Moral Permissibility of Action Plans

Felix Lindner, Robert Mattmüller, Bernhard Nebel
University of Freiburg, Germany
{lindner, mattmuel, nebel}@informatik.uni-freiburg.de

Abstract

Research in classical planning so far was mainly concerned with generating a satisficing or an optimal plan. However, if such systems are used to make decisions that are relevant to humans, one should also consider the ethical consequences generated plans can have. We address this challenge by analyzing in how far it is possible to generalize existing approaches of machine ethics to automatic planning systems. Traditionally, ethical principles are formulated in an action-based manner, allowing to judge the execution of one action. We show how such a judgment can be generalized to plans. Further, we study the computational complexity of making ethical judgment about plans.

Introduction

With the advent of autonomous machines that drive on the streets or act as household robots, it has been argued that we need to add an ethical dimension to such machines leading to the development of the research area *machine ethics* (Anderson, Anderson, and Armen 2005; Anderson and Anderson 2011). One important question is how we can align the behavior of autonomous machines with the moral judgment of humans. In this context, most often the question is whether a particular action is morally obligatory, permissible or impermissible, given a particular ethical principle (Driver 2006). Judging one action is, of course, important. However, automated planning systems (Ghallab, Nau, and Traverso 2016) are faced with the problem of making a huge number of decisions about including actions into a plan. And it does not necessarily make sense to analyze the ethical contents of each such decision in isolation, but it may be necessary to take an ethical perspective on an entire plan (and perhaps alternative plans). As an example, consider utilitarian reasoning: if every action in a plan were judged in isolation, one would not be allowed to perform an action that temporarily decreases the utility, even if this action is a necessary prerequisite for later earning a lot of utility in a globally optimal final reachable state. Judging a plan as a whole allows considering this early investment for the sake of a later benefit as permissible from a utilitarian perspective.

In this paper, we address these problems. First, we will look at what kind of additional information we need in order to be able to make moral judgments in the context of different ethical theories. Secondly, we will propose methods to judge the ethical acceptability of a plan. We will test the proposed notions using examples from the literature on moral dilemmas. Thereby, we do not limit ourselves to one particular ethical principle, but will consider a number of different principles that have the potential to be treated computationally, similar to the HERA (Lindner, Bentzen, and Nebel 2017) approach. Third, we will analyze the computational complexity of assessing the moral permissibility of a plan.

The remainder of the paper is structured as follows: In the next section, we introduce different ethical principles that have been discussed in the literature. Then, the planning formalism we will use throughout the paper will be specified. This is basically a propositional planning formalism extended by variables with non-binary domains, exogenous events, and moral valuations of actions and consequences. We then formalize the notions of causation and means to an end in the framework of our planning formalism. Based on that, we can then formalize different ethical principles, which we will use to analyze the computational problem of ethically validating a given plan. Finally, we sketch related work and conclude.

Ethical Principles

In moral philosophy, various ethical principles are investigated. Ethical principles are descriptions of abstract rules that can be used to determine the moral permissibility of concrete courses of actions. In this section, we introduce ethical principles which embrace different views on how to assess moral permissibility of actions: *utilitarianism*, *deontology*, three versions of the *do-no-harm principle*, and the *principle of double effect*.

The *utilitarian principle* focuses on consequences of actions. It says that an agent ought to perform the action amongst the available alternatives with the overall maximal utility. Thus the action which the agent ought to perform is the one which leads to the best possible situation, i.e., the highest utility. The utilitarian principle is often contrasted with *deontology*. According to deontology, an action does not get its moral value from its consequences. An action is permissible according to deontology if and only if the act itself is intrinsically morally good or indifferent.

Do-no-harm principles are consequentialist. Unlike utilitarianism, they state that an agent may not perform an action which causes any harm. In its standard version, the do-no-harm principle is satisfied in case the agent remains inactive as there will then be no caused harmful consequences. Hence, the distinction between doing and allowing is relevant to this principle, as it is the causal consequences of an action which are considered. A more restrictive version of the do-no-harm principle can be found in *Asimov's first law of robotics* forbidding robots to bring about harm by their action and forbidding inaction when harm could be avoided. As another variant of do-no-harm, we will introduce the *do-no-instrumental-harm principle*. This principle allows for harm as a mere side effect but not as a means to the agent's goals. Finally, we will consider the *principle of double effect*. Under this principle, an action is permissible if five conditions hold: 1) The action itself is morally good or neutral; 2) Some positive consequence is intended; 3) No negative consequence is intended; 4) No negative consequence is a means to the goal; and 5) The positive consequences sufficiently outweigh the negative ones. The first condition of the principle of double effect implements deontology. Thus, actions are assumed to have an inherent moral value, which does not (necessarily) stem from the effect of an action. The second and third conditions take the intentions, or goals, of the agent into consideration: An agent may not have a bad consequence as a goal, but it should intend something good. The fourth condition is an implementation of the do-no-instrumental-harm principle: Morally bad consequences are permissible as side effects only. And finally, the fifth condition is a weaker version of utilitarianism: In our interpretation, the condition requires that all in all the effects of the action must yield positive utility.

Planning Formalism

We use a planning formalism based on SAS$^+$ (Bäckström and Nebel 1995), extended with conditional effects (Rintanen 2003) and exogenous events (Fox, Howey, and Long 2005; Cresswell and Coddington 2003).

Language. A *planning task* is a tuple $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ consisting of the following components: $\mathcal{V}$ is a finite set of *state variables* v, each with an associated finite domain $\mathcal{D}_v$. A *fact* is a pair $\langle v, d \rangle$, where $v \in \mathcal{V}$ and $d \in \mathcal{D}_v$, also written as $v{=}d$ in conditions and $v{:=}d$ in effects. The set of all facts is denoted by F. We call a conjunction of facts $v_1{=}d_1 \wedge \cdots \wedge v_k{=}d_k$ *consistent* if it does not contain any two facts $v_i{=}d_i$ and $v_j{=}d_j$ such that $v_i = v_j$, but $d_i \neq d_j$. We call it a *complete conjunction*, or simply *complete* if it contains a conjunct $v{=}d$ for every variable $v \in \mathcal{V}$. Up to reordering and unnecessary repetitions of conjuncts, there is a unique complete conjunction of facts for every possible assignment of domain values to variables. Therefore, we will often identify those representations. A complete conjunction of facts s is also called a *state*, and S denotes the set of states of Π. The set A is a set of *actions*, where an action is a pair $a = \langle pre, eff \rangle$. The *precondition pre* is a conjunction of facts, and the *effect eff* is a *conditional effect* in effect normal form (ENF) (Rintanen 2003), i. e., a conjunction $eff = eff_1 \wedge \cdots \wedge eff_k$ of sub-effects eff_i of the form $\varphi_i \rhd v_i{:=}d_i$, where φ_i is a conjunction of facts, the *effect condition*, and where $v_i{:=}d_i$ is an atomic effect (a fact). Every atomic effect may occur at most once in eff. We furthermore assume that, whenever eff includes two conjuncts $\varphi_i \rhd v_i{:=}d_i$ and $\varphi_j \rhd v_j{:=}d_j$, and $v_i = v_j$, but $d_i \neq d_j$, then $\varphi_i \wedge \varphi_j$ is inconsistent, to rule out contradictory effects. If some φ_i is the trivial condition $\top$ (true), then the corresponding sub-effect is unconditional, and we write $v{:=}d$ instead of $\top \rhd v{:=}d$. The set of actions A is partitioned into a set A_{endo} of *endogenous actions* and a set A_{exo} of *exogenous actions*. We assume that the set of endogenous actions always contains the *empty action* ϵ, which has an empty precondition and effect, and we assume that each exogenous action is associated with a set of discrete time points $t(a)$ at which it will be automatically applied, provided that its preconditions is satisfied. This is similar in spirit to *timed facts* (Cresswell and Coddington 2003) that are made true exactly at their associated time point. The state $s_0 \in S$ is called the *initial state*, and the partial state $s_\star$ specifies the *goal* condition.

Semantics. An endogenous action $a = \langle pre, eff \rangle$ is applicable in state s iff $s \models pre$, i. e., the precondition pre is satisfied in s. For an exogenous action a to be applicable, we additionally require that s is the t-th state in the state sequence induced by the action sequence under consideration for some $t \in t(a)$. Let $eff = \bigwedge_{i=1}^{k}(\varphi_i \rhd v_i{:=}d_i)$ be an effect in ENF. Then the *change set* (Rintanen 2003) of eff in s, symbolically $[eff]_s$, is the set of facts $\bigcup_{i=1}^{k}[\varphi_i \rhd v_i{:=}d_i]_s$, where $[\varphi \rhd v{:=}d]_s = \{v{=}d\}$ if $s \models \varphi$, and $\emptyset$, otherwise. A change set will never contain two contradicting effects. Now, applying an applicable action a to s yields the state s' that has a conjunct $v{=}d$ for each $v{=}d \in [eff]_s$, and the conjuncts from s for all variables v that are not mentioned in the change set $[eff]_s$. We write $s[a]$ for s'.

For exogenous actions, we assume an *urgent* semantics. More specifically, whenever an exogenous action a_{exo} is applicable and its application in the current state leads to a different successor state, its application is enforced. We furthermore assume that if two or more exogenous actions are applicable in the same state, they do not interfere, i. e., neither of them disables another one, nor do they have conflicting effects. Let s be a state. Then by $\Delta_{\text{exo}}(s)$ we refer to the unique state that is obtained from s by applying all applicable exogenous actions. Since exogenous actions that are applicable in the same time step do not interfere, $\Delta_{\text{exo}}(s)$ is well-defined and is obtained by the application of finitely many exogenous action occurrences. We give the following semantics to a sequence consisting of endogenous actions $\pi = \langle a_0, \ldots, a_{n-1} \rangle$: First we extend the plan by empty actions if $n - 1 < \max \bigcup_{a \in A_{\text{exo}}} t(a)$ until the highest time step of the exogenous actions equals $n - 1$. Assume that the initial state s_0 is already closed under exogenous action application, i. e., that $\Delta_{\text{exo}}(s_0) = s_0$. Then, for $i = 0, \ldots, n - 1$, the next state s_{i+1} is obtained by first applying action a_i to state s_i (assuming that it is applicable), followed by closing under exogenous actions. More formally, $s_{i+1} = \Delta(s_i, a_i) := \Delta_{\text{exo}}(s_i[a_i])$. If a_i is inapplicable in s_i for some $i = 0, \dots, n-1$, then π is inapplicable in s_0.

A state s is a goal state if $s \models s_\star$. We denote the set of goal states by $S_\star$. We call π a *plan* for Π if it is applicable in s_0 and if $s_n \in S_\star$.

Modified semantics for counterfactual reasoning. Below, we will propose a way to answer questions of the form: "What would have happened if we had followed plan π, but without action a being part of π?", or: "What would have happened if $v{:=}d$ had not been an effect of action a?" For that, we want to be able to trace plan π while leaving out a or $v{:=}d$. Unfortunately, with the semantics above, this would often simply mean that the modified plan is no longer applicable. To avoid this, we consider an alternative semantics here. Let $\pi' = \langle a_0, \dots, a_{n-1} \rangle$ be a modified plan, possibly with some actions replaced by the empty action ϵ, or with some effects removed from actions. Let s_0 be the initial state. Then we define, for all $i = 0, ..., n-1$, that $s_{i+1} = \Delta(s_i, a_i)$, if a_i is applicable in s_i, and $s_{i+1} = \Delta(s_i, \epsilon)$, otherwise. In other words, if a_i is applicable in s_i, then we apply it, otherwise, we skip it. Notice that even if a_i remains applicable in s_i in π', the actual effects of a_i may differ from what happens when tracing the original plan π, since some effect conditions of a_i may be satisfied for π, but not for π', or the other way around.

Moral valuations of actions and consequences. Above, we defined the planning formalism we use. To define the possible *dynamics* of the system under consideration, this is sufficient. However, in order to formally capture and reason about the *ethical principles* outlined above, we also need to classify actions and facts with respect to their moral value as either morally bad, indifferent, or good. To that end, in the following, we assume that each planning task Π comes with a utility function u that maps endogenous actions and facts to utility values: $u\colon A_{\text{endo}} \cup F \to \mathbb{R}$.

We let u map to $\mathbb{R}$ instead of just $\{-1, 0, 1\}$ to allow for different degrees of how morally good or bad an action or fact may be. We need this in order to reasonably capture the utilitarian principle. We call an action a or fact f *morally bad* if $u(a) < 0$ or $u(f) < 0$, respectively. Similarly, we call an action or fact *morally indifferent* or *morally good* if its utility value is zero or greater than zero, respectively. Notice that we explicitly do *not* require that moral values of actions and facts must be consistent in any particular sense. For instance, we do not require that an action must be classified as morally bad if one (or all) of its effects are morally bad. The rationale behind this choice is that, in terms of deontology, actions are good or bad *per se*, without regard to their actual effects. We leave enforcing such consistency to the modeler where this is desired, and emphasize that occasionally, such consistency may be explicitly *not* desired.

When using a consequentialist view, we will judge the moral value of a plan by the utility value of its final state, which is defined to be the sum over the utility values of all facts in the final state: $u(s) = \sum_{\{v=d \mid s \models v=d\}} u(v{=}d)$. If we want to consider also the utility value of intermediate states of a plan, one would need to propagate the relevant facts to the final state. This again would be something the modeler is responsible for.

Formalization of Ethical Principles

Moral dilemmas. To exemplify each of the principles and to demonstrate how they come to different judgments about the moral permissibility of plans, we first introduce two versions of the *trolley problem* (Foot 1967). The classical trolley problem is a thought experiment that asks the listener to imagine they were in the following situation: "A runaway trolley is about to run over and kill five people. If you, as a bystander, throw a switch then the trolley will turn onto a sidetrack, where it will kill only one person." Using SAS$^+$, the dynamics of the trolley problem can be modeled as a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, such that:

$$
\begin{aligned}
&\mathcal{V} = \{\mathit{man}, \mathit{men}, \mathit{tram}, \mathit{lever}\}\\
&A_{endo} = \{\mathit{pull}\}, A_{exo} = \{\mathit{advance}\}\\
&\mathit{pull} = \langle \top, \mathit{lever}{=}l \triangleright \mathit{lever}{:=}r \wedge \mathit{lever}{=}r \triangleright \mathit{lever}{:=}l \rangle\\
&\mathit{advance} = \langle \top, \mathit{tram}{=}\mathit{start} \wedge \mathit{lever}{=}r \triangleright \mathit{tram}{:=}r \wedge\\
&\qquad \mathit{tram}{=}\mathit{start} \wedge \mathit{lever}{=}l \triangleright \mathit{tram}{:=}l \wedge\\
&\qquad \mathit{tram}{=}r \triangleright \mathit{men}{:=}\mathit{dead} \wedge \mathit{tram}{=}l \triangleright \mathit{man}{:=}\mathit{dead} \rangle\\
&t(\mathit{advance}) = \{1, 2\}\\
&s_0 = \mathit{man}{=}\mathit{alive} \wedge \mathit{men}{=}\mathit{alive} \wedge \mathit{tram}{=}\mathit{start} \wedge \mathit{lever}{=}r\\
&s_\star = \mathit{men}{=}\mathit{alive}\\
&u(\mathit{pull}) = u(\mathit{lever}{=}l) = u(\mathit{lever}{=}r) = u(\mathit{tram}{=}\mathit{start}) =\\
&u(\mathit{tram}{=}l) = u(\mathit{tram}{=}r) = 0, u(\mathit{man}{=}\mathit{alive}) = 1,\\
&u(\mathit{men}{=}\mathit{alive}) = 5, u(\mathit{man}{=}\mathit{dead}) = -1,\\
&u(\mathit{men}{=}\mathit{dead}) = -5
\end{aligned}
$$

In this model, the variable men models the state of the five persons on the one track (dead or alive), and man models the state of the one person on the other track. The variable tram tracks the position of the tram (start, right track r, left track l), and the variable lever represents the state of the lever (left position l or right position r). There is one endogenous action pull available to the bystander. The action switches the state of the lever. The timed exogenous action $\mathit{advance}$ changes the position of the tram at time points 1 and 2. Deaths are considered morally bad and hence they have negative utility, and survival facts are considered morally good and hence have positive utility. All other facts and actions are considered morally neutral. Depending on the state of the lever, at time point 1, the tram will move from its start position either to the left track or to the right track. At time point 2, if it is on the left track, the tram will hit the one man, and if it is on the right track, it will hit the five men. So, if the bystander's goal was to save the five men, her only chance is to execute pull at time point 0.

The classical trolley problem is often contrasted with the *footbridge trolley problem*, which reads: "A trolley has gone out of control and now threatens to kill five people working on the track. The only way to save the five workers is to push a big man currently standing on the footbridge above the track. The big man will fall onto the track thereby stopping the tram. He will die, but the five other people will survive." The footbridge trolley problem also involves a decision between one death and five deaths. For many people, however, the intuition about what is morally permissible to do turns out to be very different to that in the classical case. The SAS$^+$ model of this scenario is given by a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, such that:

$$
\begin{aligned}
&\mathcal{V} = \{man, men\}, A_{endo} = \{push\}, A_{exo} = \{advance\} \\
&push = \langle man{=}onBridge, man{:=}deadOnTrack \rangle \\
&advance = \langle \top, man{=}onBridge \triangleright men{:=}dead \rangle \\
&t(advance) = \{1\} \\
&s_0 = man{=}onBridge \wedge men{=}alive, s_\star = men{=}alive \\
&u(push) = -1, u(man{=}onBridge) = 1, \\
&u(man{=}deadOnTrack) = -1, u(men{=}dead) = -5, \\
&u(men{=}alive) = 5
\end{aligned}
$$

The variable man represents the state of the big man on the footbridge (either $onBridge$ or $deadOnTrack$), and the variable men represents the state of the five people on the track (either $dead$ or $alive$). The endogenous action $push$ is available to the decision-making agent, who reasons about whether or not to push the big man off the bridge. The timed exogenous action $advance$ changes the state of the tram. Depending on whether or not the big man is on the track, the tram will stop at time point 1 due to its collision with the big man, or it will hit the other five men. We assume that pushing is inherently morally bad, that the fact that the big man is lying dead on the track is morally bad and that him surviving on the bridge is morally good, and that the death of the five men also is morally bad but their survival is morally good. In the modeled situation, the agent's goal is to save the five men.

Deontology and Utilitarianism. The reasoning task of interest is to check possible plans for moral permissibility. To do so, we define moral permissibility of the ethical principles introduced above. We start with the two famous principles *deontology* and *utilitarianism*. The definition of the deontological principle (Def. 1) requires that all actions in a plan are intrinsically morally good or neutral.

Definition 1. *A plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the deontological principle *if and only if* $u(a_i) \geq 0$ *for all* $i = 0, \ldots, n-1$.

Consider the plans $\pi_1 = \langle pull \rangle$ for the classical trolley problem and $\pi_2 = \langle push \rangle$ for the footbridge trolley problem as modeled above. Plan π_1 does not contain any intrinsically bad action, whereas π_2 does. Therefore, according to the deontological principle, π_1 is morally permissible and π_2 is morally impermissible.

The utilitarian principle requires an agent to always do what optimizes moral utility. In the context of action plans, we call a plan morally permissible according to the utilitarian principle iff the final state of the plan is among the morally optimal states.

Definition 2. *A plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the utilitarian principle *if and only if* $u(s_n) \geq u(s')$ *for all reachable states* s'*, where* s_n *is the final state reached by* π.

Given that the $advance$ actions will be executed anyway, the set of reachable states in both the trolley problems boil down to the states reached by acting at time point 0 or by not acting at all. In the classical trolley problem, the two reachable states differ in the number of people dead. In our version of utilitarianism, the number of people harmed is morally relevant. Thus, the plan $\langle pull \rangle$ is morally permissible, but the empty plan is not. Likewise, for the footbridge trolley problem, pushing the big man off the bridge, $\langle push \rangle$, is morally permissible but the empty plan is not.

Harm avoidance. While utilitarianism allows for harm for the greater good, it has been argued that a moral agent should avoid to cause harm at all (Nevejans 2016). We take a counterfactual approach to modeling harm by saying that an action causes harm if, had the action not been performed, the harm would not have happened. The judgment appears to be more difficult when one deliberates about leaving out arbitrary parts of the original plan as opposed to simply one action. If, for example, we have two actions in the plan, one deleting a morally bad effect, which is true in the initial situation, and the second action reinstantiates the morally bad effect, then we have not lost anything compared with the initial situation. However, when executing the plan, we reach a state from which executing the second action leads to some harm. For this reason, we consider a plan only as acceptable according to the do-no-harm principle when we can guarantee that by skipping arbitrary parts we never reach a state where the harm does not hold. Definition 3 captures this idea.

Definition 3. *A plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the do-no-harm principle *if and only if for all facts* $v{=}d$*, if* $s_n \models v{=}d$ *and* $u(v{=}d) < 0$*, then for all plans* π' *obtained by replacing a subset of actions in* π *by empty actions* ϵ*,* $v{=}d$ *still holds in the final state of* π'*, where* s_n *is the final state reached by* π.

According to this definition, the plan $\langle pull \rangle$ for the classical trolley problem is morally impermissible. This is because it makes the morally bad fact $man{=}deadOnTrack$ true, which is false if $pull$ is deleted from the plan. For the analogous reason, the plan $\langle push \rangle$ for the footbridge trolley problem is impermissible, as well. Contrarily, the empty plan is permissible because the harm that results in the final state cannot be avoided by skipping actions.

Definition 4 introduces the Asimovian principle, which is more restrictive than do-no-harm. According to Asimov's first law of robotics, a robot should not cause harm and it should avoid harm to happen (Asimov 1950). Hence, whereas the empty plan is always do-no-harm permissible, doing nothing is impermissible according to the Asimovian principle if there exists a plan which prevents the harm from holding in the final state. Particularly, in both the trolley problems, no morally permissible plan exists according to the Asimovian principle.

Definition 4. *A plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according the Asimovian principle *if and only if for all* *facts $v{=}d$, if $s_n \models v{=}d$ and $u(v{=}d) < 0$, then there is no alternative plan π', such that $s'_{n'} \not\models v{=}d$, where s_n and $s'_{n'}$ are the final states reached by π and π', respectively.*

We can derive Proposition 1 that asserts that Asimovian permissibility entails do-no-harm permissibility.

Proposition 1. *Every Asimovian-permissible plan is do-no-harm permissible.*

Proof. Let π be an Asimovian-permissible plan. If no harmful fact holds in the final state, then π is also do-no-harm permissible. Otherwise, let $v{=}d$ be some harm which holds in the final state of π. By definition of Asimovian permissibility there is no alternative plan π' which prevents $v{=}d$. Particularly, no sub-plan of π can prevent $v{=}d$ to hold in the final state. Therefore, π is do-no-harm permissible. □

Instrumentality. A reasonable variation of the do-no-harm principle is the do-no-instrumental-harm principle defined in Def. 6. The idea is that harm is permissible in case it is not committed as a means to one's end but only occurs as side effect. Definition 5 captures this means-end relation formally.

Definition 5. *For a given plan π with final state s_n, a fact $v_m{=}d_m$ is called a* means to the end *$v_e{=}d_e$ ($s_\star \models v_e{=}d_e$) if and only if $s_n \models v_e{=}d_e$ and there exists a subset of actions in π such that after deleting the effect $v_m{:=}d_m$ from these actions, the resulting plan π' leads to a final state s'_n s.t. $s'_n \not\models v_e{=}d_e$.*

Definition 6. *A plan $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ is* morally permissible according to the do-no-instrumental-harm principle *if and only if for all facts $v{=}d$, if $s_n \models v{=}d$ and $u(v{=}d) < 0$, then $v{=}d$ is not a means to an end.*

Definition 5 involves considering all possible subsets of effect appearances in a plan. To see that this is necessary, assume that in a plan the electric light is switched on, illuminating the room, i.e., *roomIlluminated*=⊤. Further, a candle is lit, which also illuminates the room. One of the goals is to make an object in the room visible, i.e., *object*=*visible*, which happens, if the room is illuminated. If we now check counterfactually whether the fact *roomIlluminated*=⊤ is a means to achieve *object*=*visible*, it is not clear, for which action we should delete the fact *roomIlluminated*=⊤. Moreover, regardless of which effect we delete, the object will still be visible. Only if we delete both effects in the plan, then the object is not any longer visible. So, one could argue that the above definition should be modified by requiring that *all effects* in the plan of the form $v_m{=}d_m$ should be deleted in order to check whether $v_m{=}d_m$ is a means to achieve $v_e{=}d_e$. This requirement appears to be too strict, however. Assume a toggle switch action that has an effect *pressed*=⊤, which in turn leads through an exogenous action to toggling the light and resetting the pressed status, i.e., *pressed*=⊥. Assume two of these actions are executed in a plan. Removing all *pressed*=⊤ effects will not change the status of the light in the end, but only one removal will change the status of the light in the final state.

According to the definition of the do-no-instrumental harm principle, the plan $\langle pull \rangle$ in the classical trolley dilemma is permissible. This is because the bad effect $man{=}dead$ is not a means to the end $men{=}alive$: If, counterfactually, $man{=}dead$ was not an effect of the actions in the plan, then still $men{=}alive$ would finally hold. Contrarily, in the footbridge trolley problem, if, counterfactually, $man{=}deadOnTrack$ was not an effect of $push$, the goal $men{=}alive$ would not finally hold. Hence, the plan $\langle pull \rangle$ is morally permissible according to the do-no-instrumental harm principle, and $\langle push \rangle$ is not.

One is tempted to think that the do-no-instrumental-harm principle is at least as tolerant as the do-no-harm principle, i.e., every do-no-harm permissible plan is also do-no-instrumental-harm permissible. However, this is not so: Imagine a situation where $\pi = \langle a_0 \rangle$ is executed, initially $s_0 \models p{=}\bot \wedge h{=}\bot \wedge g{=}\bot$, action a_0 has $p{=}\top \wedge h{=}\top$ as an effect, $u(h{=}\top) = -1$, and the exogenous action e_0 makes the goal fact $g{=}\top$ hold in the final state if either $p{=}\top \wedge h{=}\top$ or $p{=}\bot \wedge h{=}\bot$ holds. Moreover, e_0 makes $h{=}\top$ true in any case. In this case, deleting the harmful fact $h{=}\bot$ from the effects of a_0 leads to the goal not to hold in the final state. Hence, the harm is a means to an end and thus do-no-instrumental-harm impermissible. However, as the harm cannot be prevented by not executing a_0 at all, the plan π is do-no-harm permissible.

Finally, we define the principle of double effect in Def. 7, which contains many of the above principles.

Definition 7. *A plan $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ is* morally permissible according to the double-effect principle *if and only if all of the following conditions are satisfied:*

1. *The plan π is morally permissible according to the deontological principle.*
2. *At least one goal fact $v{=}d$ satisfies $u(v{=}d) > 0$.*
3. *No goal fact $v{=}d$ satisfies $u(v{=}d) < 0$.*
4. *The plan π is morally permissible according to the do-no-instrumental-harm principle.*
5. *$u(s_n) > 0$, where s_n is the goal state reached by π.*

Hence, the principle of double effect contains the deontological principle as its first condition and the do-no-instrumental-harm principle as the fourth condition. The second and third conditions are constraints on the goal of the planning agent: She is not allowed to have morally bad goals, and the goal should contain something morally good. The last condition is a weaker form of utilitarianism, which requires that all in all the plan brings about more good facts than bad facts—but unlike utilitarianism, it does not require the plan's final state to be among the optimal states.

In case of the footbridge trolley problem, the first condition renders pushing the man off the bridge impermissible. However, the second and third conditions are fulfilled, because the goal of the agent only consists of one fact, viz., $men{=}alive$, and this fact is morally good. The fourth condition also is violated as we have already discussed above. The fifth condition is fulfilled, because, all in all, the good consequences yield more positive utility than the negative consequence add negative utility. Hence, using the principle of double effect, the reasoner can explain that there are two reasons why the plan $\langle push \rangle$ is morally impermissible:

Because pushing is morally bad, and because the death of the big man is used as a means. For the case of the classical trolley problem, the principle of double effect comes to the conclusion that the plan $\langle \mathit{pull} \rangle$ is morally permissible.

Ethical Validation of Action Plans

The output of a planning algorithm is a sequence of actions $\pi = \langle a_0, \dots, a_{n-1} \rangle$ and a final state s_n. Our goal is to ethically evaluate a given action plan. To this end, we here describe procedures that take a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, the utility function u, a plan π, its final state s_n, and one of the introduced ethical principles as the input and decide whether or not the principle renders the plan as morally permissible.

To check whether or not a given plan π is morally permissible according to the deontic principle (Def. 1), it needs to be checked if some of the actions in π are intrinsically bad, i.e., if for one of the action a_i in π, we have $u(a_i) < 0$. This can be apparently done in time linear in the length of π.

Proposition 2 (Deontic Validation)**.** *Deciding whether a plan is morally permissible according to the deontic principle can be done in polynomial time.*

A procedure for verifying that π is morally permissible according to the utilitarian principle (Def. 2) is much more involved than checking deontological permissibility. Recall that the utilitarian principle only permits plans that lead to reachable states with maximum utility. In so far, this is very similar to over-subscription planning (Smith 2004). Based on that, we can formulate a non-deterministic procedure for deciding the complement of the permissibility problem as follows: Compute the overall utility of s_n. Then guess another complete state s' with utility that is larger than the utility of s_n. Finally generate (non-deterministically) a plan π' to achieve s'. If successful, it demonstrates that π is not morally permissible. That this is indeed an (asymptotically) optimal procedure is shown by the following theorem.

Theorem 1 (Utilitarian Validation)**.** *Deciding whether a plan is morally permissible according to the utilitarian principle is PSPACE-complete.*

Proof. PSPACE membership follows from the arguments above, and the facts that PSPACE is closed under complement and non-determinism and that deciding plan existence is in PSPACE. PSPACE-hardness follows straight-forwardly from a reduction of plan existence in SAS$^+$ planning. Given a SAS$^+$ planning task Π, generate a new task Π' by extending the set of variables by two Boolean variables g_1 and g_2, which are both assumed to be false in s_0. Extend the set of actions by two new endogenous actions: $a_1 = \langle \top, g_1{:=}\top \rangle$ and $a_2 = \langle s_\star, g_2{:=}\top \rangle$. The new goal description of Π' is $s_\star = g_1{=}\top$. The utility function is identical to zero on all actions and facts except for g_1 and g_2, where it evaluates to 1. Clearly, the only possible plan is $\langle a_1 \rangle$ leading to state s with $u(s) = 1$. This plan is impermissible according to the utilitarian principle iff there exists a plan for the original task Π because in this case we could reach a state s' for Π' such that $u(s') = 2$. □

To check whether a given plan π is morally permissible according to the do-no-harm principle (Def. 3), we have to verify that no parts of the plan lead to avoidable harm. A non-deterministic algorithm for deciding impermissibility could be: We guess one fact $v_b{=}d_b$ with $u(v_b{=}d_b) < 0$ and a subplan π' of π leading to s' and then verify that $s_n \models v_b{=}d_b$ but $s' \not\models v_b{=}d_b$.

Theorem 2 (Do-No-Harm Validation)**.** *Deciding whether a plan is morally permissible according to the do-no-harm principle is co-NP-complete.*

Proof. The sketched non-deterministic algorithm demonstrates membership in co-NP. In order to show hardness, we use a reduction from 3SAT to the impermissibility problem. Assume a 3SAT problem over the variables $v_1, \dots, v_n$ and clauses $c_1, \dots, c_m$, where each clause consists of 3 literals l_{j1}, l_{j2}, l_{j3}. We now construct a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, where $\mathcal{V} = \{b, g, v_1, \dots, v_n, c_1, \dots, c_m\}$, $A = \{V_1, \dots, V_n, C_1, \dots, C_m, G, B\}$, $s_0 = \{v{=}\bot \mid v \in \mathcal{V}\}$, and $s_\star = \{g\}$. The actions are defined as follows: $V_i = \langle \top, v_i{:=}\top \rangle$, $C_j = \langle \top, \bigwedge_{k=1}^{3}(l_{jk} \rhd c_j) \rangle$, where $l_{jk} \equiv v_{jk}{=}\top$ if the literal l_{jk} in the original SAT problem is positive, otherwise, $l_{jk} \equiv v_{jk}{=}\bot$. Further, $G = \langle \top, g{:=}\top \wedge (\bigwedge_{j=1}^{m} c_j \rhd b{:=}\bot) \rangle$, $B = \langle \top, b{:=}\bot \rangle$. All facts have zero utility except for $b{=}\bot$, which is valued -1. The plan we want to check is $\pi = \langle V_1, \dots, V_n, C_1, \dots, C_m, G, B \rangle$. This plan obviously achieves the goal and the final state contains some harm. Moreover, the only way to avoid this harm is to delete action B. However, even without this action, we still may have harm. This harm can be avoided, if and only if we can delete a (perhaps empty) subset of the V_i actions corresponding to a variable assignment of the 3SAT problems that satisfies the original 3SAT formula, which demonstrates that impermissibility is co-NP-hard. □

For the Asimovian principle, for each harm in the final state, we have to check whether there is a plan, which avoids that harm. As for the utilitarian principle, this quantifies over all available plans, and hence checking Asimovian permissibility has the same computational complexity as utilitarian permissibility.

Theorem 3. *Deciding whether a plan is morally permissible according to the Asimovian principle is PSPACE-complete.*

Proof. We first prove membership by presenting a procedure which uses polynomial space: Consider as input a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ and a plan π. As a first step, the execution of π is simulated to obtain the final state s_n, whose size is bound by $|\mathcal{V}|$. For all harmful facts $v{=}d$ (viz., with $u(v{=}d) < 0$) that hold in s_n, a planner is used to solve the planning task $\Pi' = \langle \mathcal{V}, A, s_0, \neg v{=}d \rangle$, i.e., to see if there exists a plan, which makes $v{=}d$ false. Plan existence is known to be decidable in polynomial space. To show hardness, we reduce existence of propositional STRIPS (with one goal literal) to Asimovian permissibility: Let $\Pi = \langle \mathcal{V}, A, s_0, v_g{=}d_g \rangle$ be a propositional STRIPS planning instance. Set $u(\neg v_g{=}d_g) = -1$ and $u(v) = 0$ for all other facts. The empty plan $\pi_\epsilon = \langle\rangle$ is morally permissible according to the Asimovian principle iff there exists a plan that solves Π: If $v_g{=}d_g$ holds in the initial state, then the empty plan is both permissible and a solution to the plan-existence problem. Otherwise, if $\neg v_g{=}d_g$ holds in the initial state, then, by definition of the Asimovian principle, the empty plan is impermissible if and only if there exists plan which finally makes $v_g{=}d_g$ true. □

For the do-no-instrumental-harm principle (Def. 6), we can use a very similar method to checking for the do-no-harm principle. Instead of skipping subsets of actions, we have to delete subsets of effect occurrences in the plan. Hence, checking this principle for a given plan has the same computational complexity.

Theorem 4. *Deciding whether a plan is morally permissible according to the do-not-instrumental-harm principle is co-NP-complete.*

Proof. One can use the same non-deterministic algorithm as for the do-no-harm principle, demonstrating that deciding permissibility of plan for this principle is again in co-NP. For hardness, we can use a reduction very similar to the one in the last theorem. Instead of deleting actions we would delete effects, which are used to enable the execution of exogenous actions that regulate the assignment of the variables. □

Finally, we consider the double-effect principle. Except for the fourth condition, everything can be checked in polynomial time. The fourth condition is just the do-not-instrumental-harm principle. In other words, deciding permissibility for this principle is in co-NP.

Theorem 5. *Deciding whether a plan is morally permissible according to the double-effect principle is co-NP-complete.*

Proof. Membership is obvious. Hardness follows with the same proof as above by setting $u(g) = 2$. □

Related Work

While there exists a number of papers on machine ethics, papers that focus on generating and/or validating plans according to ethical principles are scarce.

Dennis et al. (2016) propose to establish ethical principles and ethical rules that judge the severity of violating ethical principles, whereby an ethical principle could be not to harm a human. Plans can then be ordered by comparing the worst violations of these plans. While this has an deontological flavor, in fact, plans are judged according to their ultimate consequences, and hence this appears to be a consequentialist approach. The authors do not consider the distinction between causing harm and causing instrumental harm.

Pereira and Saptawijaya (2017) use abductive logic programming in order to specify the principle of double effect and to evaluate some of the trolley scenarios. Berreby et al. (2015) similarly use logic programming (in this case ASP) in order to specify the principle of double effect and evaluate on trolley scenarios described using the event calculus. In this case, however, they do not use counterfactual reasoning to judge causality, but they use simple syntactical means to determine what is a cause of an effect. Govindarajulu and Bringsjord (2017) propose a general framework to create or verify that an autonomous system is compliant to the double effect principle. For this purpose they introduce a powerful logical formalism called *deontic cognitive event calculus*. In particular, they propose a formalization of the notion of *means to an end* in a STRIPS framework, which however does not take into account that different actions in a plan can contribute to different parts of a goal, and which does not consider that combinations of actions can be causes. Weld and Etzioni (1994) propose two versions of a do-no-harm principle for action plans. Their do-no-harm principle is fine with harm in the final state given that the harm already held in the initial state. Our do-no-harm principle does not permit to heal harm first just to reintroduce it later on. However, our formulation allows to cause harm if it is healed later on. This is not allowed in one version of Weld and Entzioni's account. We can, however, generate this behavior by introducing special harm facts into the model that become true when harm happens during plan execution and that remain true forever. Interestingly, none of the papers mentioned above address the issue that evaluating the moral permissibility of action plans might require a counterfactual analysis that is combinatorial in nature.

Conclusions

We formalized various ethical principles, which take different aspects of a plan to be morally significant. Deontology stresses the moral value of action tokens, utilitarianism requires utility optimization, the do-no-harm principle and the Asimovian principles strive for avoiding avoidable harm, and the do-no-instrumental-harm principle and the principle of double effect take serious the intuition that harm should not be used as a means to an agent's end but may be acceptable as a mere side effect.

We studied these principles in the context of *action sequences*, as opposed to the more usual way of studying them in the context of *individual actions*. Only in this way we can analyze moral permissibility of entire plans, since it is not sufficient to judge the moral permissibility of each action in isolation, but also in the context of the whole plan. We exemplified and explained our formalizations using classical moral dilemmas such as the trolley problem. Furthermore, we studied the computational complexity of verifying whether a given plan is permissible with respect to each of the five investigated principles. We saw that, with respect to our formalization, verification is PSPACE-complete for utilitarianism and for the Asimovian principle, co-NP-complete for do-no-harm, for do-no-instrumental-harm, and for the principle of double effect, and that it is polynomial-time for deontology. Verifying the do-no-harm principles involves a combinatorial reasoning over possible *sets* of actions that lead to harm or that may be instrumental towards achieving a goal condition, which makes verifying those ethical principles surprisingly hard.

We believe that our work has the potential of being useful in making autonomous systems ethical by providing them with the capability of coming up with morally permissible plans or at least being able to judge ethical permissibility of given plans.

References

Anderson, M., and Anderson, S. L., eds. 2011. *Machine Ethics*. Cambridge, UK: Cambridge University Press.

Anderson, M.; Anderson, S. L.; and Armen, C. 2005. Machine Ethics: Papers from the AAAI Fall Symposium. Technical report, AAAI Press.

Asimov, I. 1950. *Runaround*. Gnome Press.

Bäckström, C., and Nebel, B. 1995. Complexity results for SAS$^+$ planning. *Computational Intelligence* 11(4):625–655.

Berreby, F.; Bourgne, G.; and Ganascia, J. 2015. Modelling moral reasoning and ethical responsibility with logic programming. In *Proceedings of the 20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2015)*, 532–548.

Cresswell, S. N., and Coddington, A. M. 2003. Planning with timed literals and deadlines. In *Proceedings of the 21st Workshop of the UK Planning and Scheduling SIG*, 22–35.

Dennis, L. A.; Fisher, M.; Slavkovik, M.; and Webster, M. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77:1–14.

Driver, J. 2006. *Ethics: The Fundamentals*. Hoboken, NJ: Wiley-Blackwell.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review*.

Fox, M.; Howey, R.; and Long, D. 2005. Validating plans in the context of processes and exogenous events. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, 1151–1156.

Ghallab, M.; Nau, D. S.; and Traverso, P. 2016. *Automated Planning and Acting*. Cambridge University Press.

Govindarajulu, N. S., and Bringsjord, S. 2017. On automating the doctrine of double effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 4722–4730.

Lindner, F.; Bentzen, M. M.; and Nebel, B. 2017. The HERA approach to morally competent robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017)*, 6991–6997.

Nevejans, N. 2016. European civil law rules in robotics.

Pereira, L. M., and Saptawijaya, A. 2017. Agent morality via counterfactuals in logic programming. In *Proceedings of the CogSci 2017 Workshop on Bridging the Gap between Human and Automated Reasoning - Is Logic and Automated Reasoning a Foundation for Human Reasoning?*, 39–53.

Rintanen, J. 2003. Expressive equivalence of formalisms for planning with sensing. In *Proceedings of the 13th International Conference on Automated Planning and Scheduling (ICAPS 2003)*, 185–194.

Smith, D. E. 2004. Choosing objectives in over-subscription planning. In *Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS 2004)*, 393–401.

Weld, D. S., and Etzioni, O. 1994. The first law of robotics (A call to arms). In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI 1994)*, 1042–1047.