  $\gamma$ ;
- $C_{3a}$  the agent performing the action **primarily** intends only the good effects;
- $C_{3b}$  the agent does not **primarily** intend any of the bad effects, **but may secondarily intend some of them**;
- $C_4$  no **primarily** intended bad effects are used as a means to obtain the good effects, **but secondarily intended bad effects may be**.

## 6. Formal $\mathcal{DTE}$

Utilizing the  $\mathcal{DCEC}$  we now present our formalization. Let  $\Gamma$  be the set of background axioms, which include axioms for whatever our autonomous agent knows about the world. A particular situation that is in play is represented by  $\sigma$ . We use ground fluents for effects. As stated above, we assume we have a utility function  $\mu$  that maps fluents at certain times to real-number utility values. Good effects are fluents that would have a positive utility while negative effects are fluents that have a negative utility. The signature is shown below:

$$\mu : \text{Fluent} \times \text{Moment} \rightarrow \mathbb{R}$$

Additionally, we utilize the *means* operator,  $\triangleright$  from Ref. 1 which has the following signature:

$$\triangleright : \text{Formula} \times \text{Formula} \rightarrow \text{Formula}$$

The means operator is defined such that given  $\Gamma$ , a fluent  $f$  that holds at  $t_1$  is a cause or means of another fluent  $g$  at  $t_2$  where  $t_2 > t_1$  *iff* the truth condition for  $g$  changes if we were to change or remove  $f$ . An example is that we let  $f$  stand for "throwing a stone  $s$  at a window  $w$ " and  $g$  be "window  $w$  is broken". We can see that  $g$  is not a mere side-effect of  $f$  as if we were to remove  $f$  or the stone  $s$ , then  $g$  would not hold.

To formalize the  $\mathcal{DTE}$ , we need to first formalize our test of primary intention:

### Formal Conditions for Primary Intention

- $G_1$  if an agent  $a$  intends to bring about some effect  $\phi$ , and there is some means  $\psi$  to bring about  $\phi$ , then  $a$  will intend to bring about  $\psi$ . That is:

$$\begin{aligned} & \left( \mathbf{I}(a, t_1, \text{Holds}(\phi, t_2)) \wedge \triangleright(\text{Holds}(\psi, t_1), \text{Holds}(\phi, t_2)) \right) \\ & \rightarrow \mathbf{I}(a, t_1, \text{Holds}(\psi, t_1)) \end{aligned}$$

- $G_2$  if an agent  $a$  intends to bring an effect  $\phi$  about,  $a$  will pursue that effect (that is, if one way fails to bring about  $\phi$ , then  $a$  will pursue some other way). That is:

$$\begin{aligned} & \left( \mathbf{I}(a, t_1, \text{Holds}(\phi, t_1)) \wedge \neg \text{Holds}(\phi, t_1) \wedge \triangleright(\text{Holds}(\psi, t_1), \text{Holds}(\phi, t_2)) \right) \\ & \rightarrow \mathbf{I}(a, t_1, \text{Holds}(\psi, t_2)) \end{aligned}$$

- $G_3$  if an agent  $a$  intends an effect, and is rational and has consistent intentions,

then the agent will filter out any intentions that conflict. That is:

$$\begin{aligned} & \left( \triangleright (\text{Holds}(\psi, t_1), \neg \text{Holds}(\phi, t_2)) \wedge \mathbf{I}(a, t_1, \text{Holds}(\phi, t_2)) \right) \\ & \rightarrow \neg \mathbf{I}(a, t_1, \text{Holds}(\psi, t_1)) \end{aligned}$$

Hence, for an agent's intention to be a primary intention,  $\mathbf{I}_P$ , it must then pass all three conditions. If any of these conditions are false, than the intention is a secondary intention,  $\mathbf{I}_S$ .

Given the above, we now have the necessary machinery for our formalization of the  $\mathcal{DT}\mathcal{E}$ . An agent  $a$  may carry out some action type  $\alpha$  at time  $t$ , initiating some set of fluents  $\alpha_I^{a,t}$  and terminating some set of fluents  $\alpha_T^{a,t}$ . Thus, for any action  $\alpha$  taken by an agent  $a$  at time  $t$ , given some background information  $\Gamma$  in situation  $\sigma$ , this action adheres to the  $\mathcal{DT}\mathcal{E}$  up to some event horizon  $H$ , that is  $\mathcal{DT}\mathcal{E}(\Gamma, \sigma, a, \alpha, t, H)$  iff:

### Formal Conditions for $\mathcal{DT}\mathcal{E}$

**F<sub>1</sub>**  $\alpha$  carried out at  $t$  is not forbidden. That is:

$$\Gamma \not\vdash \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

**F<sub>2</sub>** The net utility is greater than a given positive real  $\gamma$ :

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

**F<sub>3a</sub>** The agent  $a$  primarily intends only the good effects. (**F<sub>2</sub>** should still hold after removing all other good effects.) There is at least one fluent  $f_g$  in  $\alpha_I^{a,t}$  with  $\mu(f_g, y) > 0$ , or  $f_b$  in  $\alpha_T^{a,t}$  with  $\mu(f_b, y) < 0$ , and some  $y$  with  $t < y \leq H$  such that the following holds:

$$\Gamma \vdash \left( \begin{array}{c} \exists f_g \in \alpha_I^{a,t} \mathbf{I}_P(a, t, \text{Holds}(f_g, y)) \\ \vee \\ \exists f_b \in \alpha_T^{a,t} \mathbf{I}_P(a, t, \neg \text{Holds}(f_b, y)) \end{array} \right)$$

**F<sub>3b</sub>** The agent  $a$  does not primarily intend any of the bad effects, but may secondarily intend some of them For all fluents  $f_b$  in  $\alpha_T^{a,t}$  with  $\mu(f_b, y) < 0$ , or  $f_g$  in  $\alpha_I^{a,t}$  with  $\mu(f_g, y) > 0$ , and for all  $y$  such that  $t < y \leq H$  the following holds:

$$\Gamma \not\vdash \mathbf{I}_P(a, t, \text{Holds}(f_b, y)) \text{ and}$$

$$\Gamma \not\vdash \mathbf{I}_P(a, t, \neg \text{Holds}(f_g, y))$$

**F<sub>4</sub>** No primarily intended bad effects can cause the good effects, but secondarily intended bad effects can be. For any bad fluent  $f_b$  holding at  $t_1$ , and any good fluent  $f_g$  holding at some  $t_2$ , such that  $t < t_1, t_2 \leq H$ , the following holds:

$$\Gamma \vdash \left( \begin{array}{c} \mathbf{I}_S(a, t, \text{Holds}(f_b, t_1)) \wedge \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2)) \\ \vee \\ \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2)) \end{array} \right)$$

## 7. Proof Sketch for the $\mathcal{DTE}$

We now apply our formal definitions for primary intentions and  $\mathcal{DTE}$  from above to a brief proof-sketch for the Loop Case. Drawing from Ref. 1 and the Push Case, we know that it is ethically impermissible to push someone onto the track to stop the trolley. Additionally, we know that we have an intention to save the pair of people,  $P_1$  and  $P_2$ , on track  $track_1$ . Intuitively, we know that to stop the trolley in the Loop Case, we must flip the switch and have the trolley hit  $P_3$ , or in other words we intend the trolley to hit  $P_3$ . To determine the permissibility of our flipping the switch, we need to determine whether the intention of hitting  $P_3$  is a primary intention or a secondary one. To do this, we need to only show one of  $\mathbf{G}_1 - \mathbf{G}_3$  to be false, and as such we will focus on proving the negation of  $\mathbf{G}_2$ :

**Proof.** Assume the agent  $a$  primarily intends the trolley to hit  $P_3$ . Also assume that  $P_3$  walks off the track at  $t_0$ . The trolley will then not hit  $P_3$  at  $t_1$  as intended. It is given that pushing  $P_3$  at  $t_x$  is a means to having  $P_3$  be hit at  $t_{x+1}$ .  $a$  will therefore push  $P_3$  at  $t_1$  so that  $P_3$  gets hit at  $t_2$ . However, it is also given that it is impermissible to push someone and therefore not allowed. As such,  $a$  cannot push  $P_3$  onto the track, and therefore  $a$  can not primarily intend for  $P_3$  to be hit.  $\square$

From this, we see that our intention of  $P_3$  being hit is a secondary one that only occurs due to the misfortune of  $P_3$  already being on the track. As such, we are allowed to pursue the bad effect of  $P_3$  being hit to accomplish the good effect, the pair not being hit, as our utility of the bad effects is less than the utility of the good effects.

## 8. Conclusion

We now quickly summarize the primary chief contributions of this work, and end by discussing promising future lines of work. In this work, we have presented a formalization of the  $\mathcal{DTE}$  within a cognitive calculi. To do this, we first created an informal definition of both the test of primary intention,  $\mathbf{D}_1 - \mathbf{D}_3$ , as well as for the  $\mathcal{DTE}$ ,  $\mathbf{C}_1 - \mathbf{C}_4$ . From this, we built the necessary formalizations,  $\mathbf{G}_1 - \mathbf{G}_3$  and  $\mathbf{D}_1 - \mathbf{D}_4$  of both. Finally, we present a proof sketch of how this formalization could be applied in determining the ethical permissibility of flipping the switch in the Loop Case of trolley problems.

For future work, there is an immediate next step of taking this formalization and building out the machinery necessary for use of the  $\mathcal{DTE}$  in moral machines, such as described in Ref. 19. Indeed, a chief goal in creating formalizations for diverse moral doctrine is to allow machines to pick and choose which moral theories it should subscribe to for a given task, or even for usage within groups of people who differ on the grounds of race, religion, politics, etc. Having said that, it is important to note the work done in Ref. 20,21 that shows that robots are held to a different standard of humans, and are expected to do actions that would be questionable if done by a human. Indeed, in proceeding with formalization of these principles, and their subsequent usage of AI agents, will be necessary to conduct more empirical studies to see how a human views various principles as applied to an AI agent versus when applied to a human.

## Acknowledgments

A grant from the Office of Naval Research to explore “moral competence in machines” (PI M. Scheutz) has provided indispensable support for the research reported herein. Crucial support also came in the form of a grant from the Air Force Office of Scientific Research to make possible “great computational intelligence” in AIs on the strength of automated reasoning (PI S. Bringsjord).

## References

1. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, (Melbourne, Australia, 2017).
2. A. McIntyre, The Doctrine of Double Effect, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2004/2014)
3. F. Cushman, L. Young and M. Hauser, The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm, *Psychological science* **17**, 1082 (2006).
4. M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A Dissociation Between Moral Judgments and Justifications, *Mind & Language* **22**, 1 (2007).
5. F. M. Kamm, *Intricate Ethics: Rights, Responsibilities, And Permissible Harm* (Oxford University Press, New York, New York, 2007).
6. A. McIntyre, Doing away with double effect, *Ethics* **111**, 219 (2001).
7. S. M. Liao, The loop case and kamm’s doctrine of triple effect, *Philosophical Studies* **146**, p. 223–231(Jul 2008).
8. L. M. Pereira and A. Saptawijaya, *Counterfactuals in Critical Thinking with Application to Morality*, in *Model-Based Reasoning in Science and Technology*, (Springer International Publishing, 2016), p. 279–289.
9. E. T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach*, 2 edn. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2015).
10. G. Gentzen, Untersuchungen über das logische Schließen I, *Mathematische Zeitschrift* **39**, 176 (1935).
11. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
12. S. Bringsjord, N. Govindarajulu, D. Thero and M. Si, Akkratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014).
13. P. Foot, The problem of abortion and the doctrine of double effect, *Oxford Review* **5**, 5 (1967).
14. J. J. Thomson, The trolley problem, *The Yale Law Journal* **94**, p. 1395(May 1985).
15. M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A Dissociation Between Moral Judgments and Justifications, *Mind & Language* **22**, 1 (2007).
16. M. Otsuka, Double effect, triple effect and the trolley problem: Squaring the circle in looping cases, *Utilitas* **20**, p. 92–110(Feb 2008).
17. M. E. Bratman, Intention, plans and practical reason, **100**(01 1987).
18. S. Bringsjord, A 21st-Century Ethical Hierarchy for Robots and Persons:  $\mathcal{EH}$ , in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, (Lisbon, Portugal, 2017).
19. P. Bello and S. Bringsjord, On How to Build a Moral Machine, *Topoi* **32**, 251 (2013), Preprint available at the URL provided here.
20. B. Malle, M. Scheutz, T. Arnold, J. Voiklis and C. Cusimano, Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents, in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI’15*, (ACM, New York, NY, 2015) pp. 117–124.
21. B. F. Malle, S. T. Magar and M. Scheutz, AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma, in *Robotics and Well-Being*, eds. M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi and E. E. KadarIntelligent Systems, Control and Automation: Science and Engineering (Springer International Publishing, Cham, 2019) pp. 111–133.