Logic-Based Ethical Planning

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

[1]IRIT, CNRS and University of Toulouse, France
[2]LAAS-CNRS, France

Abstract

In this paper we propose a model for planning with multiple values, with intended application to ethics and robotics. Our language for ethical planning combines linear temporal logic with lexicographic preference modelling, allowing us to assess plans both with respect to an agent's values and their desires and introducing the novel concept of morality level of an agent. We provide some foundational complexity results for our setting, and we discuss potential applications to robotics.

Keywords

KR and ethics, Linear temporal logic, Compact preference representation, Robotics

1. Introduction

In ethical planning the planning agent has to find a plan for promoting a certain number of ethical values. Unlike classical planning in which the goal to be achieved is unique, in ethical planning the agent can have multiple and possibly conflicting values. Consequently, in ethical planning the agent needs to evaluate and compare different plans depending on how many and which values are promoted by each of them.

Including ethical considerations in robotics planning requires (at least) two steps. First, design a language to express these considerations as values, taking in mind that they often conflict both amongst themselves, and with the goal. Such a value representation language needs to be compact and computationally tractable. Second, design an algorithm that compares 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 simplicity 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 any conflicts between goals, desires, and best practice in a unified planning model.

IJCAI 2022: Cognitive Aspects of Knowledge Representation, July 2022, Vienna, Austria
*Corresponding author.
umberto.grandi@irit.fr (U. Grandi); emiliano.lorini@irit.fr (E. Lorini); timothy.parker@irit.fr (T. Parker); rachid.alami@laas.fr (R. Alami)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3251/paper8.pdf

There is considerable research in the field of ethics and AI, see Müller (2021) for a general overview. Popular ethical theories for application are consequentialism, deontology, and virtue ethics.[1] Our approach is designed to be theory neutral and should be able to handle most ethical systems, though it is probably a most natural fit for pluralistic consequentialism [6].

In terms of practical applications of ethics to robotics, there are approaches both in terms of formal models [7] and allowing agents to learn ethical values [8]. Yu et al. (2018) 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) using a compact representation language to aid with decision-making in an ethically sensitive domain. The former are based on different methodologies including event calculus (ASP) [10], epistemic logic and preference logic [11, 12], BDI (belief, desire, intention) agent language [13], classical higher-order logic (HOL) [14]. The latter was presented in "blue sky" papers [15, 16] complemented with a technical study of distances between CP-nets [17] and, more recently, with an empirical study on human ethical decision-making [18].

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 by Arkin et al. (2012) and Vanderelst and Winfield (2018). Evans et al. (2020) introduces ethical decision-making by way of considering the competing ethical claims of various agents on a robot's behaviour. Other work helps robots to produce socially acceptable plans by assigning weights to social rules [22] .

2. 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.

2.1. LTL Language

Let $Prop$ be a countable set of atomic propositions and let Act be a finite non-empty set of action names. The set of states is $S = 2^{Prop}$. In order to represent the agent's values, we introduce the language of LTL (Linear Temporal Logic) [23], noted $\mathcal{L}_{\mathsf{LTL}}(Prop)$ (or $\mathcal{L}_{\mathsf{LTL}}$), defined by the 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. We can also add the operators "henceforth" (G) and "eventually" (F) which are defined in the usual way: $\mathsf{G}\varphi \stackrel{\text{def}}{=} \neg(\top \,\mathsf{U}\, \varphi)$ and $\mathsf{F}\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi$.

2.2. Histories, Actions and Plans

History Histories describe how a state changes over time. In our model a history describes the state of the environment after each action performed by the agent, as well as the actions themselves. We define a history to be a pair $H = (H_{st}, H_{act})$ with $H_{st} : \mathbb{N} \longrightarrow S$ and H_{act} :

[1] See Copp (2007) for a philosophical introduction, and Jenkins et al. (2017), Powers (2005), and Vallor (2016) for a discussion of these three theories in robotics.

$\mathbb{N} \longrightarrow Act$. We define $Hist$ to be the set of all possible histories. Semantic interpretation of formulas in $\mathcal{L}_{\mathsf{LTL}}$ relative to a history $H \in Hist$ and a time point $k \in \mathbb{N}$ goes as follows (boolean cases are as usual):

$$\begin{aligned}
H, k \models p \quad &\iff p \in H_{st}(k), \\
H, k \models \mathsf{X}\varphi \quad &\iff H, k+1 \models \varphi, \\
H, k \models \varphi_1 \,\mathsf{U}\, \varphi_2 \quad &\iff \exists k' \geq k : H, k' \models \varphi_2 \text{ AND} \\
&\qquad \forall k'' \geq k : \text{IF } k'' < k' \text{ THEN } H, k'' \models \varphi_1.
\end{aligned}$$

Action 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 functions, where $\gamma^+ : Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}, \gamma^- : Act \times Prop \longrightarrow \mathcal{L}_{\mathsf{PL}}$ ($\mathcal{L}_{\mathsf{PL}}$ is propositional logic).

Therefore if $H_{act}(k) = a \in Act$ (meaning that action a is performed at time k) and $H, k \vDash \gamma^+(a, p)$ then $p \in H_{st}(k+1)$ (meaning that p is true at time $k+1$). Similarly, if $H_{act}(k) = a$ and $H, k \vDash \gamma^-(a, p)$ then $p \notin H_{st}(k+1)$. If both or neither of $\gamma^+(a, p)$ and $\gamma^-(a, p)$ are true at time k (where $H_{act}(k) = a$) then $p \in H_{st}(k+1) \Leftrightarrow p \in H_{st}(k)$ (p does not change).

We also suppose that every action theory contains the special action *skip*, such $\forall a \in Act, p \in Prop, \gamma^+(a, p) = \gamma^-(a, p) = p \wedge \neg p$ (this action does nothing).

Plan Given $k \in \mathbb{N}$, a k-plan is a function $\pi : \{0, \ldots, k\} \longrightarrow Act$. In other words, a plan is a sequence of actions. Since actions are deterministic, given a plan π, an action theory γ and an initial state s_0 it is possible to create the corresponding history by setting $H_{act}(t) = \pi(t)$ for $0 \leq t \leq k$ and $H_{act}(t) = skip$ for $t > k$, setting $H_{st}(0) = s_0$ and generating the rest of H_{st} using γ. Given a set of LTL-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.

2.3. Values and Desires

Values In our setting an agent's values are represented by sets of LTL formulas ordered according to their priority level (Ω_1 are the most important and Ω_m are the least). Values can take various forms, but many values can be interpreted as saying that either a certain state of affairs must always/never hold, or should hold at some point. These can be expressed as $\mathsf{G}\varphi/\mathsf{G}\neg\varphi$ (example: "humans must not be harmed") and $\mathsf{F}\varphi$ (example: "the dog should be taken for a walk"). Since our model can handle an arbitrary number of prioritised value sets, this means we can handle values of various types, including moral values, social norms and values of best practice.

Definition 1 (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* $\Omega_k \subseteq \mathcal{L}_{\mathsf{LTL}}$ *for every* $1 \leq k \leq m$.

Following [12], we call *evaluation* the operation of computing an ideality ordering over plans from a value base. Building on classical preference representation languages [24], 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. It is also possible to define a quantitative ordering based on the number of satisfied values at each level.

Definition 2 (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:*

$$(i)\ \exists 1 \leq k \leq m \text{ s.t. } Sat(\Omega_k,\pi_1,s_0,\gamma) \subseteq Sat(\Omega_k,\pi_2,s_0,\gamma),$$
$$(ii)\ \forall 1 \leq k' < k, Sat(\Omega_{k'},\pi_1,s_0,\gamma) = Sat(\Omega_{k'},\pi_2,s_0,\gamma).$$

Desires We expect autonomous ethical agents to be driven by both ethical values and also endogenous motivations, also called *desires* or *goals*. 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 3 (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 1);*
- *$\Omega_D \subseteq \mathcal{L}_{\mathsf{LTL}}$ is the agent's set of desires or goals;*
- *$\mu \in \{0, \ldots, dg(\overline{\Omega})\}$ 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. The lower the agent's degree of morality, the higher the "goal set" is ranked relative to the agent's values. This works as follows: for morality level μ and mixed-motive planning domain $M = (\gamma, s_0, \overline{\Omega}, \Omega_D, \mu)$ the induced ethical planning domain is $M' = (\gamma, s_0, \overline{\Omega}')$ where $\overline{\Omega}' = \Omega_1, ..., \Omega_{\mu-1}, \Omega_D, \Omega_\mu, ..., \Omega_m$.

3. Complexity Results

We borrow our terminology from the work of Lang (2004) on compact preference representation, but the problems we study have obvious counterparts in the planning literature. Our first problem is Comparison, which takes as input an initial state s_0, an ethical planning domain Δ, two k-plans π_1 and π_2, and asks whether $\pi_1 \preceq_{\Delta}^{qual} \pi_2$. Our second problem is Non-Dominance, i.e., the problem of determining if given a k-plan π_1 for ethical planning domain Δ there exists a better k-plan wrt. $\preceq_{\Delta}^{qual}$.

Despite the complexity of our setting, Comparison can be solved quite efficiently (it is in P). Our second problem, Non-Dominance, like most instances of classical planning satisfaction, is PSPACE-complete. These should be interpreted as baseline results showing the computational feasibility of our setting for ethical planning with LTL. Formal results and proofs have been omitted in the interest of space and can be provided on request.

4. 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 goals 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.

With regards to the current literature on ethical planning, we feel that one of the strengths of our model is its relative simplicity and ease of understanding, which could be an important factor for the acceptance of ethical robots by the general public. A similar idea to our lexicographic ordering of values is discussed in Dennis et al. (2016), although they use propositional rather than temporal logic. Possibly the most significant feature of our model is the concept of the morality level of an agent or goal, as this appears to be a novel idea in the field of ethical planning and should allow robots to appropriately handle goals with vastly different levels of urgency/importance.

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, expand on the computational complexity analysis and, most importantly, test our model by implementing it in simple robotics scenarios.

References

[1] V. C. Müller, Ethics of Artificial Intelligence and Robotics, in: The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2021.

[2] D. Copp, The Oxford Handbook of Ethical Theory, Oxford University Press, 2007.

[3] R. Jenkins, B. Talbot, D. Purves, When Robots Should Do the Wrong Thing, in: Robot Ethics 2.0, Oxford University Press, 2017.

[4] T. M. Powers, Deontological Machine Ethics, in: Association for the Advancement of Artificial Intelligence Fall Symposium Technical Report, 2005.

[5] S. Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Oxford University Press, 2016.

[6] A. Sen, On Ethics and Economics, Basil Blackwell, 1987.

[7] L. A. Dennis, C. P. del Olmo, A Defeasible Logic Implementation of Ethical Reasoning, in: First International Workshop on Computational Machine Ethics (CME), 2021.

[8] M. Anderson, S. L. Anderson, Geneth: a general ethical dilemma analyzer, in: Paladyn (Warsaw), De Gruyter, 2018.

[9] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, Q. Yang, Building Ethics into Artificial

Intelligence, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 2018.
[10] F. Berreby, G. Bourgne, J. Ganascia, A Declarative Modular Framework for Representing and Applying Ethical Principles, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS), 2017.
[11] E. Lorini, A logic for reasoning about moral agents, in: Logique & Analyse, 2015.
[12] E. Lorini, A Logic of Evaluation, in: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2021.
[13] L. A. Dennis, M. Fisher, M. Slavkovik, M. Webster, Formal verification of ethical choices in autonomous systems, in: Robotics and Autonomous Systems, 2016.
[14] C. Benzmüller, X. Parent, L. W. N. van der Torre, Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support, in: Artificial Intelligence, 2020.
[15] A. Loreggia, F. Rossi, K. B. Venable, Modelling Ethical Theories Compactly, in: The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[16] F. Rossi, N. Mattei, Building Ethically Bounded AI, in: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI), 2019.
[17] A. Loreggia, N. Mattei, F. Rossi, K. B. Venable, On the Distance Between CP-nets, in: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), 2018.
[18] E. Awad, S. Levine, A. Loreggia, N. Mattei, I. Rahwan, F. Rossi, K. Talamadupula, J. B. Tenenbaum, M. Kleiman-Weiner, When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgement Based on Empirical Data, in: CoRR abs/2201.07763, 2022.
[19] R. C. Arkin, P. Ulam, A. R. Wagner, Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception, in: Proceedings of the IEEE, 2012.
[20] D. Vanderelst, A. F. T. Winfield, An architecture for ethical robots inspired by the simulation theory of cognition, in: Cognitive Systems Research, 2018.
[21] K. Evans, N. de Moura, S. Chauvier, R. Chatila, E. Dogan, Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project, in: Science and engineering ethics, Springer Netherlands, 2020.
[22] S. Alili, R. Alami, V. Montreuil, A Task Planner for an Autonomous Social Robot, in: Proceedings of the 9th International Symposium on Distributed Autonomous Robotic Systems (DARS), 2008.
[23] A. Pnueli, The temporal logic of programs, in: Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS), 1977.
[24] J. Lang, Logical Preference Representation and Combinatorial Vote, in: Annals of Mathematics and Artificial Intelligence, 2004.