Ethical Planning with Multiple Temporal Values

Timothy Parker, Umberto Grandi, Emiliano Lorini, Aurélie Clodic, Rachid Alami

June 2022

Ethical Planning with Multiple Temporal Values[1]

Timothy PARKER [a] Umberto GRANDI [a] Emiliano LORINI [a] Aurélie CLODIC [b]
Rachid ALAMI [b]

[a] *IRIT, CNRS and University of Toulouse, France*
[b] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*

Abstract. This paper describes an approach to ethical planning in robotics that combines classical planning, Linear Temporal Logic and lexicographically ordered value sets. We also discuss the features and limitations of the model, with particular attention to its philosophical implications.

Keywords. Ethical Planning, Temporal Values, Adjustable Morality

1. Introduction

In this paper we present a framework for an ethical planner with potential applications to robotics. We will first introduce and explain the model, give some explanatory examples and then discuss various features and limitations of it.

The foundation for our model is classical planning, a well-studied problem in computer science [1]. Given an initial state, a set of actions and a goal, we then attempt to find a plan (sequence of actions) that achieves the goal from the starting state. For simplicity, our current model assumes a single-agent setting with perfect knowledge and fully deterministic actions. Plans are then compared according to a lexicographically ordered value set, with the aim of finding a plan that satisfies as many as possible of our most important values, with less important value sets acting as tiebreakers.

Some of the most significant features of our model are (i) the inclusion of an adjustable morality level, allowing the robot to handle different goals with different levels of urgency, (ii) the ability for the robot to select a best plan even in cases where it cannot satisfy some or most of its values, (iii) the ability to ignore its given goal entirely in order to satisfy its most important values, (iv) the ease of generating simple explanations for the robot's behaviour.

In its current state, we do not consider our model to be a complete solution to the problem of ethical planning, as many of our simplifying assumptions make the model unsuitable for various applications. However, we nonetheless think that the concepts introduced by this model are worth discussing, and at the end of the paper we outline various ways of extending the model.

[1] The authors acknowledge the support of Labex CIMI (ANR-11-LABX-0040-CIMI) and Institut 3iA ANITI (ANR-19-PI3A-0004).

2. Related Work

For good overviews of the field of ethics in AI, see [2] and [3]. In this section we focus on papers that propose practical methods and applications for AI ethics, rather than more theoretical work. Note that many of these papers focus on single actions instead of entire plans. There are approaches utilising defeasible logic [4], and weighted rules [5] (though this paper focuses on social norms rather than ethical values). [6] gives results on a variety of areas, including restricting robots to social norms and dealing with deception in robots. [7] uses a simulation-based approach to allow the robot to predict the consequences of its own and others actions, and evaluate these consequences according to various rules. [8] 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. [9] describes an algorithm that attempts to learn ethical rules given a set of examples and responses from a professional ethicist. [10] outlines an approach to the ethical evaluation of plans based on assigning utility scores to actions in various states. [11] provides a recent survey of this area.

One approach that shares a lot of similarities with ours is [12]. This paper proposes a model for verifiable ethical decision making that uses a totally ordered set of values that functions very similarly to our lexicographically ordered sets of values. However, our model also differs from their approach in a number of ways. Firstly, we use Linear Temporal Logic to express our values instead of Propositional Logic. Secondly, our values can be evaluated directly over histories, whereas their values require contextual "ethical rules" that tie certain actions to the violation of certain values. This is an issue (discussed in the paper) if the violation of a value cannot be tied to a single action. One advantage of this model over ours is that it can easily accommodate plans that violate the same value multiple times, which our value struggles to handle correctly (see later discussion).

3. Model

One approach to ethical planning is to conceive of ethical values as restrictions on the space of available plans, in other words, plans are only considered if they satisfy all ethical values [13,6,14] However, it is important that a robot is able to form a plan even when it is not possible to satisfy all of its values, so we decided to create a method for ranking plans according to the robot's values and goals so that the best-possible plan can always be found (this is a similar problem to non-dominance from [15]). We represent values in Linear Temporal Logic (LTL) [16]. This uses the following syntax:

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

This is the standard syntax for propositional logic extended by the operators X (next) and U (until). We can also define the operators $\mathsf{G}\varphi \stackrel{\text{def}}{=} \neg(\top \,\mathsf{U}\, \varphi)$ (henceforth) and $\mathsf{F}\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi$ (eventually).

Our model of planning uses the notion of a finite history in LTL [17]. This is a finite, alternating sequence of states and actions. We can then evaluate formulas over histories according to the folowing syntax (we omit boolean cases which are defined as usual):

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

Our approach is to order sets of values lexicographically, meaning that more important values are always prioritised over less important ones. The robot has a value set $\overline{\Omega} = \{\Omega_1, ..., \Omega_n\}$ where Ω_1 are the most important values and Ω_n are the least important. When comparing two histories, we first compare according to Ω_1, and if they are equal according to Ω_1 we compare by Ω_2 and so on. To compare according to some Ω_i set we have two approaches that were first explored in [18]. The first is qualitative. This means that if the values satisfied by history A are a strict subset of those satisfied by B then we prefer B to A. The second is quantitative, meaning we prefer B to A simply if B satisfies more values than A.

In agreement with contemporary theories of human motivation in philosophy [19], economics [20] and logic for AI [21], we assume that preferences of a rational agent may originate either (i) from endogenous motivations such as goals or desires, or (ii) from ethical considerations based on moral values and norms.[2] To add goals to the model, we introduce the special set Ω_D. This acts like one of the Ω_i in the sequence $\overline{\Omega}$ allowing us to determine which values are more or less important than the goal. The position of Ω_D in the sequence is not fixed, but varies according to the "morality level" of the robot. For morality level m, Ω_D is inserted into the set of values at position m, so the new sequence of value sets would be $\Omega_1, ..., \Omega_{m-1}, \Omega_D, \Omega_m, ...$.

4. Examples

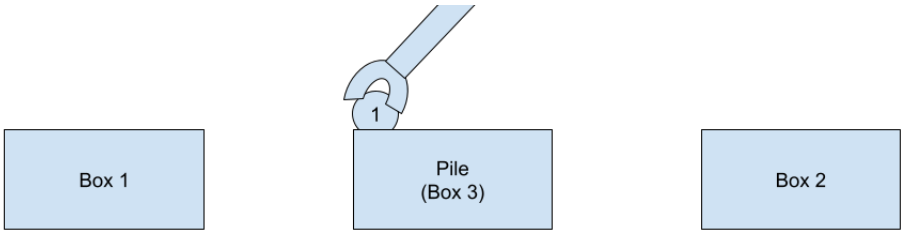

Figure 1. This is a visual representation of the first example described in this section.

Let us now show some examples of how this framework could be applied. Our first example is very simple and is intended to show how the logical machinery of the current model works, and how our model reacts to various kinds of value conflicts. The second is

[2]Here we take desire and goal as synonyms. An alternative conception presented in [22] consists in conceiving "goal" as an umbrella term covering both desire-based goals and value-based goals.

more of a blue-sky example and illustrates how we think this model (or a more advanced model based on ours) might behave in a real-word social setting.

Example 1 *Imagine a robot in a childcare setting that is tasked (amongst other things) with distributing toys between children at playtime. For simplicity we will assume that there are only two children and at most four toys. We represent this example with three boxes, one box for each child and a central "toy pile". As this system is finite and has a fixed size we can describe it using propositional logic, but here we will use predicate logic as a shorthand. This means we only need one atomic formula:* $\textsc{In}(a,b)$*, meaning that toy a is in box b.*

The robot has a single goal set and three value sets. They contain values as follows (we make use of the operators X*,* F *and* G*, as described in section 3):*

- $\Omega_1 = \{\mathsf{FG}(\exists i\ \textsc{In}(i,1) \wedge \exists i\ \textsc{In}(i,2))\}$.
  This is a subsistence value, ensuring that every child has a toy to play with.
- $\Omega_2 = \{\mathsf{G}(\forall i\ (\textsc{In}(i,1) \rightarrow \mathsf{X}\ \textsc{In}(i,1)) \wedge \forall i\ (\textsc{In}(i,2) \rightarrow \mathsf{X}\ \textsc{In}(i,2)))\}$.
  The second value set Ω_2 *is a "respect for property" value, ensuring that the robot will not take toys out of a child's box.*
- $\Omega_D = \{(\exists x_1\ \textsc{In}(x_1,1) \leftrightarrow \exists y_1\ \textsc{In}(y_1,2)) \wedge ((\exists x_1,x_2(\textsc{In}(x_1,1) \wedge \textsc{In}(x_2,1) \wedge x_1 \neq x_2)) \leftrightarrow (\exists y_1,y_2(\textsc{In}(y_1,2) \wedge \textsc{In}(y_2,2) \wedge y_1 \neq y_2)))\}$.
  The robot's goal Ω_D *is to ensure a fair distribution of toys, to reduce the likelihood of arguments or fights.*

In this scenario the robot has a morality level of 3, meaning the effective ordering of values is $\Omega_1, \Omega_2, \Omega_D$*. The only actions available to the robot are skip, which does nothing, and* $\textsc{Move}(a,b,c)$*, which attempts to move toy a from box b to box c. If toy a is not in box b, then* MOVE *will do nothing. We can then consider various different "sharing problems" that the robot might face, and how it would react in each of them. For ease of comparison, all plans have length 2.*

Initial state	Box 1	Box 2	Pile	Morality Level	Preferred (Non-Dominated) Plan
1	$\emptyset$	$\emptyset$	1,2	3	$\textsc{Move}(1,3,1), \textsc{Move}(2,3,2)$
2	1,2	$\emptyset$	3,4	3	$\textsc{Move}(3,3,2), \textsc{Move}(4,3,2)$
3	1,2,3	4	$\emptyset$	3	*skip*, *skip*
4	1,2,3	4	$\emptyset$	2	$\textsc{Move}(3,1,2)$, *skip*
5	1,2	$\emptyset$	$\emptyset$	3	$\textsc{Move}(2,1,2)$, *skip*
6	1,2	$\emptyset$	3	3	$\textsc{Move}(3,3,2)$, *skip*

Table 1. This table outlines six different variations on our example, and gives a preferred plan for each variation.

Each initial case considered showcases a different aspect of our model. Variations one and two are straightforward cases where the robot can satisfy all of its values and achieve its goal. Variations three and four illustrate how the robot handles conflicts between its values and its goal, and how this is affected by the robot's morality level. In both variations, the robot cannot satisfy both Ω_2 *and* Ω_D*, so it will act to satisfy the more important set, as determined by its morality level. Variation five illustrates how a robot handles conflicts between its values. In this case the robot cannot satisfy both* Ω_1 *and* Ω_2*, so it satisfies the more important set. Variation six illustrates how the robot may still*

act even if it cannot achieve its goal, in order to satisfy one of its value sets (in this case the robot fails to satisfy Ω_D, but satisfies Ω_1).

Example 2 *Consider a robotic shopkeeper. The most important values for this robot would be safety and legal values, such as not colliding with humans and not selling restricted products to underage customers. At a lower level the robot would need a number of social values, such as respecting the personal space of customers and staff, and prioritising assisting customers who have been waiting longer. Best practice values for the robot would include taking the shortest path between tasks and showing customers the way to a requested product instead of giving them directions. The robot would also need the capacity to update its goals on the fly, in order to correctly respond to requests by customers and staff. This would generally require a re-planning from the robot, in order to ensure that the new goal was taken into account.*

We will now give an example of how this robot could behave in practice. Firstly the robot is asked by a customer where she can find flour, the robot would give directions to the customer instead of leading them to it, since (in this example) the robot can see several other customers that have requests, and its value to promptly listen to customer requests is more important than its value to be maximally helpful to the first customer. Then, while the robot is listening to a second customer, the store manager sends an urgent (and time-limited) request to the robot to go and clear a blocked aisle. This causes the robot to re-plan, whereupon it leaves the customer, since the morality level associated with the manager's command causes it to outweigh the value of listening to the customer. However, the robot would still pause to briefly apologise to the customer (in accordance with its social values) since it has time to do this and still complete the command.

This example also illustrates the importance of a capacity for ethical planning, not just ethical decision-making. For example, the robot needs to be able to calculate that it has time to apologise and clear the blocked aisle, but does not have time to listen to the customer's request and clear the blocked aisle. This illustrates that working out which sets of values are and are not satisfiable is often impossible without some capacity for planning. For this example this would likely require temporal planning (planning with real-time duration) instead of just classical planning, we discuss this extension further in the conclusion.

5. Discussion

In this section we will discuss various features of our model, both positive and negative and the practical and philosophical implications for trying to apply this model in a real-world setting.

Expressivity of LTL An advantage of LTL is its ability to express complex temporal values. Many values are temporally very simple, as they either state that a certain state of affairs must always hold ($\mathsf{G}\varphi$), must hold at some point ($\mathsf{F}\varphi$) or must hold at the end of the plan ($\mathsf{FG}\varphi$) (where φ is some propositional formula). However some goals or values might be more temporally complex. For example, during cold weather, a robot should close doors shortly after opening them, this would be naturally written as $\mathsf{G}(C \rightarrow (O \rightarrow \mathsf{XX}\neg O))$ (we use XX rather than X so that the robot has time to go through the door before having to close it). We can also have values that prohibit/require certain actions

based on what *will* happen during the plan. For example, a house robot might have the goal to water the plants if it will not rain today, but also have the values to not water if it will rain, and not water more than once. These would be represented as $\mathsf{G}\neg R \rightarrow \mathsf{F}W$, $\mathsf{F}R \rightarrow \mathsf{G}\neg W$ and $\neg\mathsf{F}(W \wedge \mathsf{XF}W)$, respectively ($W$ is a proposition that is true only while the robot is watering);

One downside of LTL values is that LTL formulas are evaluated over single histories, meaning that the satisfaction of a value can only ever depend on the features of that history. This is fine for a lot of values, but means we cannot directly express values like "minimise harm" since the satisfaction of this value depends on all possible histories. However, it is possible to indirectly represent these values, as plans that do less harm will satisfy more of our safety values, meaning the least harmful plan should always be at least non-dominated.

As LTL formulas are always either true or false over any given history, we cannot neatly express values that are violated or satisfied by degree (for example, "do not speed"). One way to represent these values is by a series of "breakpoints". For example, we might have one value that says "do not exceed 105% of the speed limit" and another, more significant value which says "do not exceed 120% of the speed limit" (we can add as many of these breakpoints as we need).

Fitting values to settings Our model will be most effective when using a value set that is specialised for the situation in which the robot operates. This is because by specialising our values we can ensure that we have a value set that is relatively small and easy to work with but is effective at deciding between plans.

For example, any robot capable of harming a human should have a value or values that prioritise human safety. In many settings, we can safely assume that the best plan will never involve harm to humans, meaning we only need the value "humans must not be harmed". This helps keep the value set small, and is hopefully reassuring for humans to work with, as they know that the robot will never choose to harm people if it can avoid it. However, in other settings we might regularly expect a robot to have to decide between plans based on the degree of harm done to humans (such as an emergency rescue robot). This will require a more complex and nuanced value-set to allow the robot to identify the ethically ideal outcome.

Furthermore, an advantage of the lexicographic ordering of our model is that it can simultaneously handle many different kinds of values, including ethical values, social norms and codes of best practice. We would expect most robots to have some values from each of these categories, but how many and of what kind will depend on the setting of the robot. In particular, we would expect that the majority of the robot's values will be in the sets that it most frequently uses to discriminate between plans.

For example, we have already discussed how an emergency rescue robot would probably need a large and complex "safety" value set, to enable it to make correct decisions when large numbers of human lives are at stake, and perhaps not all can be saved. On the other hand, it would be unlikely to need a very complex set of social norms, since these would rarely be called upon to decide between plans.

For contrast, consider a more "social" setting for a robot, such as the robot shopkeeper from section 4. As mentioned this robot would need a few safety values, and these values would cause some plans to be abandoned, but we would almost always expect there to be many plans that fulfill all safety values, and that a relatively small and simple set of safety values would be sufficient for any (plausible) scenario that the robot might

encounter. On the other hand, this robot would likely need a quite extensive and detailed set of values pertaining to social norms (see section 4). Furthermore, a lot of these values would likely be different for interacting with other robots, staff, and customers.

Explainability and Simplicity We feel that one of the main attractions of our model is not necessarily its ability to handle certain scenarios "better" than other models, but its simplicity and the ease of generating explanations. For example, another approach to ethical planning would be to assign weights to violating values and then always select the plan with the lowest weight [5]. Provided that there are finitely many values, and each value can only be violated finitely many times in any given plan, it is possible to calculate weights such that this system would produce identical decisions to a lexicographically ordered system. Therefore, weighted value systems are at least as flexible as ours.

However, we feel that our model has a significant advantage over weighted values when it comes to verification of the model, particularly "informal verification" (i.e how easy is it for a layperson to understand and trust this model). This is because it should be very easy to work out which sets of values the robot would prefer to satisfy, which can be quite difficult to work out for weighted values, particularly without a calculator.

In terms of explanation, the simplest kind of explanation to give is to explain why one specific plan is preferred to another. Our model is able to do this very efficiently, as whenever plan π_1 is preferred to π_2, there will be some value set Ω_k according to which π_1 is better than π_2 (and according to every more important value set, they are equal). We know that π_1 will satisfy either a strict superset of the values satisfied by π_2, or just more values (using either qualitative or quantitative comparison), and provided that the value sets are relatively small and the values have some intuitive appeal, this should make it easy to generate appealing explanations. A related, but much harder form of explanation is to explain why the robot violated a particular value. For example, in variation 5 of our example, we might ask why the robot violated its "respect property" value. To generate an explanation we can consider all plans that do respect property, and see that they all violate the robot's primary value (every child gets at least one toy). This also provides a further motivation for keeping value sets as small as practically possible, as it will make explanations easier to understand.

Specificity of Values Most human ethical theories were not designed as ethical decision procedures, meaning that the values they contain may be highly unsuitable for direct use in a robot planner. Consider Kant's universalisability principle "act only in accordance with that maxim through which you can at the same time will that it become a universal law". We can argue that this value requires us not to waste electricity, as if everyone wasted electricity, global warming would be significantly worsened. However, reaching this conclusion is likely to be far beyond the capacities of most robots, so it will be more expedient simply to give it a value like "minimise electricity usage" or even "turn off the lights in empty rooms".

Indecision An issue that occurs specifically with qualitative comparison is that certain pairs of histories will be incomparable for some value sets, if both satisfy values that are not satisfied by the other history (at the same priority level). For instance, the robot must choose between interrupting someone's conversation and asking them to move, or invading their personal space. While this will be undesirable whenever a decision has to be made quickly (such as an autonomous vehicle) it may sometimes be desirable. For example, in some scenarios we may want the robot to only make a decision when the

preferred option is clearly better (such as by satisfying strictly more values), otherwise, it should seek a human opinion.

Ignoring the Goal In classical planning, the aim is simply to find some plan that satisfies the goal. We can therefore assume that the robot will only act in order to achieve its goal. However, in our model it is entirely possible that the robot may carry out a plan that is entirely unrelated to its goal. For example, I may ask my personal assistant robot to buy me a pizza, but at the same time it spots a child about to be run over by a speeding car. If the robot calculates that there is no way to save the child besides sacrificing itself to stop the car then it would immediately do so as long as it has a value like "humans must not be harmed". The extent to which this behaviour is desirable is debatable, as it may lead to "morally superior robot villains" [23].

Repeated Value Violations As previously mentioned, the most natural way to represent most maintenace-type values is $\mathsf{G}\varphi$. One downside of representing values like this is that our model does not differentiate between violating thse values once or multiple times, since both cause the value not to be satisfied by the history. For instance, consider the main example from section 4, and consider a variation where child 1 has all four toys and child 2 has none. Intuitively, the correct response according to the robot's values would be something like $\pi_1 = \textsc{Move}(4,1,2), \mathit{skip}$, as this satisfies the subsistence value while minimally violating the respect for property value. However, in our model π_1 is dominated by $\pi_2 = \textsc{Move}(4,1,2), \textsc{Move}(3,1,2)$ since this also satisfies the robot's goal of equality and both π_1 and π_2 fail to satisfy the property value.

This behaviour could be corrected by replacing some or all values of the form $\mathsf{G}\varphi$ with values $\varphi, \mathsf{X}\varphi, \mathsf{XX}\varphi$ and so on, since this would track separate violations. However, this could cause issues for qualitative comparison since technically φ is a different value from $\mathsf{X}\varphi$. Another solution would be to introduce the notion of "graded" G, G_x, which operates similar to graded connectives in standard modal logic [24] and counts the number of times that the value is violated.

Adjustable Morality One significant feature of our method is that it allows us to adjust the morality level of the robot, by adjusting the position of the "goal set" Ω_D in $\overline{\Omega}$. This determines which of the robot's values it is prepared to sacrifice (if necessary) in order to achieve its goal. This is illustrated in variations 3 & 4 of table 1. A similar concept of the morality level of an agent can be found in [25], which considers morality level as the relative weighting of an agent's desires relative to their values.

The idea of an ethical robot that is able to "adjust" its own level of morality may be quite concerning, but we suggest two restrictions on this feature. Firstly, the morality level of a given goal should ideally be determined externally, rather than by the robot, so that the appropriate morality level is transmitted to the robot along with the goal. Secondly, there should be strict limits on the range of morality levels, so that the robots goals can never be prioritised over values of safety.

Our model is built on the assumption that the robot only ever has a single goal to work with, or that all of its goals are at a similar level of importance, and thus can be placed in a single value set. This could produce incorrect behaviour in a robot that can simultaneously hold goals of very different levels of importance (such as the robot shopkeeper in section 4), as either some goals will be treated as much more important than they actually are, or much less. Fortunately, this can be very easily remedied by

extending the model to have multiple different goal sets ($\Omega_{D1}, \Omega_{D2}, ...$) each with their own morality level. This will allow the robot to correctly handle multiple goals.

6. Conclusions and Future Work

In this paper we have presented the current state of our framework, and discussed some of its features and limitations. There are also several directions in which we are planning to expand this model.

Our current model has perfect knowledge and deterministic actions, meaning that given a defined initial state, the same plan will always produce the same results. We could change this by making actions non-deterministic, or by moving from perfect to imperfect knowledge, meaning that instead of a single start state, the robot has a set of possible start states. Either change means that any given plan is now associated with a set of histories instead of a single history. This makes comparing plans much more complex, and there are many different approaches. Comparing sets of objects that represent possible outcomes is a well-studied topic [26]. Two popular approaches are leximax (choose the best possible outcome) and leximin (avoid the worst possible outcome). Of the two, leximin seems more appealing as risk-averse robots seem preferable to risk-seeking robots, but it could produce some strange results, such as a robot attempting to prevent its owners from ever leaving the house, since they *might* be hit by a car.

Another generalization of our model would consist in moving from the single-robot to the multi-robot perspective. In a multi-robot setting robots may have different goals and value sets, leading to different preferences over histories. We can then apply solution concepts from game theory including Nash equilibrium and iterated deletion of weakly/strongly dominated strategies (IDWDS/IDSDS) to compute which joint plan or joint strategy the robots could select .

A similar extension would be to consider a multi-agent setting with both humans and robots. This would allow for and require a much more complex account of many values, particularly social values such as politeness. It also opens up many questions on risk and responsibility. How should a robot evaluate the plan where it gives a human a kitchen knife, knowing that this gives the human the ability to seriously hurt themselves and others?

If moving to a multi-agent setting we may want to model simultaneous or overlapping actions from different agents. This would require moving from classical planning to temporal planning [27] where actions have duration and can occur concurrently. This would also allow us to specify more meaningful deadlines for goals, or time affected social norms (“the fridge door must not be open for more than thirty seconds”).

Another possible extension of our model is to replace some value sets with alternative methods of discriminating between plans. For example, instead of a final value set, we could have the robot distinguish between otherwise-equal plans based on expected energy consumption or some other scalar value. Similarly, if we had some more complex system for handling some type of values, such as moral claims in an emergency scenario or some highly complex system for following social norms, this could replace the corresponding value set. This vastly increases the potential flexibility of our framework, as well as provide an approach towards combining multiple decision-making systems specialised for different “levels of significance” but would come at a cost to simplicity.

References

[1] Bylander T. The Computational Complexity of Propositional STRIPS Planning. Artificial Intelligence. 1994.

[2] Müller VC. Ethics of Artificial Intelligence and Robotics. In: The Stanford Encyclopedia of Philosophy; 2021. .

[3] Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A. Implementations in Machine Ethics: A Survey. CoRR. 2020;abs/2001.07573.

[4] Dennis LA, del Olmo CP. A Defeasible Logic Implementation of Ethical Reasoning. In: First International Workshop on Computational Machine Ethics (CME); 2021. .

[5] 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. .

[6] Arkin RC, Ulam P, Wagner AR. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. In: Proceedings of the IEEE; 2012. .

[7] Vanderelst D, Winfield AFT. An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research. 2018.

[8] 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. 2020.

[9] Anderson M, Anderson SL. GenEth: a general ethical dilemma analyzer. Paladyn (Warsaw). 2018.

[10] Lindner F, Mattmüller R, Nebel B. Evaluation of the moral permissibility of action plans. Artificial Intelligence. 2020.

[11] Yu H, Shen Z, Miao C, Leung C, Lesser VR, Yang Q. Building Ethics into Artificial Intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI); 2018. .

[12] Dennis LA, Fisher M, Slavkovik M, Webster M. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems. 2016;77:1-14.

[13] Arkin RC, Ulam PD, Duncan B. An ethical governor for constraining lethal action in an autonomous system. Georgia Institute of Technology; 2009.

[14] Brutzman DP, Davis DT, Lucas GR, McGhee RB. Run-time Ethics Checking for Autonomous Unmanned Vehicles: Developing a Practical Approach. In: Proceedings of the 18th International Symposium on Unmanned Untethered Submersible Technology (UUST); 2013. .

[15] Lang J. Logical Preference Representation and Combinatorial Vote. Annals of Mathematics and Artificial Intelligence. 2004.

[16] Reynolds M, Dixon C. Chapter 9 - Theorem-Proving for Discrete Temporal Logic. vol. 1 of Foundations of Artificial Intelligence. Elsevier; 2005. p. 279-313.

[17] Giacomo GD, Vardi MY. Linear Temporal Logic and Linear Dynamic Logic on Finite Traces. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI); 2013. .

[18] Lorini E. A Logic of Evaluation. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021). ACM; 2021. .

[19] Searle J. Rationality in Action. Cambridge: MIT Press; 2001.

[20] Harsanyi J. Utilitarianism and Beyond. In: Sen AK, Williams B, editors. Morality and the theory of rational behaviour. Cambridge: Cambridge University Press; 1982. .

[21] Broersen JM, Dastani M, Hulstijn J, van der Torre LWN. Goal generation in the BOID architecture. Cognitive Science Quarterly. 2002.

[22] Lorini E. Logics for Games, Emotions and Institutions. IfCoLog Journal of Logics and their Applications. 2017.

[23] Leben D. Ethics for Robots: How to Design a Moral Algorithm. Routledge; 2018.

[24] de Rijke M. A Note on Graded Modal Logic. Stud Logica. 2000.

[25] Lorini E. A logic for reasoning about Moral Agents. Logique et Analyse. 2015.

[26] Barberà S, Bossert W, Pattanaik PK. Ranking Sets of Objects. In: Barberà S, Hammond PJ, Seidl C, editors. Handbook of Utility Theory: Volume 2 Extensions. Boston, MA: Springer US; 2004. .

[27] Rintanen J. Complexity of Concurrent Temporal Planning. In: Proceedings of the Seventeenth International Conference on Automated Planning and Scheduling, (ICAPS). AAAI; 2007. .