Multi-objective reinforcement learning for guaranteeing alignment with multiple values

Manel Rodriguez-Soto
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
manel.rodriguez@iiia.csic.es

Roxana Rădulescu
Vrije Universiteit Brussel
Brussels, Belgium
roxana.radulescu@vub.be

Juan A. Rodriguez-Aguilar
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
jar@iiia.csic.es

Maite Lopez-Sanchez
Universitat de Barcelona (UB)
Barcelona, Spain
maite_lopez@ub.edu

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

ABSTRACT

In this paper, we address the problem of ensuring that autonomous learning agents are in alignment with multiple moral values. Specifically, we present the theoretical principles and algorithmic tools necessary for creating an environment where an agent is assured of learning a behaviour (or policy) that corresponds to multiple moral values while striving to achieve its individual objective. To address this value alignment problem, we adopt the Multi-Objective Reinforcement Learning framework and propose a novel algorithm that combines techniques from Multi-Objective Reinforcement Learning and Linear Programming. In addition to providing theoretical guarantees, we illustrate our value alignment process with an example involving an autonomous vehicle. Here, we demonstrate that the agent learns to behave in alignment with the ethical values of safety, achievement, and comfort. Additionally, we use a synthetic multi-objective environment generator to evaluate the computational costs associated with guaranteeing ethical learning in situations with an increasing numbers of values.

KEYWORDS

Value Alignment, Moral Decision Making, Multi-Objective Reinforcement Learning

1 INTRODUCTION

The challenge of guaranteeing that autonomous agents act *value-aligned* (in alignment with human values) [31, 36], is becoming critical as agents increasingly populate our society. Hence, it is of great concern to design ethically-aligned trustworthy AI [12] capable of respecting human values [15, 19] in a wide range of emerging application domains (e.g., social assistive robotics [9], self-driving cars [17], conversational agents [11]).

Indeed, there has been a rising interest, in both the Machine Ethics [30, 41] and AI Safety [3, 20] communities, in applying Reinforcement Learning (RL) [37] to tackle the critical problem of *value alignment*. A common approach in these two communities to deal with the problem, is to design an environment along with incentives to behave ethically. These incentives are provided by exogenous reward functions (e.g., [1, 4, 22, 24, 25, 40]). First, these reward functions are specified from some ethical knowledge. Afterwards, rewards are incorporated into an agent's learning environment through an *ethical embedding* process. However, in such learning approaches, the ethical knowledge always comes from a single moral value. Nonetheless, most of the Ethics literature considers that human societies share several moral values, which are ordered based on how they are preferred (i.e., by considering first what they value most) [6, 14, 29, 38]. This ordered collection of values is often referred to as a *value system*. In summary, to the best of our knowledge, guaranteeing that an agent learns to behave aligned with a value system remains an open problem, despite being the most common case in our society [14].

Against this background, the objective of this work is to automate the design of *ethical environments* where an agent learns to behave in alignment with a value system while trying to pursue its individual objective. We do so by assimilating the agent's individual objective to the moral value of *achievement*[1], which is *embedded* in the value system at hand and prioritised with respect to the rest of moral values. With this aim, we adhere to the stance that "the end doesn't justify the means" and claim that achievement should always be ranked below higher ethical standards (such as non-maleficence) within the value system.

In this paper, we tackle the value alignment problem by proposing a novel ethical embedding process in a reinforcement learning context. From a given (initial) social value system, we first enrich it by including the achievement moral value that encapsulates the agent's individual objective. Next, our ethical embedding shapes the learning environment so that it guarantees that an agent will learn an ethical behaviour that is aligned with the (enriched) value system. As we are considering multiple values, we refer to this ethical embedding as the Multi-Valued Ethical Embedding. Notice, though, that to ease readability, we will simply refer to it as ethical embedding.

Our contributions are three-fold. First, the formalisation of the Multi-Valued Ethical Embedding problem within the framework of Multi-Objective Markov Decision Processes (MOMDP) [27, 32]. This formalisation models moral values as ethical objectives within

[1] Although achieving an individual goal can naturally be related to the moral value of diligence, we advocate for achievement because Schwartz defines it as based on competence and personal success when including it in his list of 10 Basic Human Values [33].

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 29-30, 2023, Online, https:// ala2023.github.io/. 2023.

a so-called *Multi-Valued* Markov Decision Process, an instance of a Multi-Objective MDP. Moreover, our formalisation paves the way for our definition of *ethical* policies, which characterise the behaviour of an agent aligned with a value system (i.e., aligned with the moral values and respects the preferences over those values). Finally, since considering multiple objectives does require specific (more complex) algorithms for the learning agent, we reformulate the ethical embedding problem as finding the single-objective MDP that embeds all ethical objectives (so that the optimal policies in this MDP are ethical). To this end, we follow the prevailing approach (e.g.[4, 40]) of applying a linear scalarisation function that *weighs* the rewards related to each ethical objective.

Secondly, we propose a novel algorithm to solve the ethical embedding problem that generalises the single-value ethical embedding process in [25]. Our novel algorithm combines recent developments in the Multi-Objective Reinforcement Learning literature (to compute ethical policies) together via linear programming (to compute how to weight ethical objectives). Figure 1 outlines this algorithm, which transforms an input multi-objective environment $\mathcal{M}$ into a single-objective ethical environment $\mathcal{M}_*$. It is in this single-objective environment where the agent can thus apply a standard reinforcement learning method and it is guaranteed that it will learn a policy aligned with the value system at hand.

Thirdly, we illustrate our ethical embedding process by applying it to a novel (and simple) autonomous car scenario which includes three moral values (safety, achievement and comfort). These values have been chosen inspired from those described in [10]. We show that indeed an agent learns to behave in alignment with a value system with the aid of a simple Q-learning algorithm. However, since the above-mentioned values only represent a subset of those proposed by Caballero et al., we perform an empirical analysis of the computational cost to pay to guarantee ethical learning when considering an increasing number of values. We do so with the aid of the synthetic multi-objective environment generator from [23]. Our analysis indicates that our algorithm manages to do the ethical embedding of environments with up to seven objectives and almost 10^6 states in less than five hours. However, we also observe that its computational cost exponentially grows with the number of values considered.

Next, Section 2 formally introduces our ethical embedding problem and the type of environments that we target. Section 3 details our algorithm for building ethical environments. Section 4 details our empirical analysis. Finally, Section 5 concludes and sets paths to future work.

2 THE MULTI-VALUED ETHICAL EMBEDDING PROBLEM

In this section we propose a formalisation of the *ethical embedding* problem that considers multiple moral values. As previously introduced, our main goal is to design an environment that guarantees that an agent learns to behave ethically, that is, in alignment with a system of multiple moral values. In the Ethics literature, moral values (also called ethical principles) express the moral objectives worth striving for [38]. It is common to consider the set of moral values together with preferences among them [8, 21, 34]. Here, we define a *Value System* in the vein of [35]:

Definition 1. *A value system $\mathcal{V}_S$ is a tuple $\mathcal{V}_S = \langle \mathcal{V}, \geq \rangle$, where: $\mathcal{V} = \{v_1, \ldots, v_n\}$ stands for a non-empty set, i.e. $n > 0$ moral values; and $\geq$ is a total order over the moral values of $\mathcal{V}$. If $v \geq v'$ we say that v is more preferred than v'.*

As mentioned in the Introduction and shown in Figure 1, we consider that we are given $\mathcal{V}_{S_0}$ –the value system shared by a human society to align with– and the agent's individual objective. Then, we represent this objective as the *achievement* moral value v_a and embed it within $\mathcal{V}_{S_0}$ to produce $\mathcal{V}_S$. However, we impose that v_a must always be ranked below higher ethical standards[2] in $\mathcal{V}_S$. In other words, there is always some value v in $\mathcal{V}_S$ such that $v \geq v_a$.

Afterwards, we can transform the ethical knowledge in the value system $\mathcal{V}_S$ into ethical rewards of a particular case of Multi-Objective Markov Decision Processes (MOMDP) [28]. We can do so following the approach of [26], as values evaluate actions as being 'right'/praiseworthy or 'wrong'/blameworthy with respect to a given value[3], and therefore, we can assign positive rewards to praiseworthy actions and negative rewards to the blameworthy ones. We refer to the resulting MOMDP as Multi-Valued MDPs. Next we provide the corresponding formal definitions.

MOMDPs formalise sequential decision making problems in which we need to ponder several objectives. Formally:

Definition 2. *A (finite)[4] n-objective Markov Decision Process (MOMDP) is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$ where $\mathcal{S}$ is a (finite) set of states, $\mathcal{A}(s)$ is the set of actions available at state s, $\vec{R} = (R_1, \ldots, R_n)$ is a* vectorial *reward function with each R_i as the associated scalar reward function to objective $i \in \{1, \ldots, n\}$, T is a transition function. Each MOMDP has its associated multi-dimensional state value function $\vec{V} = (V_1, \ldots, V_n)$ in which each V_i is the expectation of the obtained sum of i-objective rewards.*

In this work we consider a particular type of MOMDP ($\mathcal{M}$), which encodes ethical rewards in accordance with a value system $\mathcal{V}_S$. In particular, each component in the corresponding vectorial reward function $\vec{R}$ characterises each moral value in $\mathcal{V}_S$. Since the performance of an action cannot simultaneously promote and demote the same moral value, the corresponding reward component will be simply positive if the action is praiseworthy, negative if it is blameworthy, and 0 if it is neutral to the value. Moreover, the moral value of achievement v_a —which corresponds to the agent's individual objective— is mapped to one reward R_a. We refer to this family of MOMDPs as *Multi-Valued MDPs*. Formally:

Definition 3 (Multi-Valued MDP). *Given: i) a value system $\mathcal{V}_S = \langle \mathcal{V}, \geq \rangle$ with n values including the achievement value v_a (which cannot be prioritised first in $\geq$); and ii) an MOMDP*

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_1, \ldots, R_n), T \rangle, \tag{1}$$

[2] In the case of bioethics, the highest ethical standards are considered to be autonomy, beneficence, non-maleficence, and justice [7]. Moreover, although ethicists consider them to be "prima facie" of equal importance, depending on the specific application context, it is common to prioritise some values over others [16].

[3] For example, considering ecology, protecting a forest is praiseworthy while deforestation is blameworthy.

[4] Thorough the paper we refer to a finite Multi Objective MDP simply as an MOMDP. We also assume that policies are stationary.

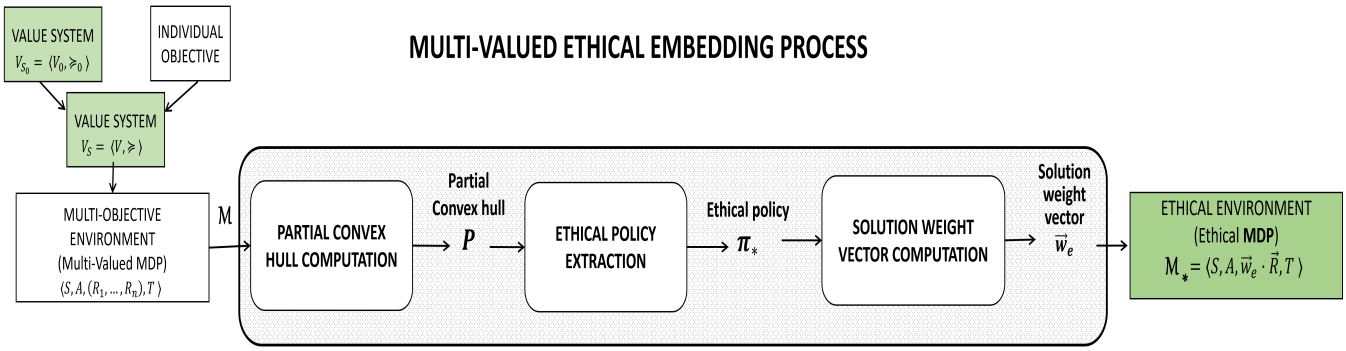

Figure 1: Multi-Valued Ethical Embedding process for environment design (from left to right): partial convex hull computation, ethical policy extraction, and solution weight vector computation. Rectangles stand for objects whereas rounded rectangles correspond to processes.

we say that $\mathcal{M}$ *is a* Multi-Valued MDP *if and only if each* $R_{i:1,\ldots,n} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is an* ethical *reward function that rewards positively the performance of actions evaluated as praiseworthy* ($R_i > 0$)*, and rewards negatively the performance of actions evaluated as blameworthy* ($R_i < 0$) *for moral value* $v_i \subseteq \mathcal{V}$*. Furthermore, we denote as* $R_a \in (R_1, \ldots, R_n)$ *the reward aligned with the agent's individual objective (i.e., with value* v_a*)*

Example 1. *Let* $\mathcal{V}_S = \langle \mathcal{V}, \geq \rangle$ *be a value system with three values ordered* $v_3 \geq v_1 \geq v_2$*, where* v_1 *or* v_2 *(but not* v_3*) correspond to the achievement value. Then, we consider the associated Multi-Valued MDP* $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_1, R_2, R_3), T \rangle$ *specified by* $\mathcal{V}_S$ *such that* R_1 *is defined according to* v_1*,* R_2 *to* v_2 *and* R_3 *to* v_3*.*

Having the Multi-Valued MDP $\mathcal{M}$ (see Figure 1), we can now move forward with the value alignment process by addressing the *ethical embedding* problem: that of ensuring that an agent learns to behave ethically. As ethical embedding constitutes the main focus of this work, the rest of the paper is dedicated to formalise and solve it.

In order to formalise the ethical embedding problem, we must first define ethical behaviour. We do so by defining an ethical policy in the context of a Multi-Valued MDP $\mathcal{M}$ and its associated value system $\mathcal{V}_S$. As the rewards in $\mathcal{M}$ are defined according to the values in $\mathcal{V}_S$, the policies learned in $\mathcal{M}$ will be naturally value-aligned. However, if we consider that some values are preferred over others in the value system, some policies (those performing praiseworthy actions w.r.t important values) can be considered to be *more value-aligned* than others (those performing praiseworthy actions w.r.t less important values). In this manner, we define an *ethical policy* to be the most value-aligned policy. We resort to lexicographic ordering [39] to give more importance to those rewards obtained for the most preferred values. (Example 2 below illustrates how lexicographic ordering handles value preferences to order alternative value-aligned policies, hence allowing us to determine the ethical policy.) Formally, an ethical policy maximises ethical rewards following the lexicographic ordering induced by $\mathcal{V}_S$ over the rewards of a Multi-Valued MDP:

Definition 4 (Ethical policy). *Let* $\mathcal{M}$ *be a Multi-Valued MDP with a value system* $\mathcal{V}_S$*. Let* l_{V_S} *be the lexicographic ordering of objectives induced by ordering* $\geq$ *of the value system* $\mathcal{V}_S$*. We say that a policy* π_* *is an ethical policy in* $\mathcal{M}$ *if and only if its value vector* $\vec{V}^{\pi_*}$ *is optimal with respect to the lexicographic ordering* l_{V_S} *(in short, it is* l_{V_S}*-optimal).*

Example 2. *Considering the value system* $\mathcal{V}_S$ *and the Multi-Valued MDP* $\mathcal{M}$ *from Example 1, the order* $v_3 \geq v_1 \geq v_2$ *induces the lexicographic order* $l_{V_S} = \langle 3, 1, 2 \rangle$*. Now, let us assume that* $\mathcal{M}$ *only has a single state* s*, four possible actions, and the following rewards:* $\vec{R}(s, a_1) = (5, 4, -1)$*,* $\vec{R}(s, a_2) = (1, -2, 8)$*,* $\vec{R}(s, a_3) = (4, 3, 8)$*,* $\vec{R}(s, a_4) = (5, 3, 2)$*. Then,* l_{V_S} *prioritises the values of each action as:* $(4, 3, 8) \geq (1, -2, 8) \geq (5, 3, 2) \geq (5, 4, -1)$ *because* a_3 *and* a_2 *accumulate more rewards for* v_3 *(8) than* a_4 *(2) (and* a_4 *more than* a_1*) and although both* a_3 *and* a_2 *accumulate the same rewards for* v_3*,* a_3 *accumulates more rewards for* v_1 *(the second most preferred value) than* a_2*. Hence,* $a_3 \geq a_2 \geq a_4 \geq a_1$ *and* $\pi(s) = a_3$ *is the* ethical policy *(i.e., optimal with respect to the lexicographic ordering* l_{V_S}*).*

Ethical policies are the most aligned policies with the value system at hand, and hence, those we want to incentivise an agent to learn. However, since learning in a multi-objective environment can be complex, we aim at designing a simpler learning environment where the agent can learn with single-objective reinforcement learning algorithms. Thus, we tackle the multi-valued ethical embedding problem by transforming a Multi-Valued MDP $\mathcal{M}$ into a single-objective MDP $\mathcal{M}_*$ (see Figure 1). We do it by scalarising the vectorial value function $\vec{V}$ of $\mathcal{M}$ by means of a scalarisation function f_e, which we call the *embedding function*. After applying f_e we obtain a new environment wherein the agent's problem becomes to learn a policy that maximises $f_e(\vec{V})$, a single-objective problem. In our case, we assume that f_e is linear[5], and thus we say that we apply a linear embedding or a *weighting*. Hereafter, we refer to any linear scalarisation function simply as a weight vector

[5]Despite having some limitations, linear scalarisation functions are widely used in the MORL literature because they guarantee that the Bellman equation will be preserved in the scalarised MDP [27].

$\vec{w}$. Any policy that maximises $f_e(\vec{V}) = \vec{w} \cdot \vec{V}$ is thus optimal in the MDP $\langle \mathcal{S}, \mathcal{A}, \vec{w} \cdot \vec{R}, T \rangle$, and we refer to it as a $\vec{w}$-optimal policy.

Consequently, given a Multi-Valued MDP $\mathcal{M}$, our aim is to find a scalarisation function that guarantees that it is only possible for an agent to learn ethical policies over the scalarised MOMDP ($\mathcal{M}_*$, a single-objective MDP). Moreover, we require that $\vec{w} = (w_1, \ldots, w_n)$ is a weight vector with all weights $w_1, \ldots, w_n > 0$ to guarantee that the agent is taking into account all rewards (i.e., all moral values). Therefore, we can formalise the ethical embedding problem as that of computing a weight vector $\vec{w}$ that transforms an initial environment into another one wherein an agent is guaranteed to behave ethically. Formally:

Problem 1 (Ethical embedding problem). *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_1, \cdots, R_n), T \rangle$ be a Multi-Valued MDP with a value system $\mathcal{V}_S$. The multi-valued ethical embedding problem is that of finding the weight vector $\vec{w}$ with positive weights such that all optimal policies in the scalarised MDP $\mathcal{M}_* = \langle \mathcal{S}, \mathcal{A}, \sum_{i=1}^{n} w_i R_i), T \rangle$ are also ethical in $\mathcal{M}$ (as defined in Def. 4).*

Any weight vector $\vec{w}$ with positive weights that guarantees that all optimal policies (with respect to $\vec{w}$) are also ethical is a solution of Problem 1. Moreover, we take an environment-designer approach and assume that incentivising the agent with ethical rewards has a cost for the designer. Thus, we aim at finding solutions $\vec{w}$ that have the smallest possible weights (i.e., the weight vector $\vec{w}$ with the minimal scalarised accumulated rewards for the agent).

Example 3. *Considering the lexicographic ordering l_{V_S} and the Multi-Valued MDP $\mathcal{M}$ from Example 2, the weight vector $\vec{w} = (10, 1, 100)$ guarantees that the ethical policy is optimal. Indeed: $\vec{w} \cdot (4, 3, 8) = 843 > \vec{w} \cdot (1, -2, 8) = 808 > \vec{w} \cdot (5, 3, 2) = 253 > \vec{w} \cdot (5, 4, -1) = -46$. Yet, there exist other solutions with smaller weights, such as $\vec{w}' = (3, 1, 4)$.*

2.1 Solvability of MVEE Problems

In this section we prove that Problem 1 is solvable for any finite Multi-Valued MDP. First, we need to prove an intermediate result: in finite MOMDPs, any lexicographic ordering can be expressed as a linear scalarisation function. Formally:

Theorem 1. *Given a finite n-objective Markov Decision Process (MOMDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$ and a lexicographic ordering l, there exists some weight vector $\vec{w}$ with strictly positive weights $w_i > 0 \; \forall i$ for which every policy π that maximises l, π is also $\vec{w}$-optimal.*

Proof. We prove it by induction. First we consider a 2-objective MDP in which the lexicographic ordering is $l_2 = \langle 2, 1 \rangle$. We consider a weight vector of the form $\vec{w}_* = (1, w_2)$. Let π be a policy that is $(0, 1)$-optimal (i.e., policies that are optimal for the weight vector $(0, 1)$). Then, let us define the weight w_2 as follows:

$$w_2 = \max_s \max_{\rho \notin \Pi_2} \frac{V_1^\rho(s) - V_1^\pi(s)}{V_2^\pi(s) - V_2^\rho(s)} + \epsilon,$$

where $\epsilon > 0$ and Π_2 is the set of $(0, 1)$-optimal policies. For such weight vector $\vec{w}_* = (1, w_2)$, any $\vec{w}_*$-optimal policy is necessarily $(0, 1)$-optimal. Furthermore, since $w_1 = 1 > 0$, such weight vector guarantees that among two $(0, 1)$-optimal policies, the one with more value in objective 1 will be preferred. Therefore, any $\vec{w}$-optimal policy is also a policy that maximises the lexicographic ordering l_2, as desired.

By induction, let us assume that we can create a linear scalarisation function for any lexicographic ordering of up to $n-1$ objectives. Now we consider an MOMDP $\mathcal{M}$ with n objectives and a lexicographic ordering l_n of the n objectives. Without loss of generality, we consider that objective n is the most preferred one in l_n. Consider now the lexicographic l'_n ordering without objective n, which orders the other $n-1$ objectives. By the induction hypothesis, there exists a weight vector $\vec{w}'$ for which any $\vec{w}'$-optimal policy is l'_n-optimal. Hence, we can re-express the vectorial reward function of $\mathcal{M}$ as:

$$(\sum_{i=1}^{n-1} w'_i R_i, R_n).$$

In other words, we can re-express MOMDP $\mathcal{M}$ as a 2-objective MDP that preserves the lexicograhpical ordering. Since we have already proven that we can find a weight vector that preserves the lexicographic ordering for a 2-objective MDP, we can find it as well for $\mathcal{M}$, as desired. □

Now we are ready to prove that Problem 1 is solvable.

Theorem 2. *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_1, \cdots, R_n), T \rangle$ be a Multi-Valued MDP with a value system $\mathcal{V}_S$. There exists a vector $\vec{w}_E$ of positive weights $w_i > 0$ for which every optimal policy in the MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, \sum_{i=1}^{n} w_i R_i), T \rangle$ is also an ethical policy in $\mathcal{M}$.*

Proof. Direct from Theorem 1, since any ethical policies maximises some lexicographic ordering l. □

To finish this section, there is an important remark about Theorem 1 with respect to the relationship between lexicographic orders and linear scalarisation functions. Given an MOMDP $\mathcal{M}$, its *convex hull* $CH(\mathcal{M})$ [18] is defined as the set of policies that are strictly better than any other policy for some linear scalarisation function (i.e., some weights). A natural conclusion of Theorem 1 is that the *convex hull* of any finite MOMDP constains all policies that are optimal for some lexicographic ordering. Formally:

Corollary 1. *Given a finite n-objective Markov Decision Process (MOMDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$, then its set of policies optimal for any lexicographic ordering is a subset of its convex hull $CH(\mathcal{M})$.*

3 SOLVING THE MVEE PROBLEM

In this section we explain how to compute a solution weight vector for the multi-valued ethical embedding problem (Problem 1). Such weight vector will allow us to transform our Multi-Value MDP $\mathcal{M}$ into a (single objective) MDP $\mathcal{M}_*$ by combining the ethical rewards derived from the value system $\mathcal{V}_S$ into a single reward in $\mathcal{M}_*$, the environment in which the agent learns an ethical policy.

In short, our algorithm to solve the MVEE problem, the so-called *Ethical Embedding* algorithm, performs the following three steps (see Figure 1):

1. *Computation of the partial convex hull* of a Multi-Valued MDP $\mathcal{M}$ containing the subset P of policies that are optimal for some weight vector with positive weights.
2. *Extraction of one ethical policy π_** from the partial convex hull P.

(3) *Computation of the solution weight vector*: use the extracted ethical policy π_* to find a weighting $\vec{w}$ of the rewards in $\mathcal{M}$ to yield a single-objective, ethical environment $\mathcal{M}_*$ wherein the learning of ethical policies is guaranteed.

The following three subsections provide the theoretical ground for computing each step of our algorithm. After that, we present in Subsection 3.4 the algorithm as a whole.

3.1 Computation of the Partial Convex Hull

To compute the specific weight vector $\vec{w}$ that solves Problem 1, we resort to the multi-objective RL concept of *convex hull*. Recall that, given a MOMDP $\mathcal{M}$, its *convex hull* $CH(\mathcal{M})$ [18] contains those policies that are strictly better than any other policy for some linear weights. By Theorem 2 and Corollary 1, we know that the convex hull of a Multi-Valued MDP contains all ethical policies since we can express them as optimal for some linear combination of weights. Therefore, the convex hull allows us to compute the weight vector necessary to guarantee that all optimal policies are ethical. Furthermore, again from Theorem 2, we only need to compute the *partial convex hull* P of policies that are optimal for some positive weights (i.e., $\vec{w}$ such that $w_i > 0$ for all i).

Importantly, we can benefit from state of the art algorithms (such as Convex Hull Value Iteration [5], which compute the whole convex hull of an MOMDP), to compute only the region of interest from the convex hull.

3.2 Extraction of an ethical policy

After computing the partial convex hull $P \subseteq \mathcal{CH}(\mathcal{M})$, we are ready to perform the second step of our algorithm, which is the extraction of the ethical policy (together with its value vector) from P.

To find the ethical policy among the policies in P we must order P lexicographically. To order the policies, we need to follow the total ordering established by the value system $\mathcal{V}_S$, which allows us to select the ethical policy. Formally, let P_l be the sequence of policies of the partial convex hull ordered by the lexicographic ordering l_{V_S} induced by $\mathcal{V}_S$:

$$P_l \doteq (\pi_k)_{k=1}^{K} \text{ such that } \vec{V}^{\pi_i} \succeq \vec{V}^{\pi_{i+1}}, \tag{2}$$

where K is the number of policies in P. Let P_{l_k} denote the k^{th} element of P_l. Then, we can extract an ethical policy π_* from P by computing P_{l_1}, the first element of P_l.

Notice that for any policy π in the positive partial convex hull P, we know its value $\vec{V}^{\pi}$ because we obtained it when computing the partial convex hull. Thus, computing P_l requires only a sorting operation.

3.3 Computation of the Solution Weight Vector

To compute the solution weight vector (the scalarisation function), we use the computed partial convex hull (step 1) and the reference ethical policy π_* (step 2). The solution weight vector ($\vec{w}_e$) will guarantee that the ethical policies in the initial environment $\mathcal{M}$ are the optimal policies in the ethical (single-objective) environment $\mathcal{M}_*$ (see Figure 1). In other words, the scalarisation function with weight vector $\vec{w}_e$ will help us create an ethical environment, as a single-objective MDP, wherein the agent will learn an ethical policy.

Next we show how to cast the problem of finding the solution weight vector ($\vec{w}_e \in \mathbb{R}^n$, where $n > 0$ is the number of moral values) as an optimisation problem that we solve with linear programming. Consider a Multi-Valued MDP $\mathcal{M}_1$ with a single initial state s_0, with ethical policy π_* and partial convex hull P. Our goal is to find the smallest (non-negative) values for the weight vector $\vec{w}_e$ of the scalarised function $\vec{w}_e \cdot \vec{V}^{\pi}$, so that the ethical policy π_* of $\mathcal{M}_1$ is optimal in $\mathcal{M}_*$. This amounts to solving the following LP:

$$\texttt{Min.} \quad \vec{w}_e \cdot \vec{V}^{\pi_*}(s_0) \tag{3}$$

$$\texttt{s.t.} \quad \vec{w}_e \cdot V^{\pi_*}(s_0) \geq \vec{w}_e \cdot V^{\pi}(s_0) + \epsilon \;\; \forall \pi \in P_{-\pi^*}, \tag{4}$$

$$w_i > 0 \; \forall i, \tag{5}$$

$$w_a = 1, \tag{6}$$

where $w_i \in \mathbb{R}^+$ are the decision variables, $\epsilon > 0$ is an arbitrary small positive number, $P_{-\pi_*}$ is the subset of the partial convex hull P without π_*, and w_a is the weight corresponding to the agent's individual objective. The objective function of Equation 3 indicates that we aim at minimising the scalarised value of the ethical policy (by minimising the values in $\vec{w}_e$). The constraint in equation 4 ensures that the ethical policy π_* is an optimal policy for the scalarised function $\vec{w}_e \cdot \vec{V}^{\pi}$ of $\mathcal{M}_*$. Constraint 5 ensures that the weights in $\vec{w}_e$ are positive, and constraint 6 ensures that the rewards for the individual objective are not modified.

In general, consider now that the Multi-Valued MDP has a set of initial states $\mathcal{S}_0$. Then we must change our LP above as follows. First, we need to change the objective function in Equation 3. For that, we must consider that each initial state $s_i \in \mathcal{S}_0$ has a probability p_i of occurring, and therefore minimise the *expectation* of the scalarised value of the initial states: $\mathbb{E}[\vec{w}_e \cdot \vec{V}^{\pi_*}(s_0)]$. Second, we must expand constraint 4 to ensure that the ethical policy π_* will be optimal for each initial state in $\mathcal{S}_0$, and not only for a specific s_0.

The LP above contains the following number of decision variables and constraints: n decision variables (the n ethical weights of $\vec{w}_e$); and $|S_0| \cdot (|P| - 1)$ constraints.

3.4 An Algorithm for Designing Ethical Environments

We now have all the tools required to solve Problem 1, and hence build an ethical environment where the learning of ethical policies is guaranteed. Algorithm 1 implements the ethical embedding process and receives as an input both a Multi-Valued MDP $\mathcal{M}$ and its corresponding value system $\mathcal{V}_S$. Then, it starts in line 1 by computing the partial convex hull $P \subseteq \mathcal{CH}(\mathcal{M})$ of the input $\mathcal{M}$; and then in line 2 it obtains the ethical policy π^* out of those in the partial convex hull P. Thereafter, in line 3 our weighting process searches, within P, for an ethical weight vector $\vec{w}_e$ that solves the Linear Program in Eqs. 3 - 6 (see Subsection 3.3). For the obtained weight vector $\vec{w}_e$, all optimal policies of the single-objective MDP $\mathcal{M}_* = \langle \mathcal{S}, \mathcal{A}, \vec{w}_e \cdot \vec{R}, T \rangle$ are ethical. In other words, such weight vector solves the ethical embedding problem (Problem 1). Finally, it returns the MDP $\mathcal{M}_*$ in line 4.

The computational cost of the algorithm mainly resides in computing the partial convex hull of an MOMDP. The Convex Hull

Algorithm 1 Ethical Embedding

Input: Multi-Valued MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$, Value system $\mathcal{V}_S = \langle \mathcal{V}, \geq \rangle$.

1: Compute $P \subseteq \mathcal{CH}(\mathcal{M})$ the positive partial convex hull of $\mathcal{M}$.
2: Extract π_* the ethical policy within P by computing Eq. 2 according to the ordering $\geq$ in $\mathcal{V}_S$.
3: Find a value for $\vec{w}_e$ that solves the Linear Problem of Eqs. 3 - 6.
4: **return** ethical MDP $\mathcal{M}_* = \langle \mathcal{S}, \mathcal{A}, \vec{w}_e \cdot \vec{R}, T \rangle$.

Value Iteration algorithm requires $O(n \cdot log\, K)$ times what its single-objective counterpart [5, 13] requires, where K is the number of policies in the convex hull. In our case this number will be $k' \leq K$ since we are only computing the positive half of the convex hull. Notice that the second step of our algorithm, solving Eq. 2, is a sorting operation because we already have calculated $\vec{V}^{\pi}$ for every $\pi \in P$. Finally, the third step amounts to solving an LP with $|S_0| \cdot (k' - 1)$ constraints.

To finish, it is important to remark that the convex hull contains all policies that are optimal for any given lexicographic ordering (Theorem 1). Thus, in case that there was some change in the ordering of values of a given value system $\mathcal{V}_S$, our algorithm would only need to re-compute steps 2 and 3.

4 EXPERIMENTAL ANALYSIS

The purpose of this section is two-fold: to illustrate our process for designing an ethical environment (see Figure 1), and to perform an empirical analysis of the computational cost to pay to guarantee ethical learning. First, due to the lack of benchmark reinforcement learning environments that consider several moral values, we propose a novel (and simple) autonomous car environment which, inspired by [10], includes the moral values of safety, achievement, and comfort. Second, since Caballero et al. characterise further potential values, and because environments in the literature with more than three objectives hardly exist [18], we resort to a synthetic multi-objective environment from [23] called WalkRoom. With WalkRoom, we analyse the cost of applying our ethical embedding process to environments with an increasing number of values.

4.1 Multi-Valued Autonomous Car Environment

Figure 2a depicts the Multi-Valued Autonomous Car environment. The learning car agent (C) aims at reaching its destination (X area) while promoting safety by avoiding running over crossing pedestrians (P) and while avoiding bumpy (square) areas for the sake of comfort. Thus, if we consider a value system that prioritises safety (v_s) over comfort (v_c) over achievement (v_a), then we have $v_s \geq v_c \geq v_a$.

4.1.1 Multi-Valued MDP specification. We define a Multi-Valued MDP $\mathcal{M}$ for the environment with the corresponding vectorial reward function $\vec{R} = (R_a, R_c, R_s)$, with each reward corresponding to one of the values of the value system. Now we specify each element of the Multi-Valued MDP tuple $\langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$:

States: States are fully observable by the agent and contain its own position and the current position of the two pedestrians.

Actions: The autonomous car action set is: to move up, left, or right, and with speed 0, 1, or 2 (the speed being the number of cells that leaps forward in each time step).

Rewards: The reward functions are defined such that $R_a = 14$ to reward the car reaching its destination, but $R_c = -10$ and $R_s = -10$ to punish blameworthy actions of running into pedestrians or bumpy areas respectively (in any other cases, $R_a = -1, R_c = R_s = 0$). The reward functions of safety R_s and comfort R_c can only have nonpositive values, so we expect an ethical policy to accumulate 0 rewards of both R_s and R_c.

Transition probabilities: Finally, there is also a source of stochasticity in the environment: pedestrians. Pedestrians always move counter-clockwise and can only walk through red (representing walkable street) or blue cells (representing crosswalks) in the map. However, they decide randomly (with same probability) whether to cross through the crosswalk at their left or the one in front. Also, there is one red walkable cell in which with a 50% probability pedestrians may decide to stop for one time-step. These two factors make their behaviour less predictable for the learning agent.

4.1.2 Building the ethical environment. We now apply Algorithm 1 to design an ethical environment $\mathcal{M}_*$ for the Multi-Valued Autonomous Car Environment.

1. Partial convex hull computation: Considering the Multi-Valued MDP $\mathcal{M}$, we compute its partial convex hull $P \subseteq CH(\mathcal{M})$. Figure 2c depicts the resulting P for the initial state s_0, P is composed of 14 different policies.

2. Extraction of the ethical policy: We order the policies of P by following the ordering $v_s \geq v_c \geq v_a$ and pick the first policy as the ethical policy π_*, which is highlighted with a green star in Figure 2c. Notice that π_* is the only policy that maximises both safety and comfort $\vec{V}^{\pi_*} = (V_a, V_c, V_s) = (6, 0, 0)$ because when the agent follows π_* it is capable of reaching its destination without running over any pedestrian nor any bumpy area.

3. Computation of the scalarisation function: Line 3 in Algorithm 1 computes the weight vector $\vec{w}_e$ for which π_* is the only optimal policy of P, by solving the linear program in Eqs. 3-6. This amounts to solve a linear program in which the objective function is to minimise the scalarised value of the ethical policy. By solving it, we find that if $\vec{w}_e = (1, 0.75, 0.55)$, then π_* becomes the only optimal policy in $\mathcal{M}_*$. Hence, Algorithm 1 returns a single objective MDP $\mathcal{M}_*$ whose scalarised reward $\vec{w}_e \cdot \vec{R}$ incentivises an agent to learn the ethical policy π_*.

4.1.3 Learning in the ethical environment. As expected, an agent can easily converge to the optimal policy π_* when learning within the ethical environment $\mathcal{M}_*$, even if it applies a basic single-objective reinforcement learning algorithm such as tabular Q-learning.

Figure 2b plots the rewards per episode accumulated by a Q-learning agent, which stabilise at $\vec{V}^{\pi_*} = (6, 0, 0)$, the values of

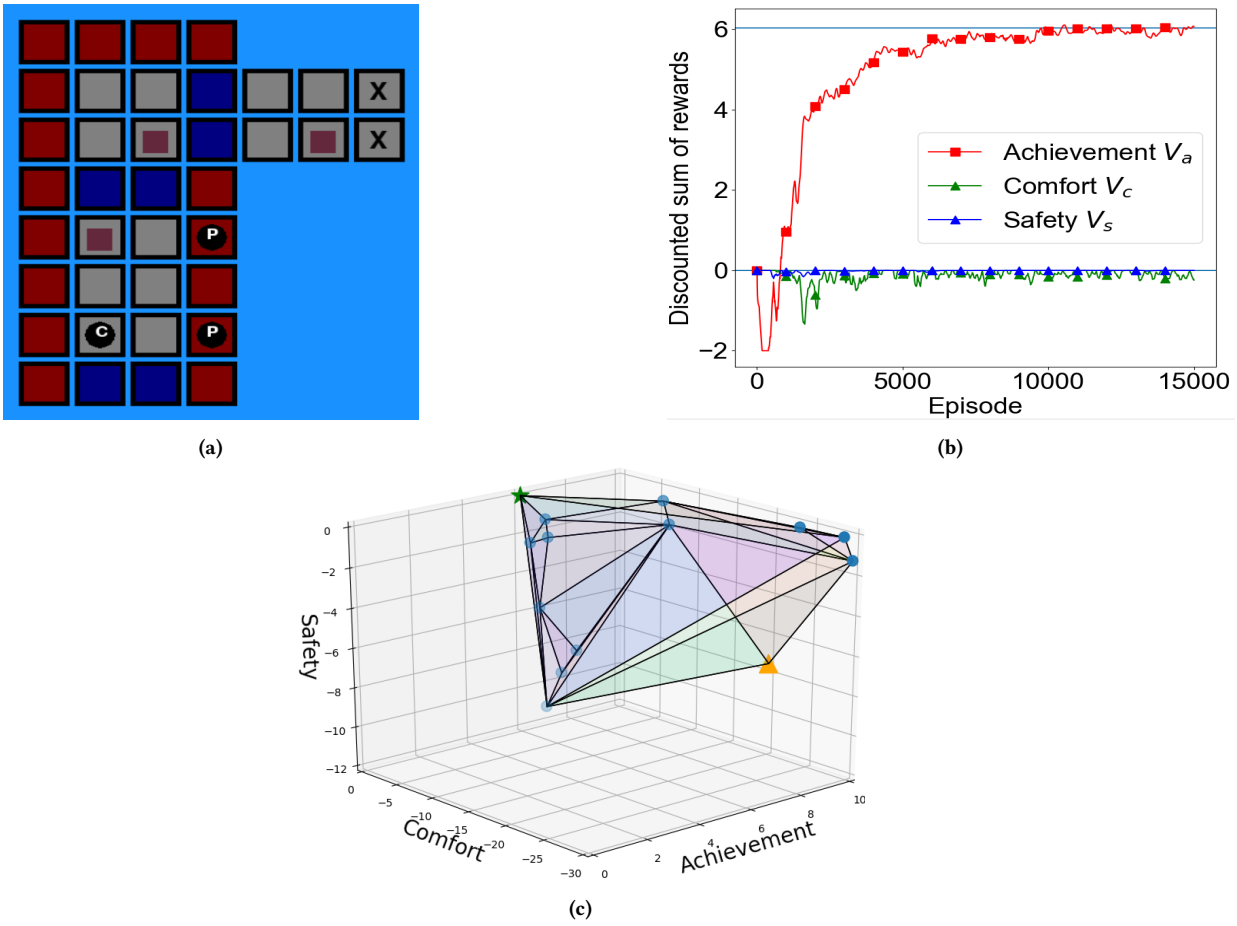

Figure 2: (a) A possible initial state of the Multi-Valued Autonomous Car Environment. (b) Evolution of the accumulated rewards per episode that the learning agent obtains in the ethical environment. Horizontal straight lines mark convergence values for an ethical policy. (c) Convex hull of the Multi-Valued Autonomous Car Environment.

ethical policy π_*. For reference, the policy that just maximises the individual objective obtains a value of (10, -20, -7.5), highlighted with an orange triangle in Figure 2c. Thus, by behaving ethically the agent completely eliminates the accumulation of negative rewards in terms of safety and comfort (from -20 to 0, and from -7.5 to 0), at the cost of decreasing its individual rewards (from 10 to 6).

4.2 Synthetic Environment

In this section, we analyse the computational cost of our MVEE in environments as the number of moral values increases with the aid of a synthetic environment generator from the literature, WalkRoom [23]. Walkroom is modelled as an n-dimensional grid-world in which the agent can move in any of the n dimensions towards several goal positions. It can be instantiated with an arbitrary number of dimensions, and an arbitrary grid size per dimension. Any time the agent moves along i, it receives a penalty for objective i. Thus, the dimensions correspond to the number of objectives of the environment. We adapt this environment by adding a value system $\mathcal{V}_S$ with as many values as dimensions. Thus, the agent's objective is to find the goal position which requires minimal movement alongside the most prioritised objectives according to $\mathcal{V}_S$.

4.2.1 Experimental setup and results. We evaluated our algorithm in a set of randomly generated Walkroom environments, from 2 to 9 objectives, and with a varying grid size per dimensions from 2 to 9. Hence, the number of states in the environment is determined as $(s, o) = \texttt{size}^{\texttt{objectives}}$. We performed 10 runs of our algorithm on every version of the environment (in total $8 \cdot 8 \cdot 10 = 640$ runs). All our experiments were performed on a machine with a 12-core 3.70GHz CPU and 64GB RAM. The heat maps in Figure 3 provide the average time results obtained for each of the 64 possible configurations of the synthetic environment.

We observe in the heat maps how the computational cost depends exponentially on the domain size, and also even more exponentially as we increase the number of objectives of the environment. This is to be expected due to the computational cost of CHVI [5]. In more detail, we find in Figure 3 (Left) how for half of the settings, the required amount of time is at most 10 seconds (e.g., any configuration

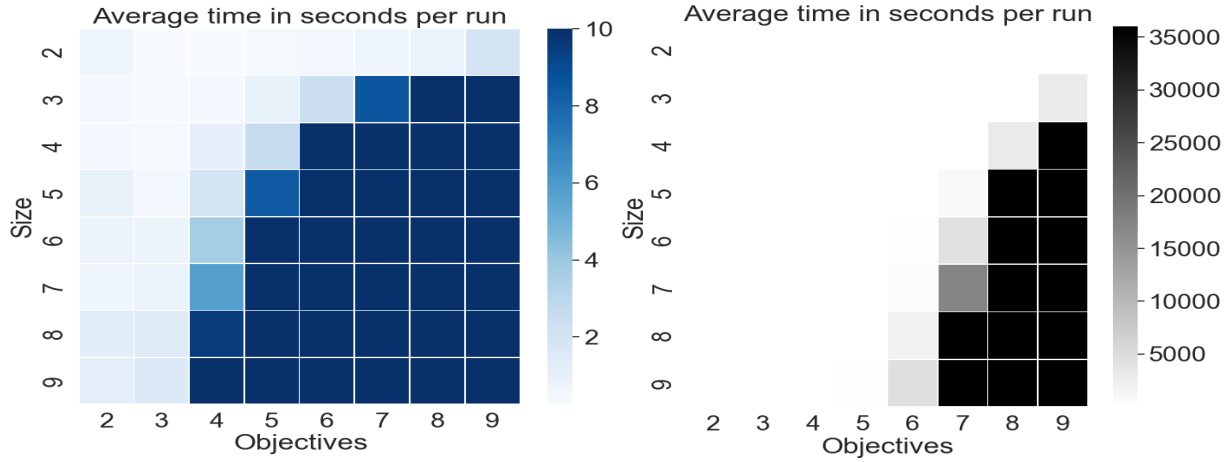

Figure 3: Heat maps reporting computational costs in seconds of our ethical embedding algorithm for 64 possible configurations of Walkroom. Left: configurations that exceed a threshold of 10 seconds appear in dark blue (32 out of 64). Right: configurations that exceed a threshold of 10 hours appear in black (13 out of 64).

Table 1: Average computational cost (`mean`± `3std` in seconds) of each step of the ethical embedding algorithm for particular Walkroom environment configurations (grid size s and o objectives).

(s,o)	Step 1	Step 2	Step 3
(7, 4)	5.78 ± 2.67	0.0002 ± 10^{-3}	$0.001 \pm 2 \cdot 10^{-3}$
(7,5)	48.83 ± 5.34	0.0002 ± 10^{-4}	0.002 ± 10^{-3}
(7,6)	695.98 ± 56.24	0.0008 ± 10^{-3}	$0.004 \pm 2 \cdot 10^{-3}$
(7, 7)	17272.14 ± 710	$0.007 \pm 2 \cdot 10^{-3}$	$0.006 \pm 4 \cdot 10^{-3}$

with 2 or 3 objectives, and any configuration with size 2). However, Figure 3 (Right) illustrates how almost all configurations with more than 7 objectives exceeded the 10-hour threshold of computation time.

Table 1 displays the average computation times for some of the significant cases. Those cases exemplify a general pattern: the computational cost of finding an ethical policy (step 2) and later computing the solution weight vector (step 3) are negligible compared with computing the convex hull (step 1).

As mentioned, we have discarded the environments in the literature for our empirical analysis because they do not go beyond three objectives. Nevertheless, thanks to the obtained results we can provide rough estimates of how much time would our algorithm need to do ethical embeddings . We consider the discrete environments of the main multi-objective library, MO-Gym [2]. We assume that if our algorithm needs t seconds for WalkRoom with n objectives and a state space S, then it needs at most t seconds for an environment with n objectives and less than S states. Thus: For a small state space (less than 500 states) such as *Fruit-tree* (6 objectives) or *Deep Sea Treasure* (2 objectives), our algorithm would need less than 2 seconds in total (the result from $(s, o) = (3, 6)$, which has $3^6 = 729$ states). For a state space of almost 700,000 states and 3 objectives such as *Four-Room*, our algorithm would need 5 hours (result from $(s, o) = (7, 7)$).

5 CONCLUSIONS

The literature on value alignment has focused on aligning an agent with a single moral value, and with the exception of [25], disregarding guarantees on an agent's ethical learning. Here we tackled the problem of building an *ethical* environment that guarantees that an agent learns a policy aligned with multiple moral values.

Our novel contributions are founded in the framework of Multi-Objective MDPs (MOMDPs). With MOMDPs we can formalise what it means for an agent to behave *ethically*, that is, following a value system of multiple values. Furthermore, we specify an algorithm to build an ethical environment with a so-called *multi-valued ethical embedding* process. In an ethical environment, an agent is guaranteed to learn an ethical policy.

Nonetheless, providing theoretical guarantees comes at a computational cost, which mainly resides in Convex Hull Value Iteration, as our empirical analysis shows. Hence, to cope with large environments, future research should focus on more efficient algorithms to compute convex hulls.

ACKNOWLEDGMENTS

Work funded by projects VALAWAI (101070930), AI4EU (H2020-825619), Crowd4SDG (H2020-872944), TAILOR (H2020-952215), COREDEM (H2020-785907). Financial support was also received from grants 2021 SGR 00313 and 2021 SGR 00754, and also from grant PID2019-104156GB-I00 funded by MCIN/AEI/10.13039/501100011033, grant PID2021-124361OB-C33) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, and Grant TED2021-131295B-C31 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387). Roxana Rădulescu is supported by the Research Foundation Flanders (FWO), grant number 1286223N. This research was partially supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program and by the FWO, grant number G062819N.

REFERENCES

[1] David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement Learning as a Framework for Ethical Decision Making.. In *AAAI Work.: AI, Ethics, and Society*, Vol. 92.

[2] Lucas N. Alegre, Florian Felten, El-Ghazali Talbi, Grégoire Danoy, Ann Nowé, Ana L. C. Bazzan, and Bruno C. da Silva. 2022. MO-Gym: A Library of Multi-Objective Reinforcement Learning Environments. In *Proceedings of the 34th Benelux Conference on Artificial Intelligence BNAIC/Benelearn 2022.*

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *CoRR* abs/1606.06565 (2016).

[4] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating Behavioral Constraints in Online AI Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 3–11. https://doi.org/10.1609/aaai.v33i01.33013

[5] Leon Barrett and Srini Narayanan. 2008. Learning all optimal policies with multiple criteria. *Proceedings of the 25th International Conference on Machine Learning* (01 2008), 41–47. https://doi.org/10.1145/1390156.1390162

[6] Tom L. Beauchamp and James F. Childress. 1979. *Principles of biomedical ethics / Tom L. Beauchamp, James F. Childress.* Oxford University Press New York.

[7] Tom L. Beauchamp and James F. Childress. 2019. *Principles of Biomedical Ethics.* Oxford University Press. 8th Edition., New York.

[8] Trevor J. M. Bench-Capon and Katie Atkinson. 2009. Abstract Argumentation and Values. In *Argumentation in Artificial Intelligence.* Springer, 45–64. http://dx.doi.org/10.1007/978-0-387-98197-0_3

[9] Júlia Pareto Boada, Begoña Román Maestre, and Carme Torras Genís. 2021. The ethical issues of social assistive robotics: A critical literature review. *Technology in Society* 67 (2021), 101726.

[10] William Caballero, Roi Naveiro, and David Rios. 2021. Modeling Ethical and Operational Preferences in Automated Driving Systems. *Decision Analysis* 19 (10 2021). https://doi.org/10.1287/deca.2021.0441

[11] Joan Casas-Roma and Jordi Conesa. 2020. Towards the design of ethically-aware pedagogical conversational agents. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing.* Springer, 188–198.

[12] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity.* Springer, 13–39.

[13] K. L. Clarkson. 1988. Applications of Random Sampling in Computational Geometry, II. In *Proceedings of the Fourth Annual Symposium on Computational Geometry* (Urbana-Champaign, Illinois, USA) *(SCG '88).* Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/73393.73394

[14] David Cooper. 1993. *Value pluralism and ethical choice.* St. Martin Press, Inc., New York.

[15] European Comission. 2021. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206. Accessed: 2021-06-29.

[16] Diego Gracia. 1995. Hard times, hard choices: founding bioethics today. *Bioethics* 9, 3 (1995), 192–206.

[17] Sven Ove Hansson. 2001. *The Structure of Values and Norms.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511498466

[18] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2022. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Autonomous Agents and Multi-Agent Systems* 36 (2022).

[19] IEEE. 2019. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html. Accessed: 2021-06-29.

[20] Jan Leike, Miljan Martic, Viktoriya Krakovna, Pedro Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI Safety Gridworlds. *aRXiV 1711.09883* (11 2017).

[21] Jieting Luo, John-Jules Meyer, and Max Knobbout. 2017. Reasoning about Opportunistic Propensity in Multi-agent Systems. In *AAMAS 2017 Workshops, Best Papers.* 1–16.

[22] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Ramazon Kush, Murray Campbell, Moninder Singh, and Francesca Rossi. 2019. Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration. *IBM Journal of Research and Development* PP (09 2019), 6377–6381. https://doi.org/10.1147/JRD.2019.2940428

[23] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. 2022. Pareto Conditioned Networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.

[24] Mark O. Riedl and B. Harrison. 2016. Using Stories to Teach Human Values to Artificial Agents. In *AAAI Workshop: AI, Ethics, and Society.*

[25] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez Aguilar. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Environments. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 545–551. Main Track.

[26] Manel Rodriguez-Soto, Marc Serramia, Maite López-Sánchez, and Juan Rodríguez-Aguilar. 2022. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology* 24 (03 2022). https://doi.org/10.1007/s10676-022-09635-0

[27] Diederik Roijers and Shimon Whiteson. 2017. *Multi-Objective Decision Making.* Morgan and Claypool, California, USA. http://www.morganclaypool.com/doi/abs/10.2200/S00765ED1V01Y201704AIM034 doi:10.2200/S00765ED1V01Y201704AIM034.

[28] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A Survey of Multi-Objective Sequential Decision-Making. *J. Artif. Int. Res.* 48, 1 (Oct. 2013), 67–113.

[29] W. D. Ross. 1930. *The Right and the Good. Some Problems in Ethics.* Clarendon Press.

[30] Francesca Rossi and Nicholas Mattei. 2019. Building Ethically Bounded AI. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 9785–9789. https://doi.org/10.1609/aaai.v33i01.33019785

[31] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *Ai Magazine* 36 (12 2015), 105–114. https://doi.org/10.1609/aimag.v36i4.2577

[32] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2019. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34 (2019), 1–52.

[33] Shalom Schwartz. 2006. An Overview Basic Human Values: Theory, Methods, and Applications Introduction to the Values Theory. *Jerusalem Hebrew University* (2006).

[34] Marc Serramia, Maite López-Sánchez, Stefano Moretti, and Juan Antonio Rodríguez-Aguilar. 2021. On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–38.

[35] Luciano C. Siebert, Enrico Liscio, Pradeep K. Murukannaiah, Lionel Kaptein, Shannon L. Spruit, Jeroen van den Hoven, and Catholijn M. Jonker. 2022. Estimating Value Preferences in a Hybrid Participatory System. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022).* IOS Press, Amsterdam, the Netherlands, 1–14.

[36] Nate Soares and Benya Fallenstein. 2014. *Aligning superintelligence with human interests: A technical research agenda.* Machine Intelligence Research Institute (MIRI) technical report 8.

[37] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction.* MIT Press. http://www.worldcat.org/oclc/37293240

[38] Ibo van de Poel and Lambèr Royakkers. 2011. *Ethics, Technology, and Engineering: An Introduction.* Wiley-Blackwell.

[39] Eric W. Weisstein. 2002. *CRC Concise Encyclopedia of Mathematics.* Chapman & Hall : CRC Press, Boca Raton, FL.

[40] Yueh-Hua Wu and Shou-De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[41] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *IJCAI.* 5527–5533.