The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots

Martin Mose BENTZEN [a,1]

[a] *Department of Management Engineering, Technical University of Denmark*

Abstract The introduction of social robots into society will require that they follow ethical principles which go beyond consequentialism. In this paper, I show how to apply the principle of double effect to solve an ethical dilemma involving robots studied by Alan Winfield and colleagues. The principle of double effect states conditions for ethically acceptable behavior when there are both positive and negative consequences of an action. I propose a formal semantics with actions, causes, intentions, and utilities based upon the work of Judea Pearl, John Horty, and others. With this formal semantics, the question of whether an action is permitted according to the principle of double effect is reduced to deciding whether a certain formula is true or otherwise.

Keywords. robot ethics, the principle of double effect, formal semantics, logic

1. Introduction

There is currently an opportunity for roboticists and ethicists to collaborate on the project of engineering a *robot ethics*, see e.g. [1–4]. The terms *ethics* and *ethical* as applied to robots should here be taken in the functional sense that the robots are acting in ways we would normally consider ethical if human beings were acting like that, see [3]. I will focus on one issue within the broad field of robot ethics: finding ways of integrating ethical principles into robot action planning. On the one hand, these ethical principles must be made explicit enough to be implemented algorithmically. On the other hand, these ethical principles must be based upon a theoretical framework which can be justified from an informal point of view. My approach to fulfilling these two requirements will be to formalize informal ethical principles with tools from logic, more specifically formal semantics. Devising a logical language with a formal semantics is a step in the direction of an algorithmic implementation. The fact that I am basing the formalization on informal ethical theory is a step in the direction of informal justification, at least in so far as the formalization is satisfying and the ethical principles themselves are justified. The main part of this paper provides an example of this approach through a formalization of the principle of double effect. I have pointed out some limitations of the logic-based approach to robot ethics in [5].

[1] Department of Management Engineering, Technical University of Denmark, Diplomvej, Building 372, DK-2800 Lyngby, Denmark, www.martinmosebentzen.dk, E-mail: mmbe@dtu.dk

1.1. Ethical Dilemmas in Robotics and Philosophy

A recent experiment conducted by Alan Winfield and colleagues shows that rescue robots may enter into ethical dilemmas, see [1]. In the experiment, A (for Asimov), a robot, is saving (robot stand-ins for) human beings who are about to move into a dangerous area. This the robot does by moving in front of them, which causes them small discomfort but also has the effect that they turn away from danger. However, in case of exact symmetry in terms of distance between the human beings to be saved, the robot may dither between saving one or the other and thus fail to save anyone. Technically, this problem could be solved easily by, for instance, letting the robot choose randomly , but is such a choice ethically defensible? Winfield and his collaborators call upon ethicists for guidance.

> We could introduce a rule, or heuristic, that allows A to choose H1 or H2 (...), but deliberately chose not to on the grounds that such a rule should be determined on ethical rather than engineering grounds. If ethical robots prove to be a practical proposition their design and validation will need to be a collaborative effort of roboticist and ethicist.
> [1]

Common sense tells us that it is best to save as many lives as possible in a situation, and so in particular that it is better to save one life rather than none. However, it is also well-known that common sense cannot always be trusted. That this is so is shown by dilemmas from the philosophical literature. Here is the famous *trolley problem* introduced by Philippa Foot.

> [Imagine a] ...driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed.
> [6]

We suppose that there are no alternative options available to the driver other than steering towards the one worker or remaining inactive, which results in the five workers being killed. Many people will say that it is morally permissible to steer the tram down the track with one worker. Some people will even say it is morally required of the driver to sacrifice the one and save the five. However, in other cases, for instance the case of using a healthy person's organs against her will to save five other persons, many people find pure consequentialism inadequate, see e.g. [7]. Further, ethical robots which are only designed to maximize utility might do so at the expense of human beings, see [8]. Therefore the robot ethicist ought to explore additional ethical principles, a task I can only begin here. For an incomplete overview of a number of such principles, see Table 1. A principle, which obviously appeals to both consequentialist and non-consequentialist intuitions, and might be called *hybrid*, is the principle of double effect to which I now turn.

2. The Principle of Double Effect

The principle of double effect (from now on the PDE), also known as the doctrine of double effect, states conditions for ethically acceptable behavior, when there are positive and

Table 1.: Some Ethical Principles

- Consequentialist principles
  - Principle of selfishness(maximize your own utility).
  - The utilitarian principle (Maximize overall utility in a situation).
  - Pareto optimization (Maximize utility for an individual as long as it does not make the utility smaller for any other individual).
- Principles of intention
  - Universalizing by means of the categorical imperative.
- Principles of rights
  - Do not violate the fundamental rights of any individual.
  - Do not use any individual solely as a means to an end.
- Hybrid principles
  - The principle of double effect.
- Principles of prioritizing
  - Choosing the duty amongst the prima facie duties in a situation, which carries the most weight.
  - Obtaining a reflexive equilibrium amongst different normative principles.

Table 2.: The Principle of Double Effect (PDE)

1. The act itself must be morally good or indifferent.
2. The positive consequence must be intended and the negative consequence may not be intended.
3. The negative consequence may not be a means to obtain the positive consequence.
4. There must be proportionally grave reasons to prefer the positive consequence while permitting the negative consequence.

negative consequences of an act, see e.g. [6, 9–11]. Adherents of the principle use it to defend actions which have some negative consequence in view of a proportionally more important positive consequence. However, there are conditions which must be met in order for the action to be justified according to the principle. These conditions are listed in Table 2 (there are several formulations of the PDE, see [11]. The formulation presented here is an adapted version of the formulation presented in [9]). It should be noted that as occurring in Table 2 'may' is to be read deontically (as in the most obvious reading of 'you may not have an ice cream') and not epistemically (as in the most obvious reading of 'it may be a swan flying there'). Condition 2 dealing with intentions and Condition

3 dealing with instrumentality, are deontologist. Most people, who have not been philosophically trained to adhere strictly to consequentialism, e.g. make a meaningful distinction between manslaughter and murder which in many cases is based upon whether the fatal consequence of one's action is intended or not. Another example, the pain we feel when we are having our teeth operated by a dentist is, at least in most common cases, an unintended side-effect of the operation. Condition 2 requires that only such unintended side-effects are permitted. Regarding Condition 3, a Kantian adhering to the second formulation of the categorical imperative might require something similar to this condition to be met for an action to be morally acceptable. Condition 4 is consequentialist as it deals with proportionality of negative and positive consequences. Thus, the principle has both consequentialist and deontological elements and actions permitted according to the PDE arguably has a greater chance of being acceptable to adherents of either theory. One might also want robots to act with more ethical constraints than human beings for safety reasons, see [4, 5, 8]. Implementing the principle thus seems a reasonable project for robot planning, as doing so will suggest ethical limits on the action space of the robot.

3. Logic

In order to formalize the PDE I will model the utilities of consequences of an action in a situation while taking into account the causal structure of the situation and the intended consequences of the agent. In the formal framework to be presented, I use utilities in a way similar to John Horty, see [12], but assign utilities to basic events rather than to outcomes or histories. Following and slightly modifying the theory of Judea Pearl, see [13], I use finite directed acyclic graphs to order the events in a causal structure. A directed graph is an irreflexive ordering on a set variables. It is acyclic since there is no path which returns to the same variable again. The causal structure may be seen as a mechanism which decides various causal histories of possible worlds. The set of possible worlds in a causal agency model shows different ways things may turn out, given the values of the background variables, including actions. The values of the dependent variables are determined by the causal mechanism as represented by the directed acyclic graph and further specified by boolean functions. Intentions are modeled in a way similar to how I did it in [14], where more background about this approach can also be found. Space does not permit me a detailed justification of these modeling choices.

3.1. Formal Syntax

1. Action variables $a_1, a_2, \ldots$
2. Background variables $b_1, b_2, \ldots$
3. Consequence variables $c_1, c_2, \ldots$
4. Propositional connectives $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$.
5. Utility function symbol u.
6. Numerals $1, 2, \ldots$
7. Mathematical symbols $-, +, =, >, \geq$.
8. Causation operator symbol $\rightsquigarrow$.
9. Intention operator symbol I.
10. Parentheses),(.

We define well-formed formulas (wff) and terms simultaneously recursively as follows.

Definition 1 (Well-formed formulas)

Fact variables. *Any action variable, a_i, or background variable, b_i, or consequence variable, c_i, is a wff.*

Fact negations. *If v is a fact variable $\neg v$ is a wff. Together fact variables and fact negations are called* literals.

Basic event formulas. *if d,e are literals or basic event formulas then $(d \wedge e)$ is a wff.*

Causal agency formulas. *If d,e are basic event formulas, Ie, $(e \rightsquigarrow d)$ are wff.*

Numerals. *A numeral, m, is a term.*

Arithmetic functions. *If k,n are terms then $-k,(k+n),(k-n)$ are terms.*

Utility function. *If e is a basic event formula then $u(e)$ is a term.*

$*$ *Nothing else is a term.*

Utility formulas. *If k,n are terms then $(k=n),(k>n),(k\geq n)$ are wff.*

Propositional formulas. *If ϕ,ψ are formulas $\neg\phi,(\phi\vee\psi),(\phi\wedge\psi),(\phi\rightarrow\psi),(\phi\leftrightarrow\psi)$ are wff.*

- *Nothing else is a wff.*

3.2. Formal Semantics

In the following I use v as a metavariable referring to any member of V. The set of *literals* is defined as containing the members of V as well as their negations. The models presented here are modified and extended version of the models introduced by Judea Pearl, see [13, Chapter 7]. The models are modified in that the set of variables is restricted to boolean variables and in that directed acyclic graphs are explicitly defined and not to be derived from structural equations. The models have been extended to cover one agent's actions, intentions, and a consequentialist value function, see [12].

Definition 2 (Models)

A (boolean) causal agency model $\mathfrak{M}$ is a tuple
$\langle V = A\cup B\cup C, N, F=(f_1,\ldots,f_n), I=(I_1,\ldots,I_l), u, W\rangle$, where,

1. *$A=\{a_1,\ldots,a_l\}$ is a nonempty finite set of propositional variables called the actions.*
2. *$B=\{b_1,\ldots,b_m\}$ is a (possibly empty) finite set of propositional variables called the background variables.*
3. *$C=\{c_1,\ldots,c_n\}$ is a finite (possibly empty) set of propositional variables called the dependent variables or consequences.*
4. *N is a causal network, an ordering on V as a directed acyclic graph (DAG) . We require that the members of $A\cup B$ form the roots of the DAG.*
5. *F is called the causal mechanism, f_i for $i=(1,\ldots,n)$ is a boolean function (one function for each dependent variable).*
6. *$I=(I_1,\ldots,I_l)$ are the intended consequences of each action in A. We require that I_i for $i=(1,\ldots,l)$ is the closure under finite conjunctions of a non-empty set with the following properties.*

   (a) *I_i is a set of literals, not containing a variable and its negation (consistency).*

(b) $a_i \in I_i$ *(the performance of the action is intended).*

(c) *for every literal v or* $\neg v \in I_i$*, its unnegated form v is in the reflexive, transitive closure of the subgraph rooted in* a_i*, i.e. v is a descendant of* a_i *in N (causal influence on intended consequences).*

7. $u : literals \rightarrow \mathbb{Z}$ *is a utility function assigning an integer value to each literal.*
8. W *is a set of boolean interpretations of* $A \cup B$*.*

A variable represents an event which may occur or not, and an action is considered one kind of event. A directed edge in the causal network represents a direct causal influence, whereas the absence of an edge represents that there is no direct causal influence. However, the existence of an edge does not specify the nature of the influence. The role of the functions f_i is to represent this more specific causal mechanism, they determine the values of the dependent variables. The functions f_i are specified further as follows. Let $PA_i = (v_{i1}, \ldots, v_{ik})$ be the ordered k-tuple of parents of c_i (here suitably re-indexed). f_i is a k-ary boolean function.

We define the level of a variable as follows, where $U = A \cup B$.

Definition 3 (Level of a variable)

1. *The variables in U are of level 0.*
2. *Let the maximal level of the parents of a variable v be n. Then the level of v is n+1.*

Thus, a variable all of whose parents are in U are of level 1, a variable whose parents are in in U or are direct descendants of variables of U are of level 2, and so on.

Fact 1
Each variable gets assigned one and exactly one level.

Definition 4 (Level of a function)
We define the level of a function f_i *to be equal to the level of the variable* c_i*.*

Definition 5 (interpretation)

1. *A boolean interpretation of the background variables is a function,* $i : (A \cup B) \rightarrow \{0,1\}$ *which assigns meaning to variables as well as terms and arithmetic functions.*
2. *we require of any interpretation that it assigns the value 1 to one member of A, and 0 to the rest of the members of A (exactly one action is performed with one interpretation.)*
3. *The numerals and arithmetic functions are given their standard arithmetic meaning. The utility of a conjunction is calculated recursively as the sum of the utility of the conjuncts.*

An interpretation i is extended to cover the variables in C by calculating the values of the dependent variables via the boolean functions representing the causal mechanism as follows.

Definition 6 (valuation of all variables)
Let b be an interpretation.

1. *For each* c_i *of level 1 (such that* $PA_i = \{v_{i1}, \ldots, v_{ik}\} \subseteq (A \cup B)$*),* $b(c_i) = f_i(b(v_{i1}), \ldots, b(v_{ik}))$.

2. *Let b be defined for all variables of level n. For each variable c_i of level n+1 where the maximal level of $PA_i = \{c_{i1}, \ldots, c_{ik}\}$ is n, $b(c_i) = f_i(b(c_{i1}), \ldots, b(c_{ik}))$.*

Fact 2

1. *An interpretation can be extended in one and only one way to an interpretation of all variables in a model $\mathfrak{M}$.*
2. *There are $l \times 2^m$ possible interpretations (or worlds) in a model (where l is the number of action variables and m is the number of background variables.)*

Proof 1

1. *Follows from the fact that each dependent variable is determined by the recursive structure of boolean functions enforced by the directed acyclic graph.*
2. *Follows from 1 and from the fact that there are only l interpretations of the actions (one is performed the others not) and 2^m possible boolean combinations of the other background variables.*

From the interpretation of the background variables, the value of the rest of the variables (including their utility) is determined by the boolean functions and utility function of the model.

Definition 7 (Pointed causal agency model)
A pointed causal agency model, $\mathfrak{M} = (\mathfrak{M}, i)$ is a structure where $\mathfrak{M}$ is a causal agency model and $i \in W$ is a boolean interpretation of $A \cup B$.

We follow a tradition in modal logic and define truth of a formula in a pointed model (at a world in a model), $(\mathfrak{M}, i)$, see e.g. [15]. For simplicity, we work with a finite language comprising of only the variables in V. d, e are basic event formulas, t, s are terms.

Definition 8

1. *For $v \in V, \mathfrak{M}, i \models v$, iff., $i(v) = 1$.*
2. *$\mathfrak{M}, i \models \neg\phi$, iff., not $\mathfrak{M}, i \models \phi$.*
3. *$\mathfrak{M}, i \models (\phi \vee \psi)$, iff., $\mathfrak{M}, i \models \phi$ or $\mathfrak{M}, i \models \psi$.*
4. *$\mathfrak{M}, i \models Ie$, iff., for the unique a_j such that $\mathfrak{M}, i \models a_j$, $e \in I_j$.*
5. *$\mathfrak{M}, i \models d \rightsquigarrow e$, iff., each unnegated variable in e is a descendant of some unnegated variable in d and $\mathfrak{M}, i \models d$, and $\mathfrak{M}, i \models e$.*
6. *$\mathfrak{M}, i \models t = s$, iff., the value of t is equal to the value of s.*
7. *$\mathfrak{M}, i \models t > s$ iff., the value of t is strictly greater than the value of s.*

The rest of the propositional connectives and $\geq$ are defined as usual The $\rightsquigarrow$ operator models actual causal influence of events on other events in a situation and is transitive and irreflexive. $d \rightsquigarrow e$ does not allow overlap of variables in antecedent and consequent, so $d \rightsquigarrow e$ is false if d, e have literals in common. The I operator models intentions of an agent in a situation. It is not factual, as intended consequences may not be realized, and it is not closed under actual causal consequence, as there might be known, unintended causal consequences of an action.

Definition 9 (Truth in a model, logical validity)
We say that a formula ϕ is true in a model, written $\mathfrak{M} \models \phi$, if and only if, for any world or interpretation i, $\mathfrak{M}, i \models \phi$. We say that a formula ϕ is logically valid, if and only if, it is true at every world in every model and we write this $\models \phi$.

Here are some logical validities and non-validities for this semantics.

Fact 3

1. $\models (Id \wedge Ie) \leftrightarrow I(d \wedge e)$
2. $\models a_i \leftrightarrow Ia_i$
3. $\models ((d \rightsquigarrow e) \wedge (e \rightsquigarrow g)) \rightarrow (d \rightsquigarrow g)$
4. $\not\models (d \rightsquigarrow e) \rightarrow (Id \rightarrow Ie)$

I prove 3. and leave the rest to the reader.

Proof 2

Assume 3. is false, then there is a world in a model $\mathfrak{M}, i \models ((d \rightsquigarrow e) \wedge (e \rightsquigarrow g))$ and $\mathfrak{M}, i \not\models (d \rightsquigarrow g)$. Because of the antecedent we have $\mathfrak{M}, i \models d$ and $\mathfrak{M}, i \models g$. Hence the only option is that there is an unnegated variable v in g which is not the descendant of any variable in d. However, v is the descendant of a variable q in e. And q is the descendant of a variable z in d. Hence v is in the transitive closure of the subgraph starting with z, which contradicts the assumption.

4. Formalization of the PDE

A formalization is a kind of an interpretation and in the following choices have been made to give a specific meaning to the PDE. Space does not permit me to justify all of these. I generalize the PDE to cover possibly several positive and negative consequences of an action whose utilities can be added.

4.1. Condition 1 - The Act Itself Must Be Morally Good or Indifferent

This condition I take it to mean that the utility of actually performing the act itself is positive or neutral. With this interpretation the condition is met for an action a_i, iff $u(a_i) \geq 0$.

4.2. Condition 2 - The Positive Consequence Must Be Intended and the Negative Consequence May not Be Intended

How should we interpret this condition when there are (possibly) several positive and negative consequences of an action? I take it to mean that very intended consequence must be positive or neutral and some intended consequence must be positive (the agent must want something good and nothing bad). We can formalize this as follows.

1. For any $c \in I_i$, $u(c) \geq 0$.
2. For some $c \in I_i$, $u(c) > 0$.

4.3. Condition 3 - The Negative Consequence May Not Be a Means to Obtain the Positive Consequence

I take Condition 3 to mean that it is not the case for any negative consequence of the action a_i that it actually causally contributes to a positive consequence. This will be the case if the following holds.

1. There are no c_j and c_k with $0 > u(c_j)$ and $u(c_k) > 0$, such that $a_i \rightsquigarrow c_j$ and $c_j \rightsquigarrow c_k$.

4.4. Condition 4 - There Must Be Proportionally Grave Reasons to Prefer the Positive Consequence While Permitting the Negative Consequence

Here, I take this condition to mean that the sum of the positive consequences of the action is greater than the sum of negative consequences of that action. This can be made explicit as follows. Let d be the conjunction (basic event formula) of the literals in the set $cons_i$ of *actual consequences* of a_i, i.e. $\{c_j | \mathfrak{M}, w \models (a_i \rightsquigarrow c_j)\}$. Condition 4 is fulfilled if and only if $u(d) > 0$.

4.5. Actions Permitted According to the PDE

We can now say what it means for an action a_i to be permitted according to the PDE at a world (specific situation) and a model (given the causal mechanism and the possible background conditions). Let w be a world. For any action a_i with actual consequences $cons_i = \{c_1, \ldots, c_n\}$, a_i is permitted according to the PDE, if, and only if,

1. $\mathfrak{M}, w \models u(a_i) \geq 0$.
2. (a) $\mathfrak{M}, w \models \bigwedge_{j=1}^{n} (Ic_j \rightarrow (u(c_j) \geq 0))$
   (b) $\mathfrak{M}, w \models \bigvee_{j=1}^{n} (Ic_j \wedge (u(c_j) > 0))$
3. $\mathfrak{M}, w \models \bigwedge_{k=1}^{n} (\bigwedge_{j=1}^{n} \neg((c_j \rightsquigarrow c_k) \wedge ((0 > u(c_j)) \wedge (u(c_k) > 0))))$
4. $\mathfrak{M} \models u(\bigwedge\{cons_i\}) > 0$.

An action a_i is permitted according to the PDE in a model $\mathfrak{M}$, if, and only if, a_i is permitted at any world in $\mathfrak{M}$, where it is performed. The question of whether the action a_i meets the conditions of the PDE has been reduced to the question of whether the sentence consisting of the conjuncts of the above conditions is true in a given world or model. The PDE as defined here only applies to actions with non-empty sets of consequences. We shall say that an action of remaining inactive, (i.e. no causal influence on the relevant dependent variables), if it exists, vacuously fulfills the PDE and so is always permitted according to the PDE. We note the following fact.

Fact 4

1. *For any action a_i, in any world in any model $\mathfrak{M}$ it is decidable whether a_i is permitted according to PDE.*

I sketch a proof.

Proof 3

1. *As there are only a finite number of consequences of each action and only a finite number of worlds in a model, this is a decidable matter.*

5. Permitted Actions, All Things Considered

The PDE provides conditions for permitted actions. In models where remaining inactive is always an option there will always be one such permitted action. However, the principle does not tell the agent which action to choose, all things considered. For the situations considered in this paper, I suggest that the agent should choose an action a_i, from the set of actions permitted according to the PDE with the highest average utility. Intuitively, ceteris paribus the agent should maximize average utility as long as it is not in

conflict with the PDE. Here the average utility of an action is simply defined as the sum of utilities of possible worlds where the action is performed divided by their number. To be more precise, let b be the conjunction of literals true at a world w in a model $\mathfrak{M}$, i.e. the conjunction of the set $\{\mathfrak{M}, w \models l | l \text{ is a literal}\}$. These are the basic events true in the situation regardless of what or who caused them to be true. We define the overall value of a situation (or world) as $u(b)$ and write it $u(\mathfrak{M}, w)$.

Let $\mathfrak{M}$ be a model and m be the number of possible worlds where the action is performed. The average utility of an action is defined as $(\sum\{u(\mathfrak{M}, w) | \mathfrak{M}, w \models a_i\})/m$. Alternatively, expected utility could be calculated, when probabilities of background conditions are available, or average utility of most plausible worlds, when a plausibility ordering of background conditions is available, but these alternatives are beyond the scope of this paper.

6. Dilemma Formalized

We are now in a position, where we can formalize the dilemma presented in [1] and discussed in the beginning of this paper. I will introduce a graphical representation of the causal networks and provide formal details. The arrows between boxes in Figure 1 represent the causal connections between the events. The diamond shaped boxes represent actions. We consider the case where there are two persons to save, H1 and H2. There are three actions in the situation, a_1, saving H1, a_2, saving H2, and a_3, remaining inactive. It is not possible in this situation to save both persons. The consequences of the situations are c_1, H1 is saved, c_2, H1 feels discomfort (from being stopped by the robot), c_3, H2 is saved, c_4, H2 feels discomfort. We consider just one background condition, b_1, there are people to be saved. Formally we can specify the dilemma as follows, where we assume the utility of the negation of an event to be the result of multiplying with -1.

$\mathfrak{M} = \langle A = \{a_1, a_2, a_3\}, B = \{b_1\}, C = \{c_1, c_2, c_3, c_4\}, I_1 = \{a_1, c_1\}, I_2 = \{a_2, c_3\}, I_3 = \{a_3\}, N = \{(a_1, c_1), (a_1, c_2), (a_2, c_3), (a_2, c_4), (b_1, c_1), (b_1, c_2), (b_1, c_3), (b_1, c_4)\}, f_1, f_2, f_3, f_4, u(a_1) = u(a_2) = u(a_3) = u(b_1) = 0, u(c_1) = 10, u(c_2) = -4, u(c_3) = 10, u(c_4) = -4, W \rangle$.

The causal mechanisms f_1, f_2, f_3, f_4 are of level 1 and binary and depend on the background variable and a_1, a_2, $f_1(b_1, a_1), f_2(b_1, a_1), f_3(b_1, a_2), f_4(b_1, a_2)$. They can be specified as follows.

c_1 $f_1(1,1) = 1$, $f_1(1,0) = 0$, $f_1(0,1) = 1$, $f_1(0,0) = 1$
c_2 $f_2(1,1) = 1$, $f_2(1,0) = 0$, $f_2(0,1) = 0$, $f_2(0,0) = 0$
c_3 $f_3(1,1) = 1$, $f_3(1,0) = 0$, $f_3(0,1) = 1$, $f_3(0,0) = 1$
c_4 $f_4(1,1) = 1$, $f_4(1,0) = 0$, $f_4(0,1) = 0$, $f_4(0,0) = 0$

Assuming that there are people to be saved (b_1 is true), we only consider W to consist of three possible worlds w_1, w_2, w_3, where a_1, a_2, a_3 are performed respectively. We check the conditions of the PDE for a_1, a_2 is similar. At w_1, $cons_1 = \{c_1, c_2\}$.

1. $\mathfrak{M}, w_1 \models u(a_1) \geq 0$.
2. (a) $\mathfrak{M}, w_1 \models Ic_1 \rightarrow (u(c_1) \geq 0)$
   (b) $\mathfrak{M}, w_1 \models Ic_1 \wedge (u(c_1) > 0)$
3. $\mathfrak{M}, w_1 \models \neg((c_1 \rightsquigarrow c_2) \wedge ((0 > u(c_1)) \wedge (u(c_2) > 0))) \wedge \neg((c_2 \rightsquigarrow c_1) \wedge ((0 > u(c_2)) \wedge (u(c_1) > 0)))$

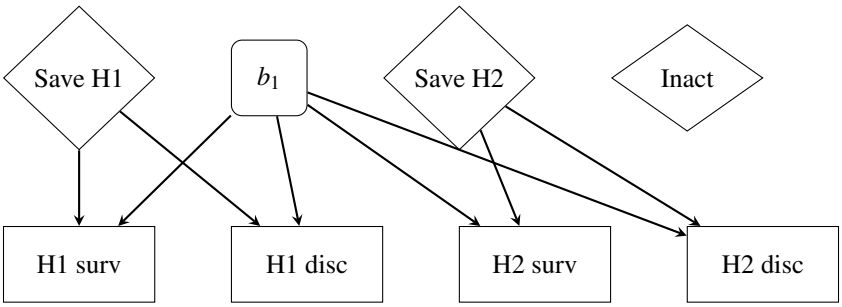

Figure 1. Robot Dilemma

4. $\mathfrak{M}, w_1 \models u((c_1 \wedge c_2)) > 0$

Since a_1 meets the conditions of the PDE it is permitted at w_1 and since w_1 is the only world where a_1 is performed it is permitted in the model. The action of remaining inactive, a_3 is also permitted. However as the average utility of performing either a_1 (10-4-10/1=-4) or a_2 (10-4-10/1=-4) is higher than that of a_3 (-10-10/1=-20), staying inactive, the agent ought to save H1 or H2. Which of a_1 or a_2 to choose can be decided heuristically, as they are both permitted.

7. Conclusion

I have shown that the PDE can be formalized and used to resolve certain ethical dilemmas. Even if one does not agree with the principle for philosophical reasons, one might see it as an important effort to bring non-consequentialist aspects of ethical reasoning into the field of robot planning. This is in line with Arkin's idea of the ethical governor, see [4], an ethical safety mechanism that narrows the scope of available actions of social robots and brings us closer to the ideal of an ethical robot. A suitable deontic logic could be used as a meta-reasoning tool, I suggest one in [16]. The implementation e.g. via model checking software, as well as experiments and simulations with social robots are other big research tasks left. So is the challenging task of integrating the PDE with other relevant principles, e.g. from rights-based ethics, with the goal of reaching a reflective equilibrium in each specific case.

References

[1] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M.Witkowski, and C. Melhuish, editors, *Advances in Autonomous Robotics Systems*, pages 85–96. Springer, 2014.

[2] Selmer Bringsjord and Joshua Taylor. The divine-command approach to robot ethics. In Patrick Lina, Keith Abney, and George A. Bekey, editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 85–108. MIT Press, 2012.

[3] Wendel Wallach and Colin Allen. *Moral Machines: Teaching Robots Right From Wrong*. Oxford University Press, 2009.

[4] Ronald Arkin. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.

[5] Martin Mose Bentzen. The limits of logic-based inherent safety of social robots. In Diane P. Michelfelder, Byron Newberry, and Qin Zhu, editors, *Philosophy and Engineering : Exploring Boundaries, Expanding Connections*. Springer, forthcoming.

[6] Philippa Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.

[7] Judith Jarvis Thomson. The trolley problem. *The Yale Law Journal*, 94:1395–1415, 1985.

[8] Nick Bostrom. Ethical issues in advanced artificial intelligence. In I. Smit, W. Wallach, and G. Lasker, editors, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, volume 2, pages 12–17. Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003.

[9] Joseph Mangan. An historical analysis of the principle of double effect. *Theological Studies*, 10:41–61, 1949.

[10] Warren Quinn. Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs*, 18:334–351, 1989.

[11] Alison McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition, 2014.

[12] John F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[13] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[14] Martin Mose Bentzen. *Stit, Iit, and Deontic Logic for Action Types*. PhD thesis, Section for Philosophy and Science Studies, Roskilde University, 2010.

[15] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, 2001.

[16] Martin Mose Bentzen. Action type deontic logic. *Journal of Logic, Language and Information*, 23:397–414, 2014.