Logic-Based Ethical Planning

Umberto Grandi, Emiliano Lorini, Timothy Parker, Rachid Alami

Logic-Based Ethical Planning

Umberto Grandi[1], Emiliano Lorini[1], Timothy Parker[1], and Rachid Alami[2]

[1] IRIT, CNRS, Toulouse University, France
[2] LAAS, CNRS, Toulouse, France

Abstract. In this paper we propose a framework for ethical decision-making in the context of planning, with intended application to robotics. We put forward a compact but highly expressive language for ethical planning that combines linear temporal logic with lexicographic preference modelling. This original combination allows us to assess plans both with respect to an agent's values and its desires, introducing the novel concept of the morality level of an agent and moving towards multi-goal, multi-value planning. We initiate the study of computational complexity of planning tasks in our setting, and we discuss potential applications to robotics.

1 Introduction

In ethical planning the planning agent has to find a plan for promoting a certain number of ethical values. The latter include both abstract values such as justice, fairness, reciprocity, equity, respect for human integrity and more concrete ones such as "greenhouse gas emissions are reduced". Unlike classical planning in which the goal to be achieved is unique, in ethical planning the agent can have multiple and possibly conflicting values, that is, values that cannot be concomitantly satisfied. It is typical of ethical planning the problem of facing a moral struggle which is "...provoked by inconsistencies between value commitments and information concerning the kinds of decision problems which arise..." [18, p. 8]. Consequently, in ethical planning the agent needs to evaluate and compare the ideality (or goodness) of different plans depending on how many and which values are promoted by each of them.

In this paper our intended application field is that of robotics. Including ethical considerations in robotics planning requires (at least) three steps. First, identify ethically sensitive situations in the robotics realm, and how are these situations represented. Planning seems to be the first candidate in which to include ethical considerations, thus we assume that values or ethical judgments are expressed about the results of plans. Second, design a language to express such values, bearing in mind that they can be, and often are, potentially conflicting in multiple ways: among values, between a value and a goal, or between a value and good practices. Such a value representation language needs to be compact and computationally tractable. Third, complete the picture of ethical planning by designing algorithms that compare plans based on the ethical values.

In this paper we put forward a framework for ethical planning based on a simple temporal logic language to express both an agent's values and goals. For ease of exposition we focus on single-agent planning with deterministic sequential actions in a known environment. Our model borrows from the existing literature on planning and combines it in an original way with research in compact representation languages for preferences. The latter is a widely studied topic in knowledge representation, where logical and graphical languages are proposed to represent compactly the preferences of an agent over a combinatorial space of alternatives, often described by means of variables. In particular, we commit to a prioritised or lexicographic approach to solve the possible arising inconsistencies among goals, desires, and good practices in a unified planning model.

2 Related Work

There is considerable research in the field of ethics and AI, see Müller [25] for a general overview. Popular ethical theories for application are consequentialism, deontology, and virtue ethics.[3] Our approach should be able to work with any notion of "good actions" but is probably a most natural fit for pluralistic consequentialism [30].

While there is a lot of work at the theoretical/abstract level, there is comparatively less that examines how ethical reasoning in artificial agents could actually be done in practice. There are approaches both in terms of formal models [12] and allowing agents to learn ethical values [2]. Yu et al. [33] provides a recent survey of this research area. The closest approaches to ours are the recent work on (*i*) logics for ethical reasoning and (*ii*) the combination of a compact representation language, such as conditional preference networks, with decision-making in an ethically sensitive domain. The former are based on different methodologies including event calculus (ASP) [6], epistemic logic and preference logic [22, 24], BDI (belief, desire, intention) agent language [11], classical higher-order logic (HOL) [5]. The latter was presented in "blue sky" papers [21, 28] complemented with a technical study of distances between CP-nets [20] and, more recently, with an empirical study on human ethical decision-making [4]. CP-nets are a compact formalism to order states of the world described by variables.

We take inspiration from these lines of work, but depart from them under two aspects. First, robotics applications are dynamic ones, and ethical principles must be expressed over time. Hence, unlike existing logics for ethical reasoning, our focus is on a specification language for values based on linear temporal logic. Second, ethical decision-making in robotic applications requires mixing potentially conflicting values with desires of the agent and to express the notion of plan, and CP-nets alone are not sufficient.

In the field of robotics, there are approaches to enabling artificial agents to compute ethical plans. The evaluative component, which consists in assessing the "goodness" of an action or a plan in relation to the robot's values, is made explicit

[3] See Copp [9] for a philosophical introduction, and Jenkins et al. [15], Powers [27], and Vallor [31] for a discussion of these three theories in robotics.

by Arkin et al. [3] and Vanderelst and Winfield [32]. Evans et al. [13] focuses on a collision scenario involving an autonomous vehicle, proposing to prioritise the ethical claims depending on the situation, e.g. by giving more priorities to the claims of the more endangered agents. Related work explores the design of planning algorithms designed to help robots produce socially acceptable plans by assigning weights to social rules [1] .

In preference-based planning by Bienvenu et al. [7] plans are compared relative to a *single* (possibly lexicographic) preference formula about temporal properties. Similarly, Lindner et al. [19] evaluate the permissibility of plans according to a *specific* ethical principle such as the deontological principle, the utilitarian principle, the do-no-harm or the double effect principle. In our approach plans are compared relative to *sets* of values. Comparison of alternatives (e.g., plans, states, histories) relative to a set of values is an essential aspect of ethics which is not considered in these two works. As we will show in Section 3.5, it opens up the possibility of formalizing the notion of moral conflict.

3 Model

In this section, we present the formal model of ethical evaluation and planning which consist, respectively, in comparing the goodness of plans and in finding the best plan relative to a given base of ethical values.

3.1 LTL Language

Let $Prop$ be a countable set of atomic propositions and let Act be a finite non-empty set of action names. Elements of $Prop$ are noted $p, q, \ldots$, while elements of Act are noted $a, b, \ldots$. We assume the existence of a special action skip. The set of states is $S = 2^{Prop}$ with elements $s, s', \ldots$

In order to represent the agent's values, we introduce the language of LTL_f (Linear Temporal Logic over Finite Traces) [26, 10], noted $\mathcal{L}_{\mathsf{LTL}_f}(Prop)$ (or $\mathcal{L}_{\mathsf{LTL}_f}$), defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathsf{X}\varphi \mid \varphi_1 \,\mathsf{U}\, \varphi_2,$$

with p ranging over $Prop$. X and U are the operators "next" and "until" of LTL_f. Operators "henceforth" (G) and "eventually" (F) are defined in the usual way: $\mathsf{G}\varphi \stackrel{\text{def}}{=} \neg(\top \,\mathsf{U}\, \varphi)$ and $\mathsf{F}\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi$. The propositional logic fragment of $\mathcal{L}_{\mathsf{LTL}_f}$ is noted $\mathcal{L}_{\mathsf{PL}}$ and is defined in the usual way. We will use $\mathcal{L}_{\mathsf{PL}}$ to describe the effect preconditions of the agent's actions.

3.2 Histories

The notion of history is needed for interpreting formulas in $\mathcal{L}_{\mathsf{LTL}_f}$. We define a k-history to be a pair $H = (H_{st}, H_{act})$ with

$$H_{st} : [0, k] \longrightarrow S \text{ and } H_{act} : [1, k] \longrightarrow Act.$$

A history specifies the actual configuration of the environment at a certain time point and the action executed by the agent that leads to the next state. The set of k-histories is noted $Hist_k$. The set of histories is $Hist = \bigcup_{k\in\mathbb{N}} Hist_k$. Semantic interpretation of formulas in $\mathcal{L}_{\mathsf{LTL}_f}$ relative to a k-history $H \in Hist$ and a time point $t \in [0,k]$ goes as follows (we omit boolean cases which are defined as usual):

$$\begin{aligned}
H,t \models p \quad &\Longleftrightarrow p \in H_{st}(t),\\
H,t \models \mathsf{X}\varphi \quad &\Longleftrightarrow t < k \text{ AND } H,t+1 \models \varphi,\\
H,t \models \varphi_1 \,\mathsf{U}\, \varphi_2 \quad &\Longleftrightarrow \exists t' \geq t : t' \leq k \text{ AND } H,t' \models \varphi_2 \text{ AND}\\
&\qquad \forall t'' \geq t : \text{IF } t'' < t' \text{ THEN } H,t'' \models \varphi_1.
\end{aligned}$$

3.3 Action Theory

We suppose actions in Act are described by an action theory $\gamma = (\gamma^+, \gamma^-)$, where γ^+ and γ^- are, respectively, the positive and negative effect precondition function $\gamma^+ : Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}$ and $\gamma^- : Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}$.

The fact $\gamma^+(a,p)$ guarantees that proposition p will be *true* in the next state when action a is executed, while $\gamma^-(a,p)$ guarantees that proposition p will be *false* in the next state when action a is executed. We stipulate that if $\gamma^+(a,p)$ and $\gamma^-(a,p)$ are concomitantly true at a given state and action a is executed, then the truth value of p will not change in the next state. The latter captures an inertial principle for fluents.

Definition 1 (Action-compatible histories). *Let $\gamma = (\gamma^+, \gamma^-)$ be an action theory and let $H = (H_{st}, H_{act})$ be a k-history. We say H is compatible with γ if the following condition holds, for every $t \in [1,k]$ and for every $a \in Act$:*

$$\begin{aligned}
&\text{IF } \; H_{act}(t) = a \text{ THEN}\\
&\quad H_{st}(t) = \Big(H_{st}(t-1) \setminus \big\{p \in Prop : H,t-1 \models \neg\gamma^+(a,p) \wedge\\
&\qquad \gamma^-(a,p)\big\}\Big) \cup \big\{p \in Prop : H,t-1 \models \gamma^+(a,p) \wedge\\
&\qquad \neg\gamma^-(a,p)\big\}.
\end{aligned}$$

The set of γ-compatible histories is noted $Hist(\gamma)$.

3.4 Plans

Let us now move from the notion of action to the notion of plan. Given $k \in \mathbb{N}$, a k-plan is a function

$$\pi : \{1,\ldots,k\} \longrightarrow Act.$$

The set of k-plans is noted $Plan_k$. The set of plans is $Plan = \bigcup_{k\in\mathbb{N}} Plan_k$. The following definition introduces the notion of history generated by a k-plan π at an initial state s_0. It is the action-compatible k-history along which the agent executes the plan π starting at state s_0.

Definition 2 (History generated by a k-plan). *Let $\gamma = (\gamma^+, \gamma^-)$ be an action theory, $s_0 \in S$ and $\pi \in Plan_k$. Then, the history generated by plan π from state s_0 in conformity with the action theory γ is the k-history $H^{\pi,s_0,\gamma} = (H^{\pi,s_0,\gamma}_{st}, H^{\pi,s_0,\gamma}_{act})$ such that:*

$$(i)\ H^{\pi,s_0,\gamma} \in Hist(\gamma),$$
$$(ii)\ H^{\pi,s_0,\gamma}_{st}(0) = s_0,$$
$$(iii)\ \forall k'\ s.t.\ 1 \leq k' \leq k : H^{\pi,s_0,\gamma}_{act}(k') = \pi(k'),$$

Given a set of LTL_f-formulas Σ, we define $Sat(\Sigma,\pi,s_0,\gamma)$ to be the set of formulas from Σ that are guaranteed to be true by the execution of plan π at state s_0 under the action theory γ. That is,

$$Sat(\Sigma,\pi,s_0,\gamma) = \big\{\varphi \in \Sigma : H^{\pi,s_0,\gamma}, 0 \models \varphi\big\}.$$

3.5 Moral Conflicts

An ethical planning agent is likely to have multiple values that it wishes to satisfy when making plans. Some of these values will be ethical in nature ("do not harm humans"), and some may not be ("do not leave doors open"). However, the more values the robot has the more likely it is to experience scenarios where it cannot satisfy all of its values with any given plan, and must violate some of them. In such a scenario, the agent must first work out which subsets of its value base are jointly satisfiable, and then which of those subsets it should choose to satisfy.

To this end we define a notion of a moral conflict (note that in line with Levi [18] we refer to any conflict between an agent's values as a "moral conflict" even if some or all of those values are not strictly moral/ethical in nature).

Definition 3 (Moral problem). *A moral problem is a tuple $M = (\Omega, \gamma, s_0)$ where:*

- *$\Omega \subseteq \mathcal{L}_{\mathsf{LTL}_f}$ is a set of values (which may or may not be strictly moral in nature);*
- *$\gamma = (\gamma^+, \gamma^-)$ is an action theory and s_0 is an initial state, as described above.*

Definition 4 (Moral conflict). *A moral problem $M = (\Omega, \gamma, s_0)$ is a moral conflict if:*

- *$\forall k \in \mathbb{N}$, there is no k-plan π such that $Sat(\Omega,\pi,s_0,\gamma) = \Omega$.*

In other words, a moral conflict occurs when it is not possible to satisfy all of our values with any given plan. In some cases, a moral conlict may not depend on any particular feature of the start state, but may result simply from the value base and action theory, or even the value base alone. This allows us to define two further notions of moral problem.

Definition 5 (Physical moral problem). *A physical moral problem is a pair* (Ω, γ) *where:*

- $\Omega \subseteq \mathcal{L}_{\mathsf{LTL}_f}$ *is a set of values;*
- γ *is an action theory.*

Definition 6 (Logical moral problem). *A logical moral problem is a set of values* $\Omega \subseteq \mathcal{L}_{\mathsf{LTL}_f}$.

We can also define moral conflict for these moral problems. A physical (logical) moral problem is a physical (logical) value conflict if for every possible start state s_0 (and every possible action theory γ), the resultant moral value problem $M = (\Omega, \gamma, s_0)$ is a moral conflict. By our definition, conflict mirrors the concept of necessity. Necessity would imply that *every* possible plan satisfies all the values in Ω, whereas conflict implies that *no* plan satisfies all values. Thus it is interesting to note that our definitions of conflict have mirrors in philosophical literature [16]. A physical moral conflict mirrors the notion of nomic necessity (necessary given the laws of nature) (at least from the perspective of the robot, for whom the action theory comprises the laws of nature) whereas a logical moral conflict mirrors the notion of logical necessity (necessary given the nature of logic).

If an agent is experiencing a moral conflict, one response would be to "temporarily forget" values until it has a satisfiable set.

Definition 7 (Contraction). *If* $M = (\Omega, \gamma, s_0)$ *is a moral problem and* $M' = (\Omega', \gamma, s_0)$ *is a moral problem, we say that* M' *is a contraction of* M *if:*

- $\Omega' \subseteq \Omega$
- M' *is not a moral conflict.*

Note that if $M = (\Omega, \gamma, s_0)$ is a moral problem, π is a plan, and $\Omega' = Sat(\Omega,\pi,s_0,\gamma)$ then $M' = (\Omega', \gamma, s_0)$ must be a contraction of M.

In this case, we refer to M' as the contraction generated by π. This also illustrates that the current notion of contraction is unhelpful for an agent attempting to select a plan in a moral conflict, as all plans generate contractions. What would be helpful is some notion of a "minimal" or "ideal" contraction that sacrifices as few values as possible.

Definition 8 (Minimal contraction). *If* $M = (\Omega, \gamma, s_0)$ *is a moral problem and* $M' = (\Omega', \gamma, s_0)$ *is a contraction of* M*,* M *is:*

- *A qual-minimal contraction if there is no contraction* $M'' = (\Omega'', \gamma, s_0)$ *such that* $\Omega' \subset \Omega''$*;*
- *A quant-minimal contraction if there is no contraction* M'' *such that* $|\Omega'| < |\Omega''|$

Proposition 1. *If* $M = (\Omega, \gamma, s_0)$ *is a moral problem and is not a moral conflict, then the only qual-minimal and quant-minimal contraction of* M *is* M.

For either notion of minimality, we will have cases where there are multiple minimal contractions of a given moral conflict. This can produce unintuitive results, as if there is some moral conflict with $\Omega = \{$"do not kill humans","do not leave the door open"$\}$ with contractions $\{$"do not kill humans"$\}$ and $\{$"do not leave the door open"$\}$ then either notion of minimality will tell you that both contractions are ideal. On the other hand, it does seem that any stronger notion of minimality should at least respect qualitative minimality, since (intuitively), if plan π_1 fulfills all of the values fulfilled by π_2, and fulfills more values, then π_1 should be preferred to π_2.

Proposition 2. *Given a moral conflict M, a contraction M' is quant-minimal only if it is qual-minimal.*

One way to resolve this is to recognise, in line with Levi [18], that some of our values are only used as tiebreakers to separate otherwise-equivalent plans, and should not be considered directly alongside our more important values. To model this, our values exist in lexicographically ordered sets, where each set is examined only if the sets above cannot deliver a verdict.

3.6 Lexicographic Value Base

Together with an action theory and an initial state, an agent's value base constitutes an ethical planning domain.

Definition 9 (Ethical planning domain). *An ethical planning domain is a tuple $\Delta = (\gamma, s_0, \overline{\Omega})$ where:*

- *$\gamma = (\gamma^+, \gamma^-)$ is an action theory and s_0 is an initial state, as specified above;*
- *$\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$ is the agent's value base with $\Omega_k \subseteq \mathcal{L}_{\mathsf{LTL}_f}$ for every $1 \leq k \leq m$.*

Ω_1 is the agent's set of values with priority 1, Ω_2 is the agent's set of values with priority 2, and so on. For notational convenience, given a value base $\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$, we note $dg(\overline{\Omega})$ its degree (or arity).

Agent's values are used to compute the *relative ideality* of plans, namely, whether a plan π_2 is at least as ideal as another plan π_1. Following [24], we call *evaluation* the operation of computing an ideality ordering over plans from a value base. Building on classical preference representation languages [17], we define the following qualitative criterion of evaluation, noted $\preceq_\Delta^{qual}$, which compares two plans lexicographically on the basis of inclusion between sets of values.

Definition 10 (Qualitative ordering of plans). *Let $\Delta = (\gamma, s_0, \overline{\Omega})$ be an ethical planning domain with $\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$ and $\pi_1, \pi_2 \in Plan$. Then, $\pi_1 \preceq_\Delta^{qual} \pi_2$ if and only if:*

$$
\begin{aligned}
&(i)\ \exists 1 \leq k \leq m \text{ s.t. } Sat(\Omega_k,\pi_1,s_0,\gamma) \subset Sat(\Omega_k,\pi_2,s_0,\gamma), \text{ and} \\
&\quad\ \forall 1 \leq k' < k, Sat(\Omega_{k'},\pi_1,s_0,\gamma) = Sat(\Omega_{k'},\pi_2,s_0,\gamma); \text{ or} \\
&(ii)\ \forall 1 \leq k \leq m, Sat(\Omega_k,\pi_1,s_0,\gamma) = Sat(\Omega_k,\pi_2,s_0,\gamma).
\end{aligned}
$$

Note that a quantitative criterion could also be defined by counting the number of satisfied values in each level and, in line with the previous definition, compare these values lexicographically.

The quantitative criterion, noted $\preceq_\Delta^{quant}$, compares two plans lexicographically on the basis of comparative cardinality between sets of values.

Definition 11 (Quantitative ordering of plans). *Let $\Delta = (\gamma, s_0, \overline{\Omega})$ be an ethical planning domain with $\overline{\Omega} = (\Omega_1, \ldots, \Omega_m)$ and $\pi_1, \pi_2 \in Plan$. Then, $\pi_1 \preceq_\Delta^{quant} \pi_2$ if and only if:*

$$\begin{aligned}
&(i)\ \exists 1 \leq k \leq m \text{ s.t. } |Sat(\Omega_k,\pi_1,s_0,\gamma)| < |Sat(\Omega_k,\pi_2,s_0,\gamma)|, \text{ and}\\
&\quad \forall 1 \leq k' < k, |Sat(\Omega_{k'},\pi_1,s_0,\gamma)| = |Sat(\Omega_{k'},\pi_2,s_0,\gamma)|; \text{ or}\\
&(ii)\ \forall 1 \leq k \leq m, |Sat(\Omega_k,\pi_1,s_0,\gamma)| = |Sat(\Omega_k,\pi_2,s_0,\gamma)|.
\end{aligned}$$

This allows us to define another notion of minimal contraction for a moral conflict, namely a minimal contraction with respect to a lexicographic value base.

Definition 12 (Lexicographic-minimal contraction). *If $M = (\Omega, \gamma, s_0)$ is a moral problem, and $\overline{\Omega} = (\Omega_1, ..., \Omega_m)$ is a value base such that $\cup\overline{\Omega} = \Omega$ then $M' = (\Omega', \gamma, s_0)$ is a $\overline{\Omega}$-qual-minimal contraction of M if and only if:*

$$\begin{aligned}
&(i)\ \Omega' \subseteq \Omega;\\
&(ii)\ M' \text{ is not a moral conflict};\\
&(iii)\ \text{If } M'' = (\Omega'', \gamma, s_0) \text{ is also a contraction of } M,\\
&\quad \nexists k : (a)\ 1 \leq k \leq m \text{ and } \Omega' \cap \Omega_k \subset \Omega'' \cap \Omega_k, \text{ and}\\
&\quad\quad (b)\ \forall 1 \leq i < k, \Omega' \cap \Omega_i = \Omega'' \cap \Omega_i.
\end{aligned}$$

Note that by combining definitions 11 and 12 we can define a notion of $\overline{\Omega}$-quant-minimal contraction.

Proposition 3. *Given a moral conflict M, a contraction M' is $\overline{\Omega}$-qual-minimal or $\overline{\Omega}$-quant-minimal only if it is qual-minimal.*

3.7 Adding Desires

The behavior of autonomous ethical agents is driven not only by ethical values aimed at promoting the good for society but also by their endogenous motivations, also called *desires* or *goals*. Following existing theories of ethical preferences in philosophy, economics and logic [29, 14, 23], we assume that (i) desires and values are competing motivational attitudes, and (ii) the agent's degree of morality is a function of its disposition to promote the fulfilment of its values at the expense of the satisfaction of its desires. The following definition extends the notion of ethical planning domain by the notions of desire and introduces the novel concept of degree of morality.

Definition 13 (Mixed-motive planning domain). *A mixed-motive planning domain is a tuple $\Gamma = (\gamma, s_0, \overline{\Omega}, \Omega_D, \mu)$ where*

- $(\gamma, s_0, \overline{\Omega})$ *is an ethical planning domain (Definition 9);*
- $\Omega_D \subseteq \mathcal{L}_{\mathsf{LTL}_f}$ *is the agent's set of desires or goals;*
- $\mu \in \{1, \ldots, dg(\overline{\Omega}) + 1\}$ *is the agent's degree of morality.*

A mixed-motive planning domain induces an ethical planning domain whereby the agent's set of desires is treated as a set of values whose priority level depends on the agent's degree of morality. Specifically, the lower the agent's degree of morality, the higher the priority of the agent's set of desires in the induced ethical planning domain. In many practical applications it is likely to be desirable to restrict the range of values that μ can take, in order to prevent (for example) the robot's goal from overriding its safety values.

Definition 14 (Induced ethical planning domain). *Let $\Gamma = (\gamma, s_0, \overline{\Omega}, \Omega_D, \mu)$ be a mixed-motive planning domain. The ethical planning domain induced by Γ is the tuple $\Delta = (\gamma, s_0, \overline{\Omega}')$ such that $dg(\overline{\Omega}') = dg(\overline{\Omega}) + 1$ with:*

$$(i)\ \Omega'_\mu = \Omega_D;$$
$$(ii)\Omega'_k = \Omega_k \text{ for } 1 \leq k < \mu;$$
$$(iii)\Omega'_k = \Omega_{k-1} \text{ for } \mu < k \leq dg(\overline{\Omega}) + 1.$$

4 An Example

Consider a blood delivery robot in a hospital. The robot mostly makes deliveries between different storage areas, and sometimes delivers blood to surgeries. The robot may have to deal with various kinds of obstacles to complete its deliveries, but we will consider only one: people blocking the robot. The robot has two methods to resolve this obstacle, it can ask for them to move and then wait for them to move (ask), or it can use a loud air-horn to "force" them to move (horn). Once the person has moved, the robot can reach its destination (move). We suppose that the robot can tell some things about its environment, it knows if it is blocked (blocked), if it is near the operating theatre (theatre) and if it has reached its destination (destination). We can then define the action model as follows:

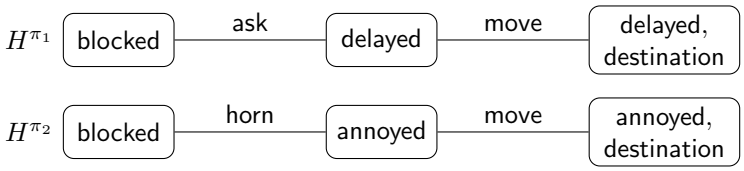
$$\begin{aligned}
\gamma^+(\mathsf{move}, \mathsf{destination}) &= \neg\mathsf{blocked} \\
\gamma^-(\mathsf{ask}, \mathsf{blocked}) &= \mathsf{blocked} \\
\gamma^+(\mathsf{ask}, \mathsf{delayed}) &= \top \\
\gamma^-(\mathsf{horn}, \mathsf{blocked}) &= \mathsf{blocked} \\
\gamma^+(\mathsf{horn}, \mathsf{annoyed}) &= \top \\
\gamma^+(\mathsf{horn}, \mathsf{dangerous}) &= \mathsf{theatre} \\
\text{otherwise}, \gamma^\pm(a, p) &= \bot
\end{aligned}$$

The propositions delayed, annoyed and dangerous are used to keep track of the robot's actions, we suppose that using the horn near the operating theatre is dangerous. The values and desires of the robot can be presented as follows:

$$\begin{aligned}\overline{\Omega} &= \{\Omega_1, \Omega_2\}\\ \Omega_1 &= \{\mathsf{G}\neg\mathsf{dangerous}\}\\ \Omega_2 &= \{\mathsf{G}\neg\mathsf{annoyed}\}\\ \Omega_D &= \{\mathsf{Fdestination}, \mathsf{F}(\mathsf{destination} \wedge \neg\mathsf{delayed})\}\end{aligned}$$

In words, the robot's goal is to reach its destination without delays, with the primary value to never do anything dangerous, and the secondary value to never be annoying. Let $\overline{\Omega}'$ be the value base induced by $\overline{\Omega}$, Ω_D and $\mu = 3$.

Now we can compare the following 2-plans $\pi_1 = (\mathsf{ask}, \mathsf{move})$ and $\pi_2 = (\mathsf{horn}, \mathsf{move})$. If we assume that in the initial state the robot is blocked but far from an operating theatre, we can represent the histories generated from these plans as follows (each block contains exactly the propositions that are true in that state):



In this case $Sat(\overline{\Omega}', \pi_1, s_0, \gamma) = \{\mathsf{G}\neg\mathsf{dangerous}, \mathsf{G}\neg\mathsf{annoyed}, \mathsf{Fdestination}\} = A \supseteq \Omega_1 \cup \Omega_2$ whereas $Sat(\overline{\Omega}', \pi_2, s_0, \gamma) = \{\mathsf{G}\neg\mathsf{dangerous}, \mathsf{Fdestination}, \mathsf{F}(\mathsf{destination} \wedge \neg\mathsf{delayed})\} = B \supseteq \Omega_1 \cup \Omega_D$. Therefore π_1 will be preferred to π_2. However, if we change the morality level to 2, perhaps to represent an urgent delivery to an ongoing surgery, then we see that the robot will choose plan π_2 rather than π_1. This illustrates how we can adjust the morality level of the robot to reflect the urgency of its goals. If we move the example to the operating theatre (so now $\mathsf{theatre} \in s_0$ instead of $\neg\mathsf{theatre} \in s_0$), then the robot would not sound its horn even if the delivery was urgent, as Ω_1 still overrides Ω_D. This also means that for this robot we should restrict μ to 2 or 3 to ensure that being safe is always prioritised over goals. Furthermore, notice that for any lexicographic value structure containing exactly these values and goals, the set of non-dominated plan will always contain either π_1, π_2 or both, since A and B are exactly the qual-minimal contractions of $\cup\overline{\Omega}'$ given an initial state where the robot is blocked.

5 Computational Complexity

In this section we initiate the study of the computational complexity of ethical planning in our setting. We borrow our terminology from the work of Lang [17] on compact preference representation, but the problems we study have obvious counterparts in the planning literature, as should be clear from the proofs. In the interest of space all proofs can be found in the appendix.

We begin by studying the problem CONFLICT, which determines if a moral problem is also a moral conflict.

CONFLICT
Input: Moral problem $M = (\Omega, \gamma, s_0)$
Question: Is there some $k \in \mathbb{N}$ such that there is a k-plan π' such that $Sat(\Omega,\pi,s_0,\gamma) = \Omega$?

Theorem 1. *CONFLICT is PSPACE-complete.*

We then study the case of contractions, in particular, determining if a given moral problem is a qual-minimal contraction.

MINIMAL-CONTRACTION
Input: Moral problem $M = (\Omega, \gamma, s_0)$, moral problem $M' = (\Omega', \gamma, s_0)$
Question: Is M' a qual-minimal contraction of M?

Theorem 2. *MINIMAL-CONTRACTION is PSPACE-complete.*

Neither of these results are particularly technically advanced, indeed CONFLICT is almost exactly equivalent to PLANSAT from classical planning [8]. The purpose of these results is to indicate that quite apart from the issue of how a robot should select the best option when faced with a moral conflict, the task of identifying that the robot is facing a moral conflict and determining all of its options is extremely computationally difficult.

On the subject of planning, we begin by studying the problem COMPARISON, which given two k-plans π_1 and π_2, asks whether $\pi_1 \preceq_{\Delta}^{qual} \pi_2$. Despite the apparent complexity of our setting this problem can be solved efficiently:

COMPARISON
Input: Ethical planning domain $\Delta = (\gamma, s_0, \overline{\Omega})$, $k \in \mathbb{N}$, k-plans π_1, π_2
Question: is it the case that $\pi_1 \preceq_{\Delta}^{qual} \pi_2$?

Theorem 3. *COMPARISON is in P.*

We then move to the problem of non-dominance, i.e., the problem of determining if given a g-plan π_1 there exists a better k-plan wrt. $\preceq_{\Delta}^{qual}$ (where $g \leq k$).

NON-DOMINANCE
Input: Ethical planning domain $\Delta = (\gamma, s_0, \overline{\Omega})$, $k \in \mathbb{N}$, g-plan π for $g \leq k$
Question: is there a k-plan π' such that $\pi \preceq_{\Delta}^{qual} \pi'$ and $\pi' \not\preceq_{\Delta}^{qual} \pi$?

We show that this problem, as most instances of classical planning satisfaction, is PSPACE-complete:

Theorem 4. *NON-DOMINANCE is PSPACE-complete.*

Proposition 4. *Given an ethical planning domain* $\Delta = (\gamma, s_0, \overline{\Omega})$*, a k-plan* π *and* $S = Sat(\cup\overline{\Omega},\pi,s_0,\gamma)$ π *is non-dominated for* Δ *if and only if* $M = (S, \gamma, s_0)$ *is a* $\overline{\Omega}$*-qual-minimal contraction for* $(\cup\overline{\Omega}, \gamma, s_0)$*.*

Theorems 3 and 4 are to be interpreted as baseline results showing the computational feasibility of our setting for ethical planning with LTL_f. One clear direction for future work would expand on the computational complexity analysis, identifying tractable fragments and exploring their expressivity in ethical applications.

An important property for an ethical planner is *explainability*. While explaining why a particular plan was chosen is difficult to do succinctly (even for humans), a simpler problem is to explain why the chosen plan was better than another proposed alternative. Our approach enables this in a way that is both computationally straightforward and intuitively understandable to humans, since by the lexicographic ordering of plans there always exists a single value or set of values that decides between two plans.

6 Conclusion

We put forward a novel setting for ethical planning obtained by combining a simple logical temporal language with lexicographic preference modelling. Our setting applies to planning situations with a single agent who has deterministic and instantaneous actions to be performed sequentially in a static and known environment. Aside from the addition of values, our framework differs from classical planning in two aspects, by having multiple goals and by allowing temporal goals. In particular, the expressiveness of LTL means that we can express a wide variety of goals and values, including complex temporal values such as "if the weather is cold, close external doors immediately after opening them", with a computational complexity equivalent to that of standard planners. As a limitation, the system is less able to express values that tend to be satisfied by degree rather than absolutely or not at all. Among the multiple directions for future work that our definitions open, we plan to study the multi-agent extension with possibly conflicting values among agents, moving from plans to strategies (functions from states or histories to actions), from complete to incomplete information, and, most importantly, test our model by implementing it in simple robotics scenarios. Furthermore, given the computational complexity of Conflict, Mininal-Contraction and Non-Dominance, it may often be the case that in practical applications we cannot guarantee finding a non-dominated plan. Therefore, it would be valuable to find more tractable algorithms that at least guarantee some degree of approximation of a non-dominated plan, or restrictions (likely to the language or action theory) that improve tractability of the problem.

Acknowledgements This work is supported by the CNRS project LEXIA ("The Logic of Explanation: From Explainable to Explaining Legal Knowledge-based Systems").

Bibliography

[1] Alili, S., Alami, R., Montreuil, V.: A task planner for an autonomous social robot. In: Proceedings of the 9th International Symposium on Distributed Autonomous Robotic Systems (DARS). Springer (2008)

[2] Anderson, M., Anderson, S.L.: Geneth: a general ethical dilemma analyzer. Paladyn (Warsaw) **9**(1), 337–357 (2018)

[3] Arkin, R.C., Ulam, P., Wagner, A.R.: Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. Proceedings of the IEEE **100**(3), 571–589 (2012)

[4] Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., Talamadupula, K., Tenenbaum, J.B., Kleiman-Weiner, M.: When is it acceptable to break the rules? Knowledge representation of moral judgement based on empirical data. CoRR **abs/2201.07763** (2022)

[5] Benzmüller, C., Parent, X., van der Torre, L.W.N.: Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. Artificial Intelligence **287**, 103–348 (2020)

[6] Berreby, F., Bourgne, G., Ganascia, J.: A declarative modular framework for representing and applying ethical principles. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS) (2017)

[7] Bienvenu, M., Fritz, C., McIlraith, S.A.: Planning with qualitative temporal preferences. In: Doherty, P., Mylopoulos, J., Welty, C.A. (eds.) Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR). pp. 134–144. AAAI Press (2006)

[8] Bylander, T.: The computational complexity of propositional STRIPS planning. Artif. Intell. **69**(1-2), 165–204 (1994)

[9] Copp, D.: The Oxford Handbook of Ethical Theory. Oxford University Press (2007)

[10] De Giacomo, G., Vardi, M.Y.: Linear temporal logic and linear dynamic logic on finite traces. In: Rossi, F. (ed.) Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI). pp. 854–860. IJCAI/AAAI (2013)

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

[12] Dennis, L.A., del Olmo, C.P.: A defeasible logic implementation of ethical reasoning. In: First International Workshop on Computational Machine Ethics (CME) (2021)

[13] Evans, K., de Moura, N., Chauvier, S., Chatila, R., Dogan, E.: Ethical decision making in autonomous vehicles: The av ethics project. Science and engineering ethics **26**(6), 3285–3312 (2020)

[14] Harsanyi, J.: Utilitarianism and beyond. In: Sen, A.K., Williams, B. (eds.) Morality and the theory of rational behaviour. Cambridge University Press, Cambridge (1982)
[15] Jenkins, R., Talbot, B., Purves, D.: When robots should do the wrong thing. In: Robot Ethics 2.0. Oxford University Press, New York (2017)
[16] Kment, B.: Varieties of Modality. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Spring 2021 edn. (2021)
[17] Lang, J.: Logical preference representation and combinatorial vote. Annals of Mathematics and Artificial Intelligence **42**(1-3), 37–71 (2004)
[18] Levi, I.: Hard Choices: Decision Making Under Unresolved Conflict. Cambridge University Press (1990)
[19] Lindner, F., Mattmüller, R., Nebel, B.: Moral permissibility of action plans. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI). pp. 7635–7642. AAAI Press (2019)
[20] Loreggia, A., Mattei, N., Rossi, F., Venable, K.B.: On the distance between cp-nets. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS) (2018)
[21] Loreggia, A., Rossi, F., Venable, K.B.: Modelling ethical theories compactly. In: The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence (2017)
[22] Lorini, E.: A logic for reasoning about moral agents. Logique & Analyse **58**(230), 177–218 (2015)
[23] Lorini, E.: Logics for games, emotions and institutions. FLAP **4**(9), 3075–3113 (2017)
[24] Lorini, E.: A logic of evaluation. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 827–835. ACM (2021)
[25] Müller, V.C.: Ethics of Artificial Intelligence and Robotics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Summer 2021 edn. (2021)
[26] Pnueli, A.: The temporal logic of programs. In: Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS) (1977)
[27] Powers, T.M.: Deontological machine ethics. In: Anderson, M., Anderson, S.L., Armen, C. (eds.) Association for the Advancement of Artificial Intelligence Fall Symposium Technical Report (2005)
[28] Rossi, F., Mattei, N.: Building ethically bounded AI. In: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI) (2019)
[29] Searle, J.: Rationality in Action. Cambridge University Press, MIT Press (2001)
[30] Sen, A.: On Ethics and Economics. Basil Blackwell (1987)
[31] Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, New York (2016)
[32] Vanderelst, D., Winfield, A.F.T.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research **48**, 56–66 (2018)
[33] Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V.R., Yang, Q.: Building ethics into artificial intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI) (2018)

Appendix: Missing proofs

Proof of Proposition 3

Proof. Let $M = (\Omega, \gamma, s_0)$ and let $\overline{\Omega}$ be a lexicographic ordering of Ω of degree m. Suppose $M' = (\Omega, \gamma, s_0)$ is a $\overline{\Omega}$-qual-minimal contraction of M. Suppose for contradiction that M' is not qual-minimal. Then there exists some contraction $M'' = (\Omega'', \gamma, s_0)$ such that $\Omega' \subset \Omega''$. Therefore there exists some value $\varphi \in \Omega$ such that $\varphi \in \Omega''$ and $\varphi \notin \Omega'$. Let p be the priority level of φ in $\overline{\Omega}$ (so $\varphi \in \Omega_p$).

Since $\Omega' \subset \Omega''$, we know that for all $1 \leq k \leq m$, either $\Omega' \cap \Omega_k \subset \Omega'' \cap \Omega_k$ or $\Omega' \cap \Omega_k = \Omega'' \cap \Omega_k$. We also know that $\Omega' \cap \Omega_p \subset \Omega'' \cap \Omega_p$. Therefore there must be some $p' \leq p$ such that:$1 \leq p' \leq m$ and $\Omega' \cap \Omega_{p'} \subset \Omega'' \cap \Omega_{p'}$ and $\forall 1 \leq k' < p', \Omega' \cap \Omega'_k = \Omega'' \cap \Omega'_k$.

By a very similar method, we can derive a contradiction if we suppose that M' is $\overline{\Omega}$-quant-minimal but not qual-minimal.

Proof of Theorem 1

Proof. To show that Conflict is PSPACE-hard, we show a reduction from the classical planning problem PLANSAT for propositional STRIPS planning [8]. In this problem we have a set of *conditions* (propositions) that can be true of false, an initial state which is a collection of conditions, a set of *operators* (actions) that have preconditions and postconditions as sets of satisfiable conjunctions of positive and negative conditions, and a single *goal* which is a conjunction of positive and negative conditions. We then attempt to find a finite sequence of operators that acheives the goal from the starting state. For a more complete description and complexity results, see [8].

To perform the reduction, set $\Omega = \{\omega\}$ where ω is our goal. Creating an action theory γ from the set of operators and a start state s_0 can be done in polynomial time. Then Conflict applied to (Ω, γ, s_0) returns true if and only if PLANSAT would return false.

To show that Conflict is PSPACE-complete, we show a reduction from Conflict to PLANSAT. Given a moral problem $M = (\Omega, \gamma, s_0)$ define φ as the conjunction of all formulas in Ω, then set φ as our goal. We can generate a set of operators from γ and an initial state from s_0 in polynomial time. Then PLANSAT returns true if and only if Conflict would return false.

Proof of Theorem 2

Proof. To show that Minimal-Contraction is PSPACE-hard, we show a reduction from Conflict . Given a value problem N, set $M = M' = N$. Then by proposition 1, Minimal-Contraction will return true if and only if Conflict would return false.

To show that Minimal-Contraction is PSPACE-complete, we provide a basic algorithm that uses polynomial space. First, use Conflict to check if M' is a conflict, if it is, return false. If not, let $A = \Omega \backslash \Omega'$. For each $a \in A$, run Conflict on $(\Omega' \cup a, \gamma, s_0)$. If Conflict returns false on any of these checks, return false, else, return true.

Proof of Theorem 3

Proof. Recall that to compare two plans π_1 and π_2 we need an ethical planning domain $\Delta = (\gamma, s_0, \overline{\Omega})$ where γ is an action theory, s_0 is an initial state and $\overline{\Omega}$ is a value base. Following Definition 10, plan π_1 is better than π_2 if the history generated by π_1 is lexicographically preferred to the history generated by π_2 according to the ranked values in $\overline{\Omega}$.

We begin by showing that generating the unique history associated to a k-plan can be done in polynomial time. Then we show that evaluating an LTL_f formula over this history can be done in polynomial time, concluding that Comparison is in P since we can give an answer by model checking all formulas in $\overline{\Omega}$ over the histories generated by the two plans.

Generating H_{act} associated to π can be done by making a copy of the plan π that returns skip whenever the input is greater than k. This can be done in polynomial time. We then set $H_{st}(0) = s_0$, then for each $H_{st}(i)$ we use γ to generate $H_{st}(i+1)$ in polynomial time by model checking all formulas in γ. We only have to do this k times.

Let us now show that LTL_f formulas can be checked in polynomial time on the history generated by a k-plan. Suppose we have an LTL_f formula φ of length n, and a history H associated with some k-plan π. We proceed by strong induction on n. For the purpose of this proof, suppose an algorithm that takes φ and H as inputs. Let H_j be the history such that $H_j{}_{act}(i) = H_{act}(i+j)$ and $H_j{}_{st}(i) = H_{st}(i+j)$.

Base case. Suppose n = 1, then $\varphi = p$ for some $p \in \mathit{Prop}$. Then we can determine if $p \in H_{st}(0)$ in polynomial time.

Inductive step. Suppose $n > 1$ and that the claim holds for all $m < n$. Then we have several options for φ.

1. $\varphi = \neg\psi$. Then by inductive hypothesis we can determine in polynomial time if $H \vDash \psi$ and thus if $H \vDash \varphi$.
2. $\varphi = \psi \wedge \chi$. Then we can determine in time P if $H \vDash \psi$ and $H \vDash \chi$.
3. $\varphi = \mathsf{X}\psi$. Then we can determine (in polynomial time) if $H_1 \vDash \psi$.
4. $\varphi = \psi \ \mathsf{U} \ \chi$. Then for $0 < j < k$ we can determine if $H_0, H_1, ..., H_{j-1} \vDash \psi$ and $H_j \vDash \chi$ in polynomial time. Therefore this whole process can be done in polynomial time.

To conclude, suppose we have an ethical planning domain $\Delta = (\gamma, s_0, \overline{\Omega})$ and k-plans π_1 and π_2. By the previous steps we can generate H^{π_1} and H^{π_2}, and for each $\omega \in \overline{\Omega}$ we can determine whether H^{π_1} and $H^{\pi_2} \vDash \omega$ in polynomial time. Therefore evaluating for every possible value can be done in polynomial time. Determining $\preceq_\Delta^{qual}$ involves checking every value a maximum of once, so we can conclude that Comparison is in P.

Proof of Theorem 4

Proof. To show that Non-Dominance is PSPACE-hard, we show a reduction from the classical planning problem PLANMIN for propositional STRIPS planning [8]. In this problem we have a set of *conditions* (propositions) that can be

true of false, a set of *operators* (actions) that have preconditions and postconditions as sets of satisfiable conjunctions of positive and negative conditions, and a single *goal* which is a satisfiable conjunction of positive and negative conditions. We then attempt to find a sequence of k or less operators that acheives the goal from the starting state. For a more complete description and complexity results, see [8].

To perform the reduction, set $\overline{\Omega} = \Omega_1$ where $\Omega_1 = \{\omega\}$ where ω is our goal. Creating an action theory from the set of operators can be done in polynomial time. Then, generate a random k-plan π and check if $H^\pi \models \omega$, if it does then we are done. If it does not then non-dominance applied to π is equivalent to PLANMIN.

To show that Non-Dominance is PSPACE-complete we provide a basic algorithm that uses polynomial space. Given $\Delta = (\gamma, s_0, \overline{\Omega})$ and k-plan π, check every possible plan π' for γ and s_0 and check Comparison against π. Terminate once a plan is found that dominates π or once all plans have been checked. This algorithm only needs two plans in memory at any one time (π and the plan being compared to π), and therefore it only requires polynomial space.