BEYOND THE DOCTRINE OF DOUBLE EFFECT: A FORMAL MODEL OF TRUE SELF-SACRIFICE*

N. S. GOVINDARAJULU*, S. BRINGSJORD, R. GHOSH and M. PEVELER

Rensselaer AI & Reasoning Lab, Rensselaer Polytechnic Institue, (RPI)
Troy, New York 12180, USA
**E-mail: naveensundarg@gmail.com*
†*E-mail: selmer.bringsjord@gmail.com*
www.rpi.edu

We have previously formalized the **doctrine of double effect** ($\mathcal{DDE}$) in a computational logic that can be implemented in robots. The doctrine of double effect is an ethical principle very commonly used to account for human judgement in moral dilemmas: situations in which all available options have good and bad consequences. $\mathcal{DDE}$, as an ethical principle for robots, is attractive for a number of reasons — (1) Empirical studies have found that $\mathcal{DDE}$ is used by untrained humans; (2) Many legal systems use $\mathcal{DDE}$; and finally, (3) the doctrine is also a hybrid of the two major opposing camps in ethics: consequentialism/utilitarianism and deontological ethics. In spite of all its attractive features, we have found that $\mathcal{DDE}$ does not fully account for human behaviour in many ethically challenging situations. Specifically, standard $\mathcal{DDE}$ fails in situations wherein humans have the option of **self-sacrifice**. Accordingly, we present a enhancement of our $\mathcal{DDE}$-formalism to handle self-sacrifice and comment on future work.

Keywords: doctrine of double effect, true self-sacrifice, law and ethics

1. Introduction

The doctrine of double effect is an ethical principle used (subconsciously or consciously) by humans in *moral dilemmas*. Moral dilemmas are situations in which all available options have both good and bad consequences. The doctrine states that an action α in such a situation 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

*We are grateful to the Office of Naval Research for funding the research presented in this paper.

intended and none of the bad consequences are intended. $\mathcal{DDE}$ is an attractive target for robot ethics for a number of reasons. Empirical studies show that $\mathcal{DDE}$ is used by untrained humans.[1,2] Secondly, many legal systems are based upon this doctrine. (For an analysis of $\mathcal{DDE}$ used in U.S. law see Ref. 3 and Ref. 4.) In addition, $\mathcal{DDE}$ is a hybrid of the two major opposing camps in ethics: *consequentialism/utilitarianism* and *deontological ethics.* In spite of this, we have found that $\mathcal{DDE}$ does not fully account for human behaviour in many ethically difficult situations. Specifically, standard $\mathcal{DDE}$ fails in situations where humans have the option of **self-sacrifice**. In some situations, but not all, actions prohibited by $\mathcal{DDE}$ become acceptable when the receiver of harm is the self rather than some other agent.

If we have to build robots that work with humans in ethically challenging scenarios and function similar to humans, rigorously formalizing the principle and incorporating self-sacrifice is vital. The situation is made more complicated by the study in Ref. 5; it shows, using hypothetical scenarios with imagined human and robot actors, that humans judge robots differently in ethical situations. In order to build well-behaved autonomous systems that function in morally charged scenarios, we need to build systems that can not only take the right action in such scenarios, but also have enough representational capabilities to be sensitive to how others might view its actions. The formal system we present in this paper has been used previously to model beliefs of other agents and is uniquely suited for this. We present an enhancement of our $\mathcal{DDE}$-formalism in order to handle self-sacrifice. [a] Our new formal model of self-sacrifice serves two purposes: (1) helps us build robots capable of self-sacrifice from *first principles* rather from manually programming in such behavior on an *ad hoc* case-by-case basis; and (2) detects when autonomous agents make real self-sacrifices rather than incidental or accidental self-sacrifices.

2. Prior Work

While there have been millennia of legends, folk stories, and moral teachings on the value of self-sacrifice, very few empirical studies in moral psychology have explored the role of self-sacrifice. The most rigorous study of self-sacrifice to date, using the well-known trolley set of problems, has been done

[a]Full formalization of $\mathcal{DDE}$ would include conditions expressing the requirement that the agent in question has certain emotions and lacks certain other emotions (e.g., the agent cannot have *delectatio morosa*. On the strength of Ghosh's Felmë theory of emotion, which formalizes (apparently all) human emotions in the language of cognitive calculus as described in the present paper, we are actively working in this direction.

by Sachdeva *et al.* in Ref. 6. Sachdeva *et al.* report that in the standard trolley class of problems, *intended* harm to oneself to save others is looked at more favorably than *intended* harm of others. This immediately catapults us beyond the confines of standard $\mathcal{DDE}$. To account for this, we present an enhanced model of $\mathcal{DDE}$ by building upon our prior work;[7] the enhanced model can account for self-sacrifice.

3. Goal

In this section we render precise what is needed from a formal model of self-sacrifice. If one is building a self-driving car or a similar robotic system that functions in limited domains, it might be "trivial" to program in self-sacrifice, but we are seeking to understand and formalize what a model of self-sacrifice might look like in *general-purpose* autonomous robotic systems. Consider a sample scenario: A team of n, ($n >= 2$), soliders from the *blue* team is captured by the *red* team.[b] The leader of the blue team is offered the choice of selecting one member from the team who will be sacrificed to free the rest of the team. Now consider the following actions:

- $\mathbf{a}_1$ The leader picks himself/herself.
- $\mathbf{a}_2$ The leader picks another soldier against their will.
- $\mathbf{a}_3$ The leader chooses a name randomly and it happens to be the leader's name.
- $\mathbf{a}_4$ The leader chooses a name randomly and it happens to be somebody else's name.
- $\mathbf{a}_5$ A soldier volunteers to die; the leader picks their name.

In addition to robotic systems with the capability for self-sacrifice in the right situations, we need systems that can understand human decisions in ethically-charged scenarios. We need a framework that can discern that: only $\mathbf{a}_1$ and $\mathbf{a}_5$ involve *true* self-sacrifice; $\mathbf{a}_3$ is *accidental* self-sacrifice; and $\mathbf{a}_2$ might be immoral.

4. The Calculus

The computational logic we use is the **deontic cognitive event calculus** ($\mathcal{DCEC}$). This logic was used by us previously in Ref. 7 to automate $\mathcal{DDE}$. While describing the calculus is beyond the scope of this paper, we give a quick overview of the system. Dialects of $\mathcal{DCEC}$ have been used to formalize and automate highly intensional reasoning processes, such as the false-belief task[8] and *akrasia* (succumbing to temptation to violate moral principles).[9] Arkoudas and Bringsjord[8] introduced the general family of

[b]The blue/red terminology is common in wargaming and offers perhaps a somewhat neutral way to talk about politically-charged situations.

cognitive event calculi to which $\mathcal{DCEC}$ belongs, by way of their formalization of the false-belief task. $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus, a first-order calculus used for commonsense reasoning. The calculus has a well-defined syntax and proof calculus; see Appendix A of Ref. 7. The proof calculus is based on natural deduction,[10] and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures.

5. Informal $\mathcal{DDE}^*$

We now informally but rigorously present $\mathcal{DDE}^*$, an enhanced version of $\mathcal{DDE}$ that can handle self-sacrifice. Just as in standard models of $\mathcal{DDE}$, assume we have at hand an ethical hierarchy of actions as in the deontological case (e.g. forbidden, neutral, obligatory); see Ref. 11. Also given to us is an agent-specific utility function or goodness function for states of the world or effects as in the consequentialist case. The informal conditions are from Ref. 7; the modifications are emphasized in bold below. For an autonomous agent a, an action α in a situation σ at time t is said to be $\mathcal{DDE}^*$-compliant *iff*:

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord,[11] 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 [**unless** a **knows that the bad effects are confined to only** a **itself**]; and
- $\mathbf{C}_5$ if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action;that is, the action is unavoidable.

Central to the formalization of $\mathcal{DDE}$ is a utility function μ that maps fluents and time points to utility values.

$$\mu : \mathsf{Fluent} \times \mathsf{Time} \rightarrow \mathbb{R}$$

Good effects are fluents with positive utility; bad effects are fluents that have negative utility. Zero-utility fluents could be neutral fluents (which do not have a use at the moment). The

above agent-neutral function suffices for classical $\mathcal{DDE}$ but is not enough for our purpose. We assume that there is a another function κ (either learned or given to us) that gives us agent-specific utilities.

$$\kappa : \mathsf{Agent} \times \mathsf{Fluent} \times \mathsf{Time} \rightarrow \mathbb{R}$$

We can then build the agent-neutral function μ from the agent-specific function ν as shown below:

$$\mu(f,t) = \sum_{a} \kappa(a,f,t)$$

For an action α carried out by an agent a at time t, let $\alpha_I^{a,t}$ be the set of fluents initiated by the action and let $\alpha_T^{a,t}$ be the set of fluents terminated by the action. If we are looking up till horizon H, then $\hat{\mu}(\alpha,a,t)$, the total utility of action α carried out by a at time t, is then:

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

Similarly, we have $\nu(\alpha,a,b,t)$, the total utility for agent b of action α carried out by agent a at time t:

$$\nu(\alpha,a,b,t) = \sum_{y=t+1}^{H} \left(\sum_{f \in \alpha_I^{a,t}} \hat{\mu}(b,f,y) - \sum_{f \in \alpha_T^{a,t}} \hat{\mu}(b,f,y) \right)$$

6. Robust Self-Representation

Modeling true self-sacrifice (as opposed to accidental self-sacrifice as discussed before) needs a robust representation system for true self-reference (called *de se* reference in philosophical literature). We now briefly go over how we can represent increasingly stronger levels of self-reference in $\mathcal{DCEC}$, with *de se* statements being the only true self-referential statements. See Ref. 12 for a more detailed presentation of the system we used here and an analysis of *de dicto* (*"about the word"*), *de re* (*"about the object"*) and *de se* statements (*"about the self"*). We have three levels of self-reference, discussed below in the box titled "**Three Levels of Self-Representation**". For representing and reasoning about true self-sacrifice, we need a Level 3 (*de se*) representation. Assume we have a robot or agent r with a knowledge base of formulae Γ.

Level-1 representation dictates that the agent r is aware of a name or description ν referring to some agent a. It is with the help of ν that the agent comes to believe a statement $\phi(a)$ about the particular agent (which happens to be itself, $r = a$). The agent need not be necessarily aware that $r = a$. Level 1 statements are not true self-referential beliefs. This is equivalent to a person reading and believing a statement about themself that uses a name or description that they do not know refers to themself.

For example, the statement *"the n^{th} tallest person in the world is taller than the $n+1^{th}$ person"* can be known by the n^{th} tallest person without that person knowing that they are in fact the n^{th} tallest person in the world, and that the statement is about this person.

Three Levels of Self-Representation

de dicto Agent r with the name or description ν has come to believe on the basis of prior information Γ that the statement ϕ holds for the agent with the name or description ν.

$$\Gamma \vdash_r \mathbf{B}\Big(\mathsf{I}_r, \mathsf{now}, \exists a : \mathsf{Agent}\ \Big[named(a,\nu) \wedge \phi(a)\Big]\Big)$$

de re Agent r with the name or description ν has come to believe on the basis of prior information Γ that the statement ϕ holds of the agent with the name or description ν.

$$\exists a : \mathsf{Agent}\ named\,(a,\nu)\ \Big[\Gamma \vdash_r \mathbf{B}\Big(\mathsf{I}_r, \mathsf{now}, \phi(a)\Big)\Big]$$

de se Agent r believes on the basis of Γ that the statement ϕ holds of itself ν.

$$\Gamma \vdash_r \mathbf{B}\Big(\mathsf{I}_r, \mathsf{now}, \phi(\mathsf{I}_\mathsf{r}\, *)\Big)$$

Level-2 representation does not require that the agent be aware of the name. The agent knows that ϕ holds for some anonymous agent a. The below representation does not dicate that the agent be aware of the name. Following the previous example, the statement *"that person is taller than the $n+1^{th}$ person"*, where *"that person"* refers to the n^{th} tallest person, can be known by the n^{th} tallest person without knowing that they are in fact the n^{th} tallest person in the world and that the statement is about them.

Level-3 representation is the strongest level of self-reference. The special function $*$ denotes a self-referential statement. We refer the reader to Ref. 12 for a more detailed analysis. Following the above two examples, this would correspond to the statement *"I myself am taller than the $n+1^{th}$ person"* believed by the n^{th} tallest person.

7. Formal $\mathcal{DDE}^*$

Assume we have an autonomous agent or robot r with a knowledge-base Γ. In Ref. 7, the predicate $\mathcal{DDE}(\Gamma, \sigma, a, \alpha, t, H)$ is formalized — and is read as "**from a set of premises Γ, and in situation σ, we can say that action α by agent a at time t operating with horizon H is $\mathcal{DDE}$-compliant.**" The formalization is broken up into four clauses corresponding to the informal clauses $\mathbf{C}_1$–$\mathbf{C}_4$ given above in Section 5:

$$\mathcal{DDE}(\Gamma, \sigma, a, \alpha, t, H) \leftrightarrow \Big(\mathbf{F}_1(\Gamma, \sigma, a, \alpha, t, H) \wedge \mathbf{F}_2(\ldots) \wedge \mathbf{F}_3(\ldots) \wedge \mathbf{F}_4(\ldots)\quad\Big)$$

With the formal machinery at hand, enhancing $\mathcal{DDE}$ to $\mathcal{DDE}^*$ is straighforward. Now, corresponding to the augmented informal definition in Section 5, we take the $\mathcal{DDE}$ predicate defined in Ref. 7 and added disjunction.

$$\mathcal{DDE}^*(\ldots) \Leftrightarrow \begin{cases} \mathcal{DDE}\left(\Gamma, \sigma, a, \alpha, t, H\right) \bigvee \\ \mathbf{F}_1 \wedge \mathbf{F}_2 \wedge \mathbf{F}_3 \wedge \mathbf{K}\left(a, t, \left(\begin{array}{c} \left[\forall b.\left(b \neq a^*\right) \rightarrow \nu(\alpha, a, b, t) \gg 0\right] \wedge \\ \nu(\alpha, a, a^*, t) \ll 0 \end{array}\right)\right) \end{cases}$$

The disjunction simply states that the new principle $\mathcal{DDE}^*$ applies when — (1) $\mathcal{DDE}$ applies; or (2) when conditions $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ apply along with the condition that the agent performing the action **knows** that all of the bad effects are directed toward itself, and the good effects are great in magnitude and apply only to other agents.

Simulation: We take a formalization of the standard trolley scenario[7] and add the option of sacrificing oneself. In this scenario, there is a train hurtling towards $n(n \geq 2)$ persons on a track. Agent a is on a bridge and has the option of pushing a spectator b onto the tracks of the train, stopping the train and preventing it from killing the n persons. Standard $\mathcal{DDE}$ prevents pushing both a or b, but empirical evidence suggests that while humans do not agree with pushing b, they are more agreeable with a sacrificing a's own life. We take the formalization of the base scenario without options for self-sacrifice, represented by a set of formulae $\Gamma_{\langle Trolley, bridge\rangle}$, and add an action describing the action of self sacrifice, giving us $\Gamma^*_{\langle Trolley, bridge\rangle}$. We simulate $\mathcal{DDE}^*$ using our quantified modal logic theorem prover, (termed **Shadow Prover**[7]). The table below summarizes some computational statistics. [c]

		Simulation Time (s)	
Scenario	$\lvert\Gamma\rvert$	$\mathcal{DDE}$ (push b)	$\mathcal{DDE}^*$ (push a^*)
$\Gamma_{\langle Trolley, bridge\rangle}$	38	[✗] 1.48 (s)	not applicable
$\Gamma^*_{\langle Trolley, bridge\rangle}$	39	[✗] 3.37 (s)	[✓] 3.37 + 0.2 = 3.57 (s)

8. Conculusion

As our $\mathcal{DDE}^*$ model builds upon a computational model of $\mathcal{DDE}$, it can be readily automated. While this model can explain the results in Ref. 6, we

[c]The code is available at https://goo.gl/JDWzi6. For further experimentation with and exploration of $\mathcal{DDE}$, we are working on *physical*, 3D simulations, rather than only virtual simulations in pure software. Space constraints make it impossible to describe the "cognitive polysolid framework" in question (which can be used for simple trolley problems), development of which is currently principally the task of Peveler.

have not yet explored or applied this model to more realistic cases. Doing so will be challenging, as such cases, though important, are unique in a number of ways . For future work, we will look at applying $\mathcal{DDE}^*$ to a slew of such cases and explore self-sacrifice in other related ethical principles, such as the **doctrine of triple effect**.[13]

References

1. F. Cushman, L. Young and M. Hauser, *Psychological science* **17**, 1082 (2006).
2. M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, *Mind & Language* **22**, 1 (2007).
3. M. E. Allsopp, The Doctrine of Double Effect in US Law: Exploring Neil Gorsuchs Analyses *The National Catholic Bioethics Quarterly* **11**2011.
4. R. Huxtable, Get Out Of Jail Free? The Doctrine Of Double Effect In English Law *Palliative Medicine* **18** (Sage Publications, 2004).
5. B. F. 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*, (Portland, USA, 2015).
6. S. Sachdeva, R. Iliev, H. Ekhtiari and M. Dehghani, The Role of Self-Sacrifice in Moral Dilemmas *PloS one* **10** (Public Library of Science, 2015).
7. 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). Preprint available at this url: https://arxiv.org/abs/1703.08922.
8. 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, Hanoi, Vietnam, 2008).
9. S. Bringsjord, N. S. 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.
10. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterdam, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version.
11. 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).
12. S. Bringsjord and N. S. Govindarajulu, Toward a Modern Geography of Minds, Machines, and Math, in *Philosophy and Theory of Artificial Intelligence*, ed. V. C. Müller, Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol. 5 (Springer, New York, NY, 2013) pp. 151–165.
13. F. M. Kamm, *Intricate Ethics: Rights, Responsibilities, And Permissible Harm* (Oxford University Press, New York, New York, 2007).