Generating preferred plans with ethical features

Martin Jedwabny,[1] **Pierre Bisquert,**[1, 2] **Madalina Croitoru**[1]
[1]LIRMM, Inria, Univ Montpellier, CNRS, Montpellier, France [2]IATE, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
martin.jedwabny@lirmm.fr, pierre.bisquert@inrae.fr, madalina.croitoru@lirmm.fr

Abstract

Normative ethics has been shown to help automated planners take ethically aware decisions. However, state-of-the-art planning technologies don't provide a simple and direct way to support ethical features. Here, we propose a new theoretical framework based on a construct, called ethical rule, that allows to model preferences amongst ethically charged features and capture various ethical theories. We show how the framework can model and combine the strengths of these theories. Then, we demonstrate that classical planning domains extended with ethical rules can be compiled into soft goals in PDDL.

Introduction

We place ourselves in the intersection of normative ethics and automated planning. Past research in this subject aimed to apply ideas from the field of normative ethics, the subfield of ethics that studies the admissibility of actions, to make autonomous agents take into account the decision process behind diverse ethical theories (Berreby, Bourgne, and Ganascia 2017; Lindner, Mattmüller, and Nebel 2019; Cointe, Bonnet, and Boissier 2016; Dennis and Fisher 2018). Still, none provide a direct way to support ethical features in PDDL (Gerevini et al. 2009) which profits from its state-of-the-art planning algorithms. As a result, it has given rise to many controversial arguments around the manner in which ethics should be embedded into AI systems (Brundage 2014). One of which is whether the welfare of society can be interpreted as a maximization of perceived utility in the context of AI systems. Here, we will focus on this last point by taking planning problems featuring ethical elements and reducing them to a maximization of utilities in such a way that it can profit from PDDL planners.

Concretely, our model gives autonomous systems the ability to capture (i) the *ethical features* considered in a decision making or classical planning problem, and (ii) *qualitative model* to represent their relative importance, and determine what sequence of actions is the best, capturing and combining the intuitions prescribed by well-known ethical theories.

First, we show how to extend *classical planning problems* like tasks with ethical rules, a construct that models the conditions under which certain actions or plan outcomes have an ethical feature to be taken into account. This construct is based on (Cointe, Bonnet, and Boissier 2016), but is adapted for STRIPS like domains and extended with ranks i.e. levels of importance. We chose to use ranks as in (Feldmann, Brewka, and Wenzel 2006), as they provide a simple and direct way of modeling qualitative preferences amongst plans presenting ethical features. We call this extended model STRIPS*E. Then, we show how the three main normative ethical theories, namely consequentialist, deontological and virtue ethics can be represented with our framework.

Second, we demonstrate how STRIPS*E planning tasks can be compiled into STRIPS*U planning tasks i.e. tasks with soft goals and utility-based preferences.

Third, we implement our notions by defining a novel extension for STRIPS like tasks encoded using the PDDL3 (Gerevini et al. 2009) language, that models our ethical rules. We provide an implementation that compiles these tasks into PDDL tasks with utility-based preferences.

Finally, we showcase how a state-of-the-art planner (Coles and Coles 2011) can be applied to STRIPS*E tasks by compiling away the ethical rules.

Planning framework

To represent planning tasks, we will use a STRIPS-like representation (STRIPS*), which will correspond to classical STRIPS tasks (Fikes and Nilsson 1971), but extended with conditional effects as in ADL (Pednault 1989).

A **STRIPS* problem** is a 4-tuple $T = \langle F, s_0, s_*, A\rangle$ that describes all the relevant information that characterizes the states of the domain, the changes actions bring in the form of transitions between states, the initial state and the final conditions a plan has to reach. More precisely:

- F is a finite set of propositions called **fluents**, which represent the characterizing properties of a state.
- s_0 denotes the **initial state**. A *state* is a set of fluents in F that correspond to the propositions that hold in a particular state. All fluents that are not present are considered to be false in a state.
- s_* is a set of fluents called the **goal conditions**.
- A is a finite set of **actions**, also called operators, in the form of pairs $a = \langle Pre(a), Eff(a)\rangle$, which consist of a set of fluents $Pre(a)$ denoting the preconditions,

and a finite set of effects $Eff(a)$. An effect is a pair $\langle Cond, Aff \rangle$ where $Cond$ is a set of fluents called the effect condition, and Aff is a set of literals denoting the affected fluents. A *literal* is either a fluent $f \in F$, or its negation $\neg f$.

Given a state s and an action a, the **successor state** $Succ_T(a, s)$ obtained by applying a is defined, or *possible*, iff $Pre(a) \subseteq s$ and there are no two effects $e_1, e_2 \in Eff(a)$ such that $e_1 = \langle Cond_1, Aff_1 \rangle$, $e_2 = \langle Cond_2, Aff_2 \rangle$, $Cond_1 \subseteq Pre(a)$, $Cond_2 \subseteq Pre(a)$, $f \in Aff_1$ and $\neg f \in Aff_2$. If an action is indeed possible, for every fluent $f \in F$, it holds that $f \in Succ_T(a, s)$ iff there is some effect $\langle Cond, Aff \rangle \in Eff(a)$ such that $Cond \subseteq s$ and $f \in Aff$, or $f \in s$ and there is no effect $\langle Cond, Aff \rangle \in Eff(a)$ such that $Cond \subseteq s$ and $\neg f \in Aff$.

A **plan** is a sequence of actions $\pi = [a_0, a_1, \ldots, a_n]$ with $n \geq 0$ and $a_0, a_1, \ldots, a_n \in A$ which satisfies the goal conditions. The final state of a plan is defined as $Succ_T(\pi, s_0) = Succ_T(a_n, Succ_T(\ldots, Succ_T(a_1, Succ_T(a_0, s_0))))$.
A plan satisfies the goal conditions if and only if $s_* \subseteq Succ_T(\pi, s_0)$.

A STRIPS-like problem with utility based soft goals, denoted **STRIPS*U**, is a tuple $T = \langle F, s_0, s_*, A, u \rangle$, where $\langle F, s_0, s_*, A \rangle$ is a STRIPS* problem, and u is a partial function $u : Form(F) \mapsto \mathbb{R}^+$ that maps propositional formulas (called the soft goals) into positive reals. We denote $Form(F)$ the set of all propositional formulas that can be constructed using the propositions from F, the negation symbol ($\neg$) and the conjunction symbol ($\wedge$). We will also use the consequence symbol ($\models$) as usual in propositional logic. The utility of a state s is obtained as the sum of the utilities of its soft goals: $u(s) = \sum_{\phi \in Dom(u) : s \models \phi} u(\phi)$.

Likewise, the utility of a plan π corresponds to the utility of its final state i.e. $u(\pi) = u(Succ_T(\pi, s_0))$.

An **optimal** plan π for a STRIPS*U problem is one for which no other plan π' has a higher utility.

Both STRIPS* and STRIPS*U planning problems are captured and extended by the PDDL3.0 planning language (Gerevini et al. 2009) featured in the IPC5.

Example 1. *(Hospital) An autonomous vehicle is tasked to get its passengers quickly from their house to a hospital as one of them has suffered an injury. The vehicle can get to the hospital either through a highway (fast) or a normal road (slow). To take the highway, the vehicle has to pass through a toll and present its id. If it presents its own id 'A', it will have to pay a fine, as the id is not authorized for this highway. If the vehicle presents another id 'B' i.e. if it lies about its identity, no fine will be paid, but someone else will have to pay for it.*

We can represent this problem with a STRIPS task $\langle F, s_0, s_*, A \rangle$ as follows:*

- $F = \{atHouse, atToll, atRoad, atHospital, atHighway, barrierOpen, presentedIdA, presentedIdB, tookHighway\}$,
- $A = \{a_0 = \langle \{atHouse\}, \{\langle \emptyset, \{atRoad\} \rangle\} \rangle,$
  $a_1 = \langle \{atRoad\}, \{\langle \emptyset, \{atHospital\} \rangle\} \rangle,$
  $a_2 = \langle \{atHouse\}, \{\langle \emptyset, \{atToll\} \rangle\} \rangle,$
  $a_3 = \langle \{atToll, barrierOpen\},$
  $\{\langle \emptyset, \{atHighway, tookHighway\} \rangle\} \rangle,$
  $a_4 = \langle \{atHighway\}, \{\langle \emptyset, \{atHospital\} \rangle\} \rangle,$
  $a_5 = \langle \{atToll\},$
  $\{\langle \emptyset, \{barrierOpen, presentedIdA\} \rangle\} \rangle,$
  $a_6 = \langle \{atToll\},$
  $\{\langle \emptyset, \{barrierOpen, presentedIdB\} \rangle\} \rangle\}$,
- $s_0 = \{atHouse\}$,
- $s_* = \{atHospital\}$.

We can consider following three plans:

- $\pi_1 = \langle a_0, a_1 \rangle$ *i.e. take the normal road,*
- $\pi_2 = \langle a_2, a_5, a_3, a_4 \rangle$ *i.e. take the highway with id 'A', and*
- $\pi_3 = \langle a_2, a_6, a_3, a_4 \rangle$ *i.e. take the highway with id 'B'.*

Representing ethical features

In the context of decision making and planning for autonomous systems, ethics can be used to imbue agents with ethical values and a theory of the right. Yet, autonomous systems deal with problems that are fundamentally different from real-life human decision making tasks. For once, AI systems rely on frameworks, such as the one presented above, that replicate real-life scenarios with several extensions or simplifications that depend on the system designer.

This is why, our approach separates the process of determining ethically correct choices into two steps. The first is recognizing the ethical features of a plan. As such, a set of rules that characterize when an action in a particular context entails a feature (such as stealing, sharing, or killing) that should be judged on ethical terms and its relative level of importance, must be given as an input. The second is applying these features, and it consists of taking the ethical features induced by a sequence of actions, and using them to compare the possible plans an autonomous agent might take. By doing this, we are able to separate the (model based) action selection process from the ethical reasoning of the agent.

Recognizing ethical features

We introduce a construct that permits to recognize the states and actions that should be judged ethically. We denote E the set of **ethical features**, i.e. the ethical characteristics that an action entails e.g. $E = \{killing, lying, stealing\}$.

For the following, let $T = \langle F, s_0, s_*, A \rangle$ be a planning domain and E a set of ethical features.

Definition 1. *An **ethical rule** is a triple $r = \langle Id(r), Pre(r), Act(r) \rangle$, where:*

- $Id(r) \in E$ *is the identifier i.e. the ethical feature of r,*
- $Pre(r)$ *is a set of fluents of F called the preconditions,*
- $Act(r) \in A \cup \{\text{final}\}$*, is either an action or the constant symbol* final*, called the activation condition.*

An ethical rule defines the conditions under which it is necessary to judge an action ethically. These conditions are represented by the precondition $Pre(r)$ and the activation condition $Act(r)$. The intuition behind them is that a plan that goes from a state s_i satisfying $Pre(r)$ to another state by executing an action specified by $Act(r)$, is assigned the ethical feature $Id(r)$. Moreover, in the special case $Act(r) = \text{final}$, the feature is assigned when the final state of the plan satisfies $Pre(a)$.

Let $R = \{r_1, r_2, \ldots, r_k\}$ with $k \in \mathbb{N}_0$ be a set of ethical rules, and $\pi = [a_0, a_1, \ldots, a_n]$ a plan passing through states $s_0, s_1, \ldots, s_{n+1}$ where $s_{i+1} = Succ_T(a_i, s_i)$:

Definition 2. *The set of ethical features assigned to π with respect to the planning domain T and the ethical rules R is denoted $E_T^R(\pi)$, or more concisely E_π, and defined as:*

$$E_T^R(\pi) = \{Id(r) : r \in R, \text{ and } (\exists i \in \{1, 2, ..., n\} \text{ such that } Pre(r) \subseteq s_i \text{ and } Act(r) = a_i), \text{ or } (Pre(r) \subseteq s_{i+1}, \text{ and } Act(r) = \text{final})\}$$

Capturing ethical theories

In what follows, we will exemplify how recognizing ethical features in a plan allows to extract the information required by different ethical theories, and then compare plans with respect to their ethical characteristics.

All three main branches of normative ethics, namely consequentialism, deontological ethics and virtue ethics have been studied to some degree in the context of automated planning. Some works focus on particular theories, while others more closely related to this work, try to combine the mechanisms of several of them as in (Cointe, Bonnet, and Boissier 2016; Lindner, Bentzen, and Nebel 2017; Lindner, Mattmüller, and Nebel 2019; Bonnemains, Saurel, and Tessier 2016; Berreby, Bourgne, and Ganascia 2017).

For recent surveys on existing implementations and challenges of applying ethical concepts into artificial intelligence systems, we refer the reader to (Yu et al. 2018; Dennis and Fisher 2018).

Consequentialist ethics In this theory, actions are evaluated upon their consequences. The precise method to determine which action is right varies between branches of consequentialist ethics. Some of the most prominent contrast points are the way in which consequences are determined, the perspective from which consequences are evaluated, and how consequences are compared. For an overview of consequentialism's branches, refer to (Haines 2006).

Here, the perspective from which consequences are determined will be the welfare of society, also called utilitarianism. In addition, we consider that an action is better than another if the overall consequences are better (Singer 1977).

To represent consequentialism in our framework, we introduce an ethical rule for each ethically relevant fluent.

Example 2. *(Hospital continued) The ethical rules that characterize the overall ethically relevant consequences of a plan in T can be modeled as:*

$$R_{con} = \{r_0 = \langle fast, \{tookHighway\}, \text{final}\rangle, r_1 = \langle paysFine, \{presentedIdA\}, \text{final}\rangle\}.$$

Then, the ethical features assigned to the plans are $E_{\pi_1} = \emptyset$, $E_{\pi_2} = \{fast, paysFine\}$, and $E_{\pi_3} = \{fast\}$.

Deontological ethics It asserts that an action should be judged on whether it complies with a set of duties and obligations, rather than based on the consequences of the action. As such, deontological ethics is applied to automatic systems by constructing and enforcing restrictions that characterize what is permitted and what forbidden.

Example 3. *(Hospital continued) The main restriction we would want to impose deals with lying about the agent's identity when passing by the toll to access the highway. This can be modeled as follows using our framework:*

$R_{deo} = \{r_2 = \langle lying, \{atToll\}, a_6\rangle\}$.

Then, the ethical features assigned to the plans presented before are $E_{\pi_1} = \emptyset$, $E_{\pi_2} = \emptyset$, and $E_{\pi_3} = \{lying\}$.

Virtue ethics In contrast to the other theories, virtue ethics relies on the moral values of an agent. As such, an agent is deemed ethical when it acts according to some moral values e.g. fairness, honesty and compassion. An agent that reasons ethically according to this theory should exhibit the characteristics of a virtuous agent i.e. perceived to favor others.

Example 4. *(Hospital continued) The main property we wish to capture is the virtue behind carrying an injured person to the hospital as fast as possible i.e. compassion. Also, we can say that only by presenting the id 'A' at the toll we are being honest.*

$R_{vir} = \{r_3 = \langle honesty, \{atToll\}, a_5\rangle, r_4 = \langle compassion, \{atToll\}, a_3\rangle\}$.

Then, the ethical features assigned to the plans presented before are $E_{\pi_1} = \emptyset$, $E_{\pi_2} = \{compassion, honesty\}$, and $E_{\pi_3} = \{compassion\}$.

A model for ethical preferences

We turn our attention to the challenge of providing a framework that allows to reason with conflicting ethical features from one or many different ethical theories.

An important requirement we demand of our language is that it must allow for qualitative preferences. It has been emphasized (Brundage 2014) that AI systems in which ethics is useful, can take dangerous decisions in situations of extreme trade-offs. This could be a problem if the preference language was strictly quantitative. For instance, we want to be able to model that a set of ethical rules R takes precedence over an arbitrarily large set of rules R' whenever R presents a critical rule $r \in R$ that precedes features in R'.

Here, we will use ranked knowledge bases (Feldmann, Brewka, and Wenzel 2006), as our preference representation model, as it combines naturally with our ethical rules and is concise and easy to elicit from external sources. A ranked knowledge base is a model to represent qualitative preferences amongst sets of alternatives. In contrast to the original work, we will not define the preferences for arbitrary formulae, but rather for ethical rules.

Definition 3. *An ethical ranked base (ERB) is a function $erb(r) = \langle Type(r), Rank(r)\rangle$ that maps an ethical rule $r \in$*

R to a pair consisting of a symbol $Type(r) \in \{+, -\}$ representing the type i.e. whether activating the rule is ethically right or wrong, and a non-negative integer $Rank(r) \in \mathbb{N}$, which denotes the rank of the rule i.e. its level of importance.

The idea behind the type and the rank of an ethical rule is to make it possible to compare plans on ethical terms with respect to the rules they satisfy or break.

Definition 4. *Given a STRIPS* problem T, a set of ethical rules R, an ethical ranked base erb over R, and π a plan of T, let $R_i(\pi) = \{r \in R : (Id(r) \in E_\pi \iff Type(r) = +)$ and $Rank(r) = i\}$ for $i \in \mathbb{N}$, then π is at least as preferred as another plan π' of T, denoted $\pi \succeq_{erb} \pi'$ iff:*

$$\forall i \in \mathbb{N}, \text{ it holds that } R_i(\pi) = R_i(\pi'), \text{ or}$$
$$\exists i \in \mathbb{N}, \text{ such that } R_i(\pi) \supset R_i(\pi'), \text{ and}$$
$$\forall j > i : R_j(\pi) = R_j(\pi').$$

We denote $\succ_{erb}$ and $=_{erb}$ as usual: $\pi \succ_{erb} \pi'$ iff $\pi \succeq_{erb} \pi'$ and $\pi' \not\succeq_{erb} \pi$; $\pi =_{erb} \pi'$ iff $\pi \succeq_{erb} \pi'$ and $\pi' \succeq_{erb} \pi$.

Now, we define our concept of an ethical planning problem by extending STRIPS* tasks with a set of ethical rules and an ethical ranked base as follows:

Definition 5. *A **STRIPS*E** problem is a tuple $T = \langle F, s_0, s_*, A, R, erb\rangle$ where $\langle F, s_0, s_*, A\rangle$ is a STRIPS* problem, R is a finite set of ethical rules over a set E of ethical features, and erb is an ethical ranked base over R.*

Then, $\succeq_{erb}$ will model which plans are more ethically correct than others according to our framework:

Definition 6. *Let $T = \langle F, s_0, s_*, A, R, erb\rangle$ be a STRIPS*E problem and π a plan for T, π is **optimal** if and only if, for any other plan π', it holds that $\pi \succeq_{erb} \pi'$.*

Example 5. *(Hospital continued) Given the planning task T as defined before, we can extend it with the set of ethical rules $R = R_{con} \cup R_{deo} \cup R_{vir}$ over a set of ethical features $E = \{fast, paysFine, honesty, compassion, lying\}$, and an ethical ranked base erb where:*

$$erb(r_0) = \langle +, 1\rangle \quad erb(r_1) = \langle -, 1\rangle \quad erb(r_2) = \langle -, 4\rangle$$
$$erb(r_3) = \langle +, 2\rangle \quad erb(r_4) = \langle +, 3\rangle$$

In this scenario, we have that:

$$E_{\pi_1} = \emptyset$$
$$E_{\pi_2} = \{fast, paysFine, compassion, honesty\}$$
$$E_{\pi_3} = \{fast, lying, compassion\}$$

Then, $\pi_1 \succ_{erb} \pi_3$ and $\pi_2 \succ_{erb} \pi_3$ since $R_4(\pi_1) = R_4(\pi_2) = \{lying\} \supset R_4(\pi_3) = \emptyset$, and $\pi_2 \succ_{erb} \pi_1$ because $R_3(\pi_2) = \{compassion\} \supset R_3(\pi_1) = \emptyset$.

This makes sense according to our defined semantics as the plan that violates the highest ranked ethical rule concerning 'lying' (π_3) is weaker than those plans which don't (π_1, π_2). Similarly, plan π_2 is preferred to π_1, as it satisfies the next most important rule concerning 'compassion'.

Planning with ethical preferences

It is simple to show that the $\succeq_{erb}$ preference relation is reflexive and transitive, thus it is a preorder. However, certain plans may satisfy different elements at level $n \in \mathbb{N}$, so we can't say this order is total. We follow the work of (Feldmann, Brewka, and Wenzel 2006), who propose to use linearizations. A linearization of $\succeq_{erb}$ is a total preorder $\succeq^{lin}_{erb}$ that extends the first in such a way that $\succeq_{erb} \subseteq \succeq^{lin}_{erb}$, $=_{erb} \subseteq =^{lin}_{erb}$ and $\succ_{erb} \subseteq \succ^{lin}_{erb}$. This extension is useful as there is always a linearization for any preorder and it can be constructed by using a valuation function as follows:

Definition 7. *Given a STRIPS*E problem $T = \langle F, s_0, s_*, A, R, erb\rangle$ and a plan π of T, let $maxval_0 = 0$, then $\forall i \in \{1, \ldots, n\}$:*

$$val_i = maxval_{i-1} + 1$$
$$maxval_i = |\{r \in R : Rank(r){=}i\}| \times val_i + maxval_{i-1}$$
$$val(\pi) = \sum_{i \in \mathbb{N}} |R_i(\pi)| \times val_i$$

(Ross 1930) argues that while (*prima facie*) duties can conflict, no true dilemma can occur since one of these duties will always be the strongest, leading to a linearization of the duties in any context. We assume that the ranking of the duties is given, and we use (Feldmann, Brewka, and Wenzel 2006) as the linearization mechanism for $\succeq_{erb}$:

Proposition 1. *Let the preference relation between two plans $\succeq^{lin}_{erb}$ be defined as $\pi \succeq^{lin}_{erb} \pi'$ iff $val(\pi) \geq val(\pi')$, then it is indeed a linearization of $\succeq_{erb}$.*

Proof. Suppose that $\pi =_{erb} \pi'$, then trivially $val(\pi) = val(\pi')$ because $R_i(\pi) = R_i(\pi')$. In the case $\pi \succ_{erb} \pi'$, then $val(\pi) > val(\pi')$ by construction of val_i because there is an i such that $R_i(\pi) \supset R_i(\pi')$ and $\forall j > i : R_j(\pi) = R_j(\pi')$, and $|R_i(\pi)| \times val_i > \sum_{k=1}^{i} |R_k(\pi')| \times val_k$. Then, $\succeq_{erb} \subseteq \succeq^{lin}_{erb}$ follows from the two previous cases. Finally, because val assigns an integer to every plan, $\succeq^{lin}_{erb}$ is total. □

Example 6. *(Hospital continued) Following our running example, it holds that $val_1 = 1, val_2 = 3, val_3 = 6$ and $val_4 = 12$, then:*

$$val(\pi_1) = |R_1(\pi_1)| \times 1 + |R_4(\pi_1)| \times 12 = 13.$$
$$val(\pi_2) = |R_1(\pi_2)| \times 1 + |R_3(\pi_2)| \times 6 + |R_4(\pi_2)| \times 12 = 19.$$
$$val(\pi_3) = |R_1(\pi_3)| \times 1 + |R_3(\pi_1)| \times 6 = 8.$$

In order to find an optimal plan, we will show that any STRIPS*E can be transformed into an equivalent STRIPS*U problem. In what follows, we make the following assumptions for STRIPS*E problems:

- Ethical rules with activation condition 'final' refer to different fluents i.e. $\nexists r, r' \in R$ s.t. $Act(r) = Act(r') =$ final and $Pre(r) = Pre(r')$,
- Ethical features are *unique* i.e. $\nexists r, r' \in R$ s.t. $Id(r) = Id(r')$, and *distinct* from the fluents i.e. $\nexists r \in R$ s.t. $Id(r) \in F$.

Proposition 2. *Given a STRIPS*E problem $T = \langle F, s_0, s_*, A, R, erb\rangle$, two plans π, π' of T, and let $T' = \langle F', s_0, s_*, A', u\rangle$ be a STRIPS*U problem where:*

- $F' = F \cup \{Id(r) : r \in R \wedge Act(r) \in A\}$,
- $A' = \{\langle Pre(a), Eff(a) \cup \{\langle Pre(r), Id(r)\rangle : Act(r) = a\}\rangle : a \in A\}$, *and*
- *The utilities* u *are defined as follows:*
  1. $u(Id(r)) = val_{Rank(r)}$ *if* $\exists r \in R$ *s.t.* $Type(r) = +$ *and* $Act(r) \in A$,
  2. $u(\neg Id(r)) = val_{Rank(r)}$ *if* $\exists r \in R$ *s.t.* $Type(r) = -$ *and* $Act(r) \in A$,
  3. $u(f_1 \wedge \ldots \wedge f_n) = val_{Rank(r)}$ *if* $\exists r \in R$ *s.t.* $Type(r) = +$, $Act(r) = \text{final}$ *and* $Pre(r) = \{f_1, \ldots, f_n\}$,
  4. $u(\neg(f_1 \wedge \ldots \wedge f_n)) = val_{Rank(r)}$ *if* $\exists r \in R$ *s.t.* $Type(r) = -$, $Act(r) = \text{final}$ *and* $Pre(r) = \{f_1, \ldots, f_n\}$, *and*
  5. $u(\phi)$ *is undefined otherwise.*

Then, $\pi \succeq_{erb} \pi'$ *w.r.t.* T *iff* $u(\pi) \geq u(\pi')$ *w.r.t.* T'.

Proof. It is trivial to see that any plan π of T is a plan in T' and vice versa, due to the fact that the goal conditions are left unchanged, the initial state is the same, the fluents of T are included in those of T', preconditions of actions don't change, and the changes in the effects of actions only deal with the new fluents in F' (as they are *distinct* from the fluents in F by assumption). Next, let π be a plan of T:

a. For any rule $r \in R$ such that $Act(r) \in A$, it holds that $Id(r) \in E_\pi$ if and only if $Id(r) \in Succ_{T'}(\pi, s_0)$ because $Id(r)$ is only added to a state s according to A' iff action $Act(r)$ is executed while $Pre(r) \subseteq s$.

b. Given any rule $r \in R$ such that $Act(r) = \text{final}$, it holds that $Pre(r) \subseteq Succ_T(\pi, s_0)$ if and only if $Succ_{T'}(\pi, s'_0) \models Pre(r)$ (treating the set $Pre(r)$ as a conjunction) as none of the original fluents are affected by the transformation (as they are *distinct* from the new).

Furthermore, let $s = Succ_{T'}(\pi, s'_0)$ be the final state of plan π, for every ethical rule $r \in R$ such that $Act(r) \in A$, due to property (a) it holds that $s \models Id(r)$ iff $Id(r) \in E_\pi$. Also, due to property (b) for every ethical rule $r \in R$ such that $Act(r) = \text{final}$, it holds that $s \models Pre(r)$ iff $Id(r) \in E_\pi$. Thus, let $C_1 = \{r \in R : Type(r) = + \wedge Act(r) \in A \wedge s \models Id(r)\}$, $C_2 = \{r \in R : Type(r) = - \wedge Act(r) \in A \wedge s \not\models Id(r)\}$, $C_3 = \{r \in R : Type(r) = + \wedge Act(r) = \text{final} \wedge s \models Pre(r)\}$, $C_4 = \{r \in R : Type(r) = - \wedge Act(r) = \text{final} \wedge s \not\models Pre(r)\}$, due to the definition of utility (1-5) of a plan, $u(\pi) = u(s) = \sum_{r \in C_1 \cup C_2 \cup C_3 \cup C_4} val_{Rank(r)} = \sum_{\{r \in R : Id(r) \in E_\pi \iff Type(r) = +\}} val_{Rank(r)} = val(\pi)$.

Finally, let π' be another plan s.t. $\pi \succeq_{erb} \pi'$, then $u(\pi) = val(\pi)$ and $u(\pi') = val(\pi')$, thus $u(\pi) \geq u(\pi')$ as val induces a linearization over $\succeq_{erb}$ from Proposition 1. □

In terms of computational costs, each ethical rule r with $Act(r) \in A$ will add a fluent to the planning domain and a conditional effect to its corresponding action. Additionally, each ethical rule will induce utilities to be checked by the planner.

Implementation

One of the main benefits of our approach is that by transforming the STRIPS*E problem into another with only soft goals and utilities, it is possible to apply PDDL planners designed for this purpose. This is why, in order to exemplify our framework we have implemented (i) an extension of STRIPS* in the syntax of the language PDDL3.0 that models ethical rules and our qualitative preference model, and (ii) a translation routine from domains with ethical rules into an equivalent one with utilities applying Proposition 2.

We chose the IPC planning definition language PDDL3.0 (Gerevini et al. 2009), which models STRIPS domains, conditional effects, utilities and other extensions. We tested our framework using the planner in (Coles and Coles 2011). The definition of ethical rules should be included in the domain file of a PDDL representation:

```
<eth-rule> ::= (:ethical-rule <name>
        :type <eth-type>
        :precondition φ
        :activation <eth-actv>
        :rank <positive integer>)
<eth-type> ::= + | -
<eth-actv> ::= <action-name> | final
```

Where ϕ is an atomic formulae over the (grounded) predicates of the domain with no comparison or numeric terms.

Our translation routine then parses a domain and problem file and generates a pair of updated files. This routine has been implemented using Python and is publicly available[1].

Example 7. *(Hospital continued) Back to our ongoing example, we represent the ethical rules* r_0 *and* r_2 *as:*

```
(:ethical-rule fast
    :type + :precondition (tookHighway)
    :activation final :rank 1)
(:ethical-rule lied
    :type - :precondition (atToll)
    :activation presentB :rank 4)
```

Then, our translation routine will update the domain and problem definition files with a new fluent 'lied' and change the effects of action 'presentB' (see Proposition 2):

```
(:action presentB
    :parameters ()
    :precondition  (atToll)
    :effect (and (<original effect>)
        (when (atToll) (lied))))
```

And add the following preferences:

```
(:goal (and (<original goal>)
    (preference p_fast (tookHighway))
    (preference p_lied (not (lied)))))
(:metric minimize (+
    (* (is-violated p_fast) 1)
    (* (is-violated p_lied) 4)))
```

Notice that instead of maximizing the utility like in the previous section, PDDL3.0 preferences specify utilities using the operator 'is-violated', which forces us to (equivalently) invert the problem to minimize ethical rule violations.

[1] https://github.com/martinjedwabny/pddl-ethical

Related work

Regarding normative ethics in the context of AI, (Cointe, Bonnet, and Boissier 2016) introduce a framework for ethical reasoning and planning based on BDI agents, that presents a construct called moral rule, which is related to our ethical rules, but used in a different fashion. (Berreby, Bourgne, and Ganascia 2017) present a modular framework that implements several ethical theories in answer set programming, as well as different mechanisms to combine them. Deontic logics has been applied to produce ethics-aware systems based on the event calculus (Govindarajulu and Bringsjord 2017; Hashmi, Governatori, and Wynn 2014; Marín and Sartor 1999). An extension for (transition system defining) language C+ is presented in (Sergot 2004), in which fluents and actions (under specified circumstances) can be forbidden, states and transitions are thus labeled permitted or not, according to whether any of those rules are broken. In (Panagiotidi and Vázquez-Salceda 2011), restrictions are characterized using context-dependent norms, and applied to STRIPS-based planning domains. However, these approaches were not designed for PDDL planners, which prevent them from being more practical in real use cases.

Discussion

We have shown a flexible framework that allows automated planners to represent the three main normative ethical theories. Qualitative preferences between the ethical features are represented using ranks as in (Feldmann, Brewka, and Wenzel 2006). Our model has the advantage that it can be translated into utility-based preferences. One possible way of extending this work would be by providing a more general language to express preferences as described in (Brewka 2004). A qualitative framework is imperative in ethical domains e.g. consider a problem in which an agent is assigned a unit of utility for executing a trivial action, such as giving away ice creams, while on the other hand is assigned a thousand units, or any other fixed utility to not killing a person. Summing up simple utilities would effectively compensate killing a person if the agent gives away enough ice creams. In contrast, the rank-based approach prohibits such kind of behaviour as we showed earlier. In addition, a rank-based approach makes it easier to elicit priorities amongst ethical rules from external sources, as one doesn't need to specify the utility for every combination of ethical rules, but for each ethical rule separately.

Finally, we have shown how our framework profits from PDDL planners, which opens up interesting avenues of research for ethical planning for real world use-cases. For future work, an in-depth analysis of the computational costs induced by our norms is necessary.

References

Berreby, F.; Bourgne, G.; and Ganascia, J.-G. 2017. A declarative modular framework for representing and applying ethical principles. In *IFAAMAS 2017*.

Bonnemains, V.; Saurel, C.; and Tessier, C. 2016. How Ethical Frameworks Answer to Ethical Dilemmas: Towards a Formal Model. In *EDIA@ ECAI 2016*, 44–51.

Brewka, G. 2004. A rank based description language for qualitative preferences. In *ECAI 2004*, volume 16, 303.

Brundage, M. 2014. Limitations and risks of machine ethics. *JEAIL* 26(3):355–372.

Cointe, N.; Bonnet, G.; and Boissier, O. 2016. Ethical judgment of agents' behaviors in multi-agent systems. In *AAMAS 2016*, 1106–1114.

Coles, A., and Coles, A. 2011. Lprpg-p: Relaxed plan heuristics for planning with preferences. In *ICAPS 2011*.

Dennis, L., and Fisher, M. 2018. Practical challenges in explicit ethical machine reasoning. *arXiv preprint arXiv:1801.01422*.

Feldmann, R.; Brewka, G.; and Wenzel, S. 2006. Planning with prioritized goals. In *KR*, 503–514.

Fikes, R. E., and Nilsson, N. J. 1971. Strips: A new approach to the application of theorem proving to problem solving. *AIJ* 2(3-4):189–208.

Gerevini, A. E.; Haslum, P.; Long, D.; Saetti, A.; and Dimopoulos, Y. 2009. Deterministic planning in the fifth international planning competition: Pddl3 and experimental evaluation of the planners. *AIJ* 173(5-6):619–668.

Govindarajulu, N. S., and Bringsjord, S. 2017. On automating the doctrine of double effect. *arXiv preprint arXiv:1703.08922*.

Haines, W. 2006. Consequentialism. In *Internet Encyclopedia of Philosophy*.

Hashmi, M.; Governatori, G.; and Wynn, M. T. 2014. Modeling obligations with event-calculus. In *RuleML 2014*, 296–310. Springer.

Lindner, F.; Bentzen, M. M.; and Nebel, B. 2017. The HERA approach to morally competent robots. In *IROS 2017*, 6991–6997. IEEE.

Lindner, F.; Mattmüller, R.; and Nebel, B. 2019. Moral permissibility of action plans. In *AAAI 2019*, volume 33, 7635–7642.

Marín, R. H., and Sartor, G. 1999. Time and norms: a formalisation in the event-calculus. In *ICAIL 1999*, 90–99.

Panagiotidi, S., and Vázquez-Salceda, J. 2011. Towards practical normative agents: a framework and an implementation for norm-aware planning. In *COIN 2011*, 93–109. Springer.

Pednault, E. P. 1989. ADL: Exploring the Middle Ground Between STRIPS and the Situation Calculus. *KR 1989* 89:324–332.

Ross, W. 1930. *The Right and the Good*. Oxford University Press.

Sergot, M. 2004. An action language for modelling norms and institutions. *Technical Report 2004/8*. Publisher: Imperial College London.

Singer, M. G. 1977. Actual consequence utilitarianism. *Mind* 86(341):67–77.

Yu, H.; Shen, Z.; Miao, C.; Leung, C.; Lesser, V. R.; and Yang, Q. 2018. Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*.