Ethical Markov Decision Processes with Moral Worth as Rewards

Extended Abstract

Mihail Stojanovski
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
mihail.stojanovski@unicaen.fr

Nadjet Bourdache
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
nadjet.bourdache@unicaen.fr

Grégory Bonnet
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
gregory.bonnet@unicaen.fr

Abdel-Illah Mouaddib
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
abdel-illah.mouaddib@unicaen.fr

ABSTRACT

We propose an expressive framework for specifying ethical behaviours, called Ethical Markov Decision Processes (E-MDPs) that extends classical MDPs with the explicit representation of moral values – positive or negative – that the agent's decisions may promote or demote.

KEYWORDS

Computational ethics; Planning; Markov decision processes

ACM Reference Format:
Mihail Stojanovski, Nadjet Bourdache, Grégory Bonnet, and Abdel-Illah Mouaddib. 2024. Ethical Markov Decision Processes with Moral Worth as Rewards: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

With the development of automated agents, many issues may come forth in certain cases due to their lack of an ethical component. For instance, in autonomous vehicle settings, driving at certain speeds can be considered as more or less prudent depending on the circumstances, e.g. driving near a school if classes have just ended. If we want to express prudence as an moral value, driving at a reduced speed near a school is preferred to driving at the speed limit. This would not be a strict constraint, but rather an ethical guideline which the agent should adhere to. Thus, it is of interest to embed automated agent decision-making processes with moral values, moral rules or ethical rules on which it can reason. Many approaches in the literature focus on qualitative logic-based models, e.g. value-based argumentation [1], modal logic [5], non-monotonic logic [2], or BDI architectures [3]. However in this article, we focus on the particular quantitative decision-making processes of *Markov Decision Processes* (MDPs). Dealing with ethics within this kind of models is relatively new and, for the time being, the proposed approaches still lack of genericity [4, 6–8]. To integrate ethics more easily into the reasoning of automated agents, it is necessary to create a generic model, which can express different kinds of ethical behaviours with the same basic components. We propose the *Ethical Markov Decision Process* (E-MDP) model which extends MDPs with explicit moral values and uses a multi-criteria reward function that distinctly represents the satisfaction (promotion) and violation (demotion) of these values. This offers a generic way of representing ethical principles by focusing on their underlying values, and gives the model sufficient expressivity to capture multiple ethical principles which can focus on different values at the same time.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 ETHICAL MDPS

Ethical decision-making necessitates an *ethical context* which is specific to the decision-maker. This context represents the moral values of the agent, which can be positive (prudence, honesty, generosity) but also negative (greed, dishonesty, selfishness). It is important to note that the polarity of these values can differ from one agent to another. For example *obedience* is a moral value which can be positive or negative depending on the agent's ethics.

Definition 1 (Ethical context). *Let $\mathcal{V} = \{v_1, \ldots, v_k\}$ be a set of moral values. An ethical context C is a tuple $\{C_1, \ldots, C_k\}$ where $C_i \in \{1, -1, 0\}$. The valuation 1 (resp. -1 and 0) means that the value is considered as positive (resp. negative and neutral) for the agent. Let $\mathcal{G}_C$ (resp. $\mathcal{B}_C$ and $\mathcal{N}_C$) be the set of positive (resp. negative and neutral) values in the context C: $\mathcal{G}_C = \{v_i \in \mathcal{V} : C_i = 1\}$ (resp. $\mathcal{B}_C = \{v_i \in \mathcal{V} : C_i = -1\}$ and $\mathcal{N}_C = \{v_i \in \mathcal{V} : C_i = 0\}$).*

Notice that values do not differ in importance between each other: one positive value cannot be considered better than another (resp. negative, worse). Dealing with hierarchical values is let for future works. *Ethical Markov Decision Processes* consist of MDPs extended with morals, ie. values associated to transitions, and ethics, which dictates how the agent decides with respect to morals.

Definition 2 (Ethical MDP). *An E-MDP is a six tuple $\langle S, A, \mathcal{T}, C, \mathcal{E}, R\rangle$ where S, A and $\mathcal{T}$ are the classical set of states, set of actions, and transition function, C is an ethical context, $\mathcal{E}$ is a moral worth function and R is an ethical reward function.*

To express ethical principles, we need to describe the alignment of the agent's behaviour with these values: a value can be promoted or demoted by a decision. Promoting a value means that the agent's behaviour is aligned with it, while demoting a value means that it is violated by the agent's behaviour (whether the value is positive or negative). Hence, we introduce the moral evaluation (moral worth) of transitions, which is a measure of whether the transition promotes, demotes, or is irrelevant to a given value of the context.

Definition 3 (Transition moral worth). *Each transition (s, a, s') $\in$ $S \times A \times S$ where $\mathcal{T}(s, a, s') > 0$ is associated with a tuple $\mathcal{E}(s, a, s') = \langle \mathcal{E}(s, a, s')_1, \ldots, \mathcal{E}(s, a, s')_k \rangle$ representing its moral worth. The i-th element $\mathcal{E}(s, a, s')_i$ of $\mathcal{E}(s, a, s')$ takes value in $\{1, -1, 0\}$, which designates whether the moral value $v_i \in \mathcal{V}$ is respectively promoted, demoted or irrelevant.*

E-MDPs' reward function describes how the moral values of the agent's ethical context are aligned with his behaviour. The agent can promote or demote a value, giving four distinct behaviours: promoting a positive or negative value – i.e. *causing good or harm*, and demoting a positive or negative value – i.e. *repairing good or harm*. Here, repairing consists in a decision that does not allow good or harm to continue to exist. It means that, to repair, the agent must have at least caused some amount of good or harm in the past. Treating pre-existing good or harm is let for further work.

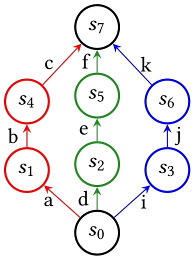
Definition 4 (Ethical reward function). *The reward function R outputs a quadruple where $\triangle$ counts caused good, ∇ caused harm, $\overline{\triangle}$ repaired good, $\overline{\nabla}$ repaired harm. Hence, $R(s, a, s') = \langle \Delta, \nabla, \overline{\Delta}, \overline{\nabla} \rangle$ with $\Delta, \nabla, \overline{\Delta}, \overline{\nabla}$ being:*

$$\Delta = \sum_{v_i \in \mathcal{G}_C} x_i \textit{ and } \nabla = \sum_{v_i \in \mathcal{B}_C} x_i \textit{ where } x_i = \begin{cases} 1 & \textit{if } \mathcal{E}(s, a, s')_i = 1, \\ 0 & \textit{otherwise.} \end{cases}$$

$$\overline{\Delta} = \sum_{v_i \in \mathcal{G}_C} x_i \textit{ and } \overline{\nabla} = \sum_{v_i \in \mathcal{B}_C} x_i \textit{ where } x_i = \begin{cases} 1 & \textit{if } \mathcal{E}(s, a, s')_i = -1, \\ 0 & \textit{otherwise.} \end{cases}$$

To obtain a policy an ethical value function ($V^{\pi}_{\star}$ where $\star \in \{\triangle, \nabla, \overline{\triangle}, \overline{\nabla}\}$) is used, which is an adaptation of the Bellman equation and uses the ethical reward function in place of the classical reward function. As causing and repairing harm and good are distinct and potentially conflicting, the notion of optimality becomes subjective. Thus, we want to be able to express them explicitly and do not aggregate all four aspects. At this point, a question arises: "How should an agent make a decision based on an ethical context?" It seems natural that ethical decision-making should maximise the promoted positive moral values and minimise the promoted negative values. Furthermore, it should maximise the demoted negative values and minimise the demoted positive values. But do the ends justify the means? In other words, do we seek to produce the greater good, which may result from doing harm, or do we focus on doing the least harm possible, which may in some circumstances prevent us from doing a lot of good? As there is no hierarchy between values and as we do not explicitly express a trade-off between harm and good, we do not seek for a "fair" and "balanced" optimisation criterion. Hence, we choose that it is ethical to firstly avoid doing harm as much as possible, and then to focus on doing the most good in the remaining decision space. To this end, we use a lexicographical order that consists of giving preference to the policies that cause the least harm, and then, from the set of policies with the least harm, we choose those that do the most good.

$$\pi^*_B \in \operatorname*{argmin}_{\pi} V^{\pi}_{\nabla}(s) - V^{\pi}_{\overline{\nabla}}(s) + \epsilon V^{\pi}_{\overline{\nabla}}(s) \text{ where } 0 < \epsilon < 1. \quad (1)$$



T	$\mathcal{E}(T)_n$	$\mathcal{E}(T)_p$
(s_0, a, s_1)	1	1
(s_1, b, s_4)	1	1
(s_4, c, s_7)	0	1
(s_0, d, s_2)	1	1
(s_2, e, s_5)	1	1
(s_5, f, s_7)	-1	0
(s_0, i, s_3)	1	1
(s_3, j, s_6)	0	1
(s_6, k, s_7)	0	0

Figure 1: Example for policy preference

To prevent the agent from being absolved from doing harm by repairing the caused harm (totally or partially) we weighted the value by the repaired harm. This motivates the agent towards avoiding harm altogether, instead of causing harm intentionally just to repair it later. In addition to this, if only a subtraction between the caused and repaired harm was considered, policies in which harm was caused and fully repaired would become indistinguishable from ones in which no harm was done at all. This also extends to the optimal policies with regard to the positive context values, the equation for which is found below.

$$\pi^*_G \in \operatorname*{argmax}_{\pi \in \pi^*_B} V^{\pi}_{\triangle}(s) - V^{\pi}_{\overline{\triangle}}(s) - \epsilon' V^{\pi}_{\overline{\triangle}}(s) \text{ where } 0 < \epsilon' < 1. \quad (2)$$

It should be noted that the optimal policy w.r.t. good is chosen from the policies which are already optimal w.r.t. harm. This choice constrains the notion of the optimal policy w.r.t. good, however it assures that the agent will not consider doing harm as a means to obtain a greater good. Once the two criteria have been satisfied we obtain a set of policies in which the agent causes the most possible good while causing the least amount of harm. To better understand how the optimal policies are ranked, Figure 1 represents three policies: red, green and blue. The table shows the positive ($\mathcal{E}(T)_p$) and negative ($\mathcal{E}(T)_n$) moral worth for each transition of these policies. Here, we can see that no policy optimises all the criteria at the same time. Indeed, the red policy maximises the caused good ($\Delta = 3$ against $\Delta = 2$ for the blue and green policies), the green policy maximises the repaired harm ($\overline{\nabla} = 1$ against $\overline{\nabla} = 0$ for the red and blue policies), while the blue policy minimises the caused harm ($\nabla = 1$ against $\nabla = 2$ for the red and green policies). Using Equations 1 and 2 we deduce the following preference order: the blue policy is the most preferred (it causes the least harm), the green policy is second (it causes as much harm as the red policy, but repairs some of it), and the red policy is the worst (even though it does the most good).

3 CONCLUSION

We proposed the E-MDP model that explicitly integrates ethics into MDPs as positive and negative moral values, which can be promoted or demoted. It uses a quadruplet reward function to optimise first causing as little harm as possible, then causing as much good as possible. This model is generic and can be set up to integrate different ethical frameworks by changing the ethical context and the moral worth rewards.

REFERENCES

[1] Trevor Bench-Capon. 2003. Persuasion in Practical Argumentation Using Value-Based Argumentation Frameworks. *J. Log. Comput.* 13, 3 (2003), 429–448.

[2] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2017. A Declarative Modular Framework for Representing and Applying Ethical Principles. In *16th AAMAS*. 96–104.

[3] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. 2016. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *15th AAMAS*. 1106–1114.

[4] Nelson De Moura, Raja Chatila, Katherine Evans, Stéphane Chauvier, and Ebru Dogan. 2020. Ethical decision making for autonomous vehicles. In *IVS*. 2006–2013.

[5] Emiliano Lorini. 2012. On the Logical Foundations of Moral Agency. In *11th DEON (LNCS, Vol. 7393)*. Springer-Verlag, 108–122.

[6] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically Compliant Planning within Moral Communities. In *4th AIES*. 188–198.

[7] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez-Aguilar. 2020. A Structural Solution to Sequential Moral Dilemmas. In *19th AAMAS*. 1152–1160.

[8] Justin Svegliato, Samer Nashed, and Shlomo Zilberstein. 2021. Ethically Compliant Sequential Decision Making. In *35th AAAI*. 11657–11665.