Multi-Objective Reinforcement Learning for Designing Ethical Environments

Manel Rodriguez-Soto[1] , **Maite Lopez-Sanchez**[2] , **Juan A. Rodriguez-Aguilar**[1]
[1]Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Spain
[2]Universitat de Barcelona (UB), Barcelona, Spain
{manel.rodriguez, jar}@iiia.csic.es, maite_lopez@ub.edu

Abstract

AI research is being challenged with ensuring that autonomous agents learn to behave ethically, namely in alignment with moral values. A common approach, founded on the exploitation of Reinforcement Learning techniques, is to design environments that incentivise agents to behave ethically. However, to the best of our knowledge, current approaches do not theoretically guarantee that an agent will learn to behave ethically. Here, we make headway along this direction by proposing a novel way of designing environments wherein it is formally guaranteed that an agent learns to behave ethically while pursuing its individual objective. Our theoretical results develop within the formal framework of Multi-Objective Reinforcement Learning to ease the handling of an agent's individual and ethical objectives. As a further contribution, we leverage on our theoretical results to introduce an algorithm that automates the design of ethical environments.

1 Introduction

As artificial agents become more intelligent and pervade our societies, it is key to guarantee that situated agents act *value-aligned*, that is, in alignment with human values [Soares and Fallenstein, 2014; Russell *et al.*, 2015]. Otherwise, we are prone to potential ethical risk in critical areas as diverse as elder caring [Barcaro *et al.*, 2018], personal services [Wynsberghe, 2016], and automated driving [Lin, 2015]. As a consequence, there has been a growing interest in the Machine Ethics [Yu *et al.*, 2018; Rossi and Mattei, 2019] and AI Safety [Amodei *et al.*, 2016; Leike *et al.*, 2017] communities in the use of Reinforcement Learning (RL) [Sutton and Barto, 1998] to deal with the urging problem of *value alignment*.

Among these two communities, it is common to find proposals to tackle the value alignment problem by designing an environment that incentivises ethical behaviours (or penalises unethical ones) by means of some exogenous reward function (e.g., [Riedl and Harrison, 2016; Abel *et al.*, 2016; Wu and Lin, 2017; Noothigattu *et al.*, 2019; Balakrishnan *et al.*, 2019; Rodriguez-Soto *et al.*, 2020]). We observe that this approach consists in a two-step process: first, the ethical knowledge is encoded as rewards (*reward specification*); and then, these rewards are incorporated into the agent's learning environment (*ethical embedding*).

The literature is populated with embedding solutions that use a linear scalarisation function for *weighting* the agent's individual reward with the ethical reward (e.g. [Wu and Lin, 2017; Rodriguez-Soto *et al.*, 2020]). However, to the best of our knowledge, there are no studies following the linear scalarisation approach that offer theoretical guarantees regarding the learning of ethical behaviours. Furthermore, [Vamplew *et al.*, 2018] point out some shortages of adopting a linear ethical embedding: the agent's learnt behaviour will be heavily influenced by the relative scale of the individual rewards. This issue is specially relevant when the ethical objective must be wholly fulfilled (e.g., a robot in charge of buying an object should never decide to steal it [Arnold *et al.*, 2017]). For those cases, the embedding must be done in such a way that ethical behaviour is prioritised, providing theoretical guarantees for the learning of ethical policies.

Against this background, the objective of this work is twofold: (1) to offer theoretical guarantees for the linear embedding approach so that we can create an *ethical environment*, that is, an environment wherein it is ensured that an agent learns to behave ethically while pursuing its individual objective; (2) and to automate the design of such ethical environment. We address such goals within our view of ethical environment design process, as depicted in Figure 1. According to our view, a reward specification task takes the individual and ethical objectives to yield a multi-objective environment. Thereafter, an ethical embedding task transforms the multi-objective environment into an ethical environment, which is the one wherein an agent learns. Within the framework of such ethical environment design process, we address the goals above, focusing on the ethical embedding task, to make the following novel contributions.

Firstly, we characterise the policies that we want an agent to learn, the so-called *ethical policies*: those that prioritise ethical objectives over individual objectives. Thereafter, we propose a particular ethical embedding approach, and formally prove that the resulting learning environment that it yields is ethical. This means that we guarantee that an agent will always learn ethical policies when interacting in such environment. Our theoretical results are based on the formalisation of the ethical embedding process within the framework

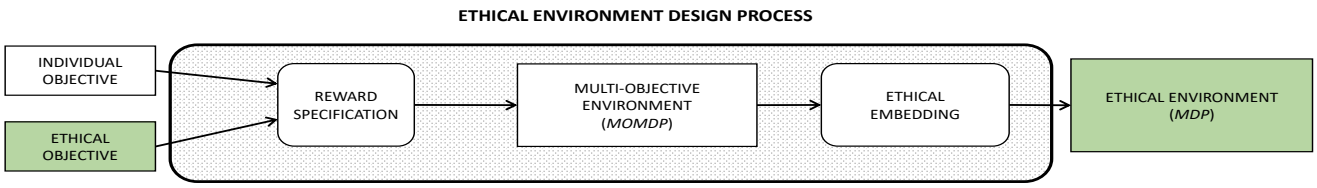

Figure 1: The process of designing an ethical environment is performed in two steps: a reward specification and an ethical embedding. Our algorithm computes the latter. Rectangles stand for objects whereas rounded rectangles correspond to processes.

of Multi-Objective Reinforcement Learning (MORL)[Roijers *et al.*, 2013], which provides Multi-objective MDPs (MOMDPs) to handle both individual and ethical objectives. Thus, MOMDPs provide the model for the multi-objective environment that results from reward specification (Figure 1).

Secondly, based on our theoretical results, we propose an algorithm to implement our ethical embedding. This novel algorithm tailors current developments in the MORL literature to build an ethical environment as a single-objective MDP from the multi-objective MDP that stems from the reward specification process. Since the resulting single-objective MDP encapsulates the ethical rewards, the agent can thus apply a basic RL method to learn its optimal policy there. Specifically, we ground ethical embedding algorithm on the computation of convex hulls (as described in [Barrett and Narayanan, 2008]) as the means to find ethical policies.

To summarise, in this paper we make headway in building ethical environments by providing two main novel contributions: (i) the theoretical means to design the learning environment so that an agent's ethical learning is guaranteed; and (ii) algorithmic tools for automating the configuration of the learning environment.

In what follows, Section 2 presents our formalisation of the ethical embedding problem that we must solve to create an ethical environment. Next, Section 3 studies how to guarantee the learning of ethical policies in ethical environments, and Section 4 introduces our algorithm to build ethical environments. Subsequently, Section 5 illustrates our proposal by means of a simple example, the public civility game. Finally, Section 6 concludes and sets paths to future work.

2 Formalising the ethical embedding problem

In this section we propose a formalisation of the *ethical embedding* of value alignment problems in which an ethical objective must be fulfilled and an individual objective is pursued. Our main goal is to guarantee that an agent will learn to behave ethically, that is, to behave in alignment with a moral value. In the Ethics literature, moral values (also called ethical principles) express the moral objectives worth striving for [van de Poel and Royakkers, 2011].

As mentioned above, the value alignment problem can be divided in two steps: the *reward specification* (to transform ethical knowledge into ethical rewards) and the *ethical embedding* (to ensure that these rewards incentivise the agent to be ethical). Although both are critical problems in the Machine Ethics and AI Safety community, in this paper we focus on the ethical embedding problem, and likewise we assume that we already have a reward specification in the form of a Multi-Objective Markov Decision Processes (MOMDP) [Roijers *et al.*, 2013]. This way we can handle an ethical objective and an agent's individual objective within the same learning framework. Precisely, MOMDPs formalise sequential decision making problems in which we need to ponder several objectives. Formally:

Definition 1. *A (finite)*[1] *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 order to transform an MOMDP into a single-objective MDP, the vectorial reward function $\vec{V}$ can be scalarised by means of a *scalarisation* function f. With f, the agent's problem becomes to learn a policy that maximises $f(\vec{V})$, a single-objective problem. It is specially notable the particular case in which f is linear, because in such case the scalarised problem can be solved with single-objective reinforcement learning algorithms. We refer to any linear f simply as a weight vector $\vec{w}$. Any policy that maximises $f(\vec{V}) = \vec{w} \cdot \vec{V}$ is thus optimal in the MDP $\langle \mathcal{S}, \mathcal{A}, \vec{w} \cdot \vec{R}, T \rangle$.

We define an *ethical MOMDP* as an MOMDP encoding the reward specification of a value alignment problem in which the agent must consider both its individual objective and an ethical objective. The first component in the corresponding vectorial reward function characterises the individual agent's objective (as usually done in RL), whereas the subsequent components represent the ethical objective [Horgan and Timmons, 2010]. Following the Ethics literature [Chisholm, 1963; Frankena, 1973; van de Poel and Royakkers, 2011; Etzioni and Etzioni, 2016], we define an ethical objective through two dimensions: (i) a *normative dimension*, which punishes the violation of normative requirements; and (ii) an *evaluative dimension*, which rewards morally praiseworthy actions. Formally:

Definition 2 (Ethical MOMDP)**.** *Given a MOMDP*

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle, \tag{1}$$

where R_0 corresponds to the reward associated to the individual objective, we say that $\mathcal{M}$ is an ethical MOMDP *if and*

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

only if:

- $R_{\mathcal{N}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^-$ *is a* normative *reward function penalising the violation of normative requirements; and*
- $R_E : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ *is an* evaluative *reward function that (positively) rewards the performance of actions evaluated as praiseworthy.*

Having two separate ethical reward functions allows us to avoid the ethical problem of an agent learning to maximise its accumulation of praiseworthy actions while disregarding some of its normative requirements.

In the ethical embedding, we transform an ethical MOMDP into a single-objective MDP (in which the agent will learn its policy) by means of scalarisation function f_e, which we call the *embedding function*. In the particular case that f_e is linear, we say that we are applying a linear embedding or a *weighting*.

Ethical MOMDPs pave the way to characterise our notion of ethical policy: an *ethical policy* is a policy that abides to all the norms while also behaving as praiseworthy as possible. In other words, it is a policy that adheres to the specification of the ethical objective. We capture this notion by means of the normative and evaluative components of the value function in an ethical MOMDP:

Definition 3 (Ethical policy)**.** *Let $\mathcal{M}$ be an ethical MOMDP. We say that a policy π_* is an ethical policy in $\mathcal{M}$ if and only if its value function $\vec{V}^{\pi_*} = (V_0^{\pi_*}, V_{\mathcal{N}}^{\pi_*}, V_E^{\pi_*})$ is optimal for its ethical objective (i.e., both its normative $V_{\mathcal{N}}$ and evaluative V_E components):*

$$V_{\mathcal{N}}^{\pi_*} = \max_{\pi} V_{\mathcal{N}}^{\pi},$$
$$V_E^{\pi_*} = \max_{\pi} V_E^{\pi}.$$

For the sake of simplicity, we refer to a policy that is not ethical in the sense of Definition 3 as an *unethical* policy.

With ethical policies, we can now define formally *ethical-optimal* policies: the policies that we want an agent to learn. Ethical-optimal policies correspond to those policies in which the individual objective is pursued subject to the ethical objective being fulfilled. Specifically, we say that a policy is *ethical-optimal* if and only if it is ethical and it also maximises the individual objective V_0 (i.e., the accumulation of rewards R_0). Formally:

Definition 4 (Ethical-optimal policy)**.** *Given an MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$, a policy π_* is* ethical-optimal *in M if and only if*

$$V_0^{\pi_*} = \max_{\pi \in \Pi_e} V_0^{\pi},$$

where Π_e is the set of ethical policies.

Given an MOMDP encoding individual and ethical rewards, our aim is to find an embedding function that guarantees that it is only possible for an agent to learn ethical-optimal policies over the scalarised MOMDP (as a single-objective MDP). Thus, we must design an embedding function that scalarises the rewards received by the agent in such a way that ensures that ethical-optimal policies are optimal for the agent. In its simplest form, this embedding function will have the form of a linear combination of individual and ethical objectives

$$f(\vec{V}^{\pi}) = \vec{w} \cdot \vec{V}^{\pi} = w_0 V_0^{\pi} + w_{\mathcal{N}} V_{\mathcal{N}}^{\pi} + w_E V_E^{\pi} \qquad (2)$$

where $\vec{w} = (w_0, w_{\mathcal{N}}, w_E)$ is a weight vector with all weights $w_0, w_{\mathcal{N}}, w_E > 0$ to guarantee that the agent is taking into account all rewards (i.e., both objectives). Without loss of generality, we fix the individual weight to $w_0 = 1$.

Therefore, we can formalise the ethical embedding problem as that of computing a weight vector $\vec{w}$ that incentivises an agent to behave ethically while still pursuing its individual objective. Formally:

Problem 1 (Ethical embedding)**.** *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$ be an ethical MOMDP. Compute a weight vector $\vec{w}$ with positive weights such that all optimal policies in the MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, w_0 R_0 + w_N R_{\mathcal{N}} + w_E R_E, T \rangle$ are also ethical-optimal 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-optimal is a solution of Problem 1. The next section proves that such solutions always exist for any ethical MOMDP.

3 Solvability of the ethical embedding problem

This section is devoted to describe the minimal conditions under which there always exists a solution to Problem 1, and to prove that such solution actually exists. This solution (a weight vector) will allow us to apply the ethical embedding process to produce an ethical environment (a single-objective MDP) wherein an agent learns to behave ethically (i.e., an ethical-optimal policy).

For all the following theoretical results, we assume the following condition for any ethical MOMDP: if we want the agent to behave ethically, it must be actually possible for it to behave ethically[2]. Formally:

Condition 1 (Ethical policy existence)**.** *Given an ethical MOMDP, there is at least one ethical policy (as defined by Def. 3).*

If Condition 1 holds, next Theorem guarantees that Problem 1 is always solvable, or in other words, that it is always possible to guarantee that the learnt behaviour of an agent will be ethical if we give a reward incentive that is large enough. Furthermore, this Theorem also dictates that, without loss of generality, we can assume that the normative and evaluative weights in the solution weight vector $\vec{w}$ are identical ($w_{\mathcal{N}} = w_E$). We will be referring thus to w_E as the *ethical weight*. Formally:

Theorem 1 (Solution existence)**.** *Given an ethical MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$ for which Condition 1 is satisfied, there exists a weight vector $\vec{w} = (1, w_E, w_E)$ with $w_E > 0$ for which every optimal policy in the MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, w_0 R_0 + w_N R_{\mathcal{N}} + w_E R_E, T \rangle$ is also ethical-optimal in $\mathcal{M}$.*

[2] In the Ethics literature this condition is summarised with the expression *Ought implies can* [Duignan, 2018].

Proof. We provide a sketch of the proof. The proof is done in two steps: (1) First we prove that if for a weight vector $\vec{w}$ there is a deterministic $\vec{w}$-optimal policy ρ that is an unethical policy, then we can always increase the weight w_E in $\vec{w}$ enough so that ρ is strictly worse than an ethical policy π (which exists thanks to Condition 1), so ρ is no longer an $\vec{w}$-optimal policy.

(2) Once the first step is proven, we can identify the unethical policy ρ_* that requires the greatest increase of w_E in order to be $\vec{w}$-suboptimal. After increasing w_E for ρ_*, all unethical policies will become $\vec{w}$-suboptimal. However, since there always exists at least one deterministic $\vec{w}$-optimal policy, by this process of elimination all remaining $\vec{w}$-optimal policies must be ethical policies (and at least one exists thanks to Condition 1), and therefore, they will be ethical-optimal. □

4 Solving the ethical embedding problem

This section is devoted to explaining how to compute a solution weight vector $\vec{w}$ for the ethical embedding problem (Problem 1). Such weight vector $\vec{w}$ allows us to combine individual and ethical rewards into a single reward to create an ethical environment in which the agent can learn its behaviour, that is, an ethical-optimal policy.

In what follows we detail an algorithm to solve the ethical embedding problem, the so-called *Ethical Embedding* algorithm. Specifically, our algorithm performs the following three steps:

1. *Computation of the partial convex hull* containing a subset P of policies of an ethical MOMDP $\mathcal{M}$ that are optimal for some weight vector.
2. *Extraction of the ethical-optimal policies* Π_* from the partial convex hull P.
3. *Computation of the embedding function*: use the reference policies Π_* to find a linear weighting $\vec{w}$ of the rewards pondering individual and ethical objectives to yield an ethical environment wherein the learning of ethical policies is guaranteed.

The following three subsections provide the theoretical grounds for computing each step of our algorithm. Then, Subsection 4.4 presents the algorithm as a whole.

4.1 Computation of the partial convex hull

Our algorithm applies a linear ethical embedding (a weight vector) to solve Problem 1. Theorem 1 determines a structure for the solution weight vector $\vec{w}$ of Problem 1. In order to compute a specific value for $\vec{w}$, we resort to the multi-objective RL concept of *convex hull*.

Given a MOMDP $\mathcal{M}$, its *convex hull* [Roijers *et al.*, 2013] is composed of those policies that are strictly better than any other policy for some linear weights. Formally:

Definition 5 (Convex hull)**.** *Given an MOMDP $\mathcal{M}$, its convex hull CH is the subset of policies $\Pi^{\mathcal{M}}$ for which there exists a weight vector $\vec{w}$ for which the linearly scalarised value function is maximal:*

$$CH(\mathcal{M}) = \{\pi_* \in \Pi^{\mathcal{M}} \mid \exists \vec{w} :\ \pi_* \in \arg\max_{\pi} \vec{w} \cdot \vec{V}^{\pi}\}. \quad (3)$$

The convex hull of an ethical MOMDP naturally contains all ethical-optimal policies by definition. Thus, it allows us to derive the weight vector necessary to guarantee that all optimal policies are ethical-optimal, which we know that exist thanks to Theorem 1. However, computing the whole convex hull of an MOMDP can be computationally demanding. Fortunately, Theorem 1 naturally characterises the minimal convex hull that we need to compute to find the solution of the ethical embedding problem, hence avoiding the computation of the whole convex hull. Formally:

Theorem 2. *Given an ethical MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T\rangle$ in which Condition 1 is satisfied, let $P \subseteq CH(\mathcal{M})$ be the region of the convex hull of $\mathcal{M}$, limited to weight vectors of the form $\vec{w} = (1, w_E, w_E)$ with $w_E > 0$. Then, P contains all ethical-optimal policies.*

Proof. From Theorem 1, we know that at least one ethical-optimal policy is optimal for a weight vector $\vec{w}$ of the form $\vec{w} = (1, w_E, w_E)$ with $w_E > 0$. Notice that by definition, all ethical-optimal policies share the same vectorial reward function and thus, all of them are optimal for the same weight. Therefore, all of them belong to this partial region P of the convex hull $CH(\mathcal{M})$. □

Henceforth, when referring to the *partial convex hull*, we are referring to this particular region P shown in Theorem 2.

To finish this subsection, we remark that this partial region of the convex hull can be computed by adapting state of the art algorithms such as Convex Hull Value Iteration [Barrett and Narayanan, 2008] –which compute the whole convex hull of an MOMDP– to only compute a region of the convex hull.

4.2 Extraction of the ethical-optimal policies

After computing the partial convex hull $P \subseteq CH(\mathcal{M})$, we are ready to perform the second step of our algorithm, which is the extraction of ethical-optimal policies from P. Notice that a policy in P is ethical-optimal if and only if is ethical. Thus, in order to know which policies in P are ethical-optimal, we have to find the ones that maximise both the normative and evaluative reward functions ($V_{\mathcal{N}}$ and V_E respectively) of the ethical MOMDP. This corresponds to the process of *ethical-optimal policy computation*. Formally, to obtain the ethical-optimal policies within P we must compute:

$$\Pi^* = \arg\max_{\pi \in P} (V^{\pi}_{\mathcal{N}}(s) + V^{\pi}_{E}(s)) \text{ for every state } s. \quad (4)$$

Here, Π^* is the set of all ethical-optimal policies of P, which thanks to Theorem 2 it is also in fact the set of all ethical-optimal policies of the ethical MOMDP $\mathcal{M}$. Notice that $\vec{V}^{\pi}_{\mathcal{N}}$ and $\vec{V}^{\pi}_{E}$ are already available for any policy π in the partial convex hull P because their computation was required in order to obtain P.

4.3 Computation of the embedding function

In the last step of our algorithm, the computation of the *embedding function* (the weight vector), we use the computed partial convex hull and the ethical-optimal policies to find the solution weight vector $\vec{w} = (1, w_E, w_E)$ that guarantees that optimal policies are ethical-optimal. In other words, such

weight vector $\vec{w}$ will create an ethical environment (a single-objective MDP) in which the agent will learn an ethical-optimal policy.

Finding the actual values of such weight vector is not straightforward because $\vec{w} \in \mathbb{R}^3$. However, thanks to our previous result in Theorem 2, we can reduce our search space from $\mathbb{R}^3$ to $\mathbb{R}$. In more detail, in order to find our targeted $\vec{w} = (1, w_E, w_E)$, we only need to consider the problem of finding the ethical weight w_E that guarantees that ethical-optimal policies are optimal in the partial convex hull P. Formally, we need to find a value for $w_E \in \vec{w}$ such that:

$$\vec{w} \cdot V^{\pi_*}(s) > \max_{\pi \in P \setminus \Pi_*} \vec{w} \cdot V^{\pi}(s), \tag{5}$$

for every state $s \in \mathcal{S}$. Here, Π_* is the set of ethical-optimal policies and π_* is any policy within Π_*.

Notice that in Eq. 5 the only unknown variable is w_E. This amounts to solving a system of $n \cdot |\mathcal{S}|$ linear inequalities (where n is the number of policies in P) with a single unknown variable.

4.4 An algorithm for designing ethical environments

At this point we now count on all the tools for solving Problem 1, and hence build an ethical environment where the learning of ethical policies is guaranteed. Algorithm 1 implements the ethical embedding outlined in Figure 1. The algorithm starts in line 2 by computing the partial convex hull $P \subseteq CH(\mathcal{M})$ of the input ethical MOMDP $\mathcal{M}$ (see Subsection 4.1); and then in line 3 it obtains the set of ethical-optimal policies Π^* out of those in the partial convex hull P (see Subsection 4.2). Thereafter, in line 4 our weighting process searches, within P, for an ethical weight w_E that satisfies Equation 5 (see Subsection 4.3). For the obtained weight vector $\vec{w} = (1, w_E, w_E)$, all optimal policies of the single-objective MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, w_0 R_0 + w_N R_{\mathcal{N}} + w_E R_E, T \rangle$ will be ethical. In other words, such weight vector will solve the ethical embedding problem (Problem 1). Finally, the algorithm returns the MDP $\mathcal{M}'$ in line 5.

Algorithm 1 Ethical Embedding

1: **function** EMBEDDING(Ethical MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$)
2: Compute $P \subseteq CH(\mathcal{M})$ the partial convex hull of $\mathcal{M}$ for weight vectors $\vec{w} = (1, w_E, w_E)$ with $w_E > 0$.
3: Find Π^* the set of ethical-optimal policies within P by solving Eq. 4.
4: Find a value for w_E that satisfies Eq. 5.
5: Return MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, R_0 + w_E(R_{\mathcal{N}} + R_E), T \rangle$.
6: **end function**

The computational cost of the algorithm mainly resides in computing the partial convex hull of an MOMDP. The Convex Hull Value Iteration algorithm requires $O(n \cdot log\, n)$ times what its single-objective Value Iteration counterpart [Clarkson, 1988; Barrett and Narayanan, 2008] requires, where n is the number of policies in the convex hull. In our case this number will be $n' \leq n$ since we are just allowing a particular form of weights, as explained in previous subsections. Notice that after computing $P \subseteq CH$, solving Eq. 4 is a sorting operation because we already have calculated $\vec{V}^{\pi}$ for every $\pi \in P$. Similarly, solving Eq. 5 requires to solve $n \cdot |\mathcal{S}|$ inequalities and then sort them to find the ethical weight w_E.

5 Example: the Public Civility Game

This section illustrates our process of designing an ethical environment (Algorithm 1) with a simple example. We use a single-agent version[3] of the *Public Civility Game* [Rodriguez-Soto *et al.*, 2020], a value alignment problem where an agent learns to behave according to the moral value of civility. This example can be seen as an ethical adaptation of the *irreversible side effects* environment from [Leike *et al.*, 2017].

Figure 2 (left) depicts the environment, wherein two agents (L and R) move from their initial positions to their respective goal destinations (GL and GR). Since the L agent finds garbage (small red square) blocking its way, it needs to learn how to handle the garbage civically while moving towards its goal GL. The civic (ethical) behaviour we expect agent L to learn is to push the garbage to the bin without throwing it to agent R, which, in our setting, has a fixed behaviour.

5.1 Reward specification

The Public Civility Game represents an ethical embedding problem where civility is the moral value to embed in the environment. As such, we encode it as an ethical MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$ in which the agent's individual and ethical objectives have been specified as follows.

On the one hand, the **agent's individual objective** is to reach its destination as fast as possible. Thus, the individual reward function R_0 returns a positive reward of 20 to the agent whenever located at its goal. Otherwise, it returns -1.

On the other hand, the **ethical objective** is to promote civility by means of:

- An evaluative reward function R_E that rewards the agent when performing the praiseworthy action of pushing the garbage inside the bin with a positive reward of 10. It returns 0 in any other circumstance.
- A normative reward function $R_{\mathcal{N}}$ that punishes the agent with a negative reward for not complying with the moral requirement of being respectful with other agents. Thus, agent L will be punished with a negative reward of -10 if it throws the garbage to agent R. Otherwise, it returns 0.

5.2 Ethical embedding

We now apply Algorithm 1 to design an ethical environment for the Public Civility Game. In what follows, we detail the three processes involved in obtaining this new environment.

Partial convex hull computation: Considering $\mathcal{M}$, our ethical MOMDP, we compute the partial convex hull $P \subseteq CH$. Figure 2 (centre) depicts the resulting P for the initial state s_0. It is composed of 3 different policies named after the behaviour they encapsulate: (1) an **U**nethical (uncivil) policy,

[3] Programmed in Python. Code available at https://gitlab.iiia.csic.es/Rodriguez/morl-for-ethical-environments.

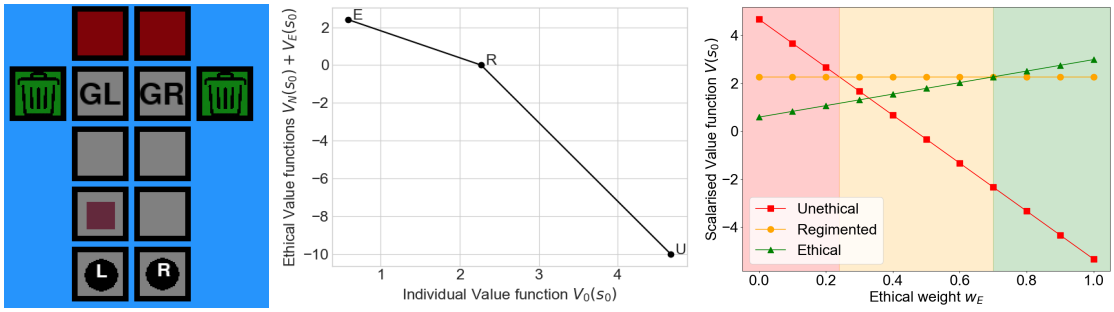

Figure 2: Left: Initial state of the public civility game. The agent on the left has to deal with the garbage obstacle, which has been located in front of it. Centre: Visualisation in Objective Space of the partial convex hull of $\mathcal{M}$ composed by 3 policies: E (Ethical), R (Regimented) and U (Unethical). Right: Visualisation in Weight Space of the partial convex hull of $\mathcal{M}$. Painted areas indicate which policy is optimal for the varying values of the ethical weight w_E.

Policy π	**Value** $\vec{V}^{\pi}(s_0)$	w_E **ranges**
Unethical	(4.67, -10, 0)	[0.0, 0.24]
Regimented	(2.27, 0, 0)	[0.24, 0.7]
Ethical	(0.59, 0, 2.4)	[0.7, ∞)

Table 1: Policies π within the partial convex hull of the Public Civility Game and their associated values $\vec{V}^{\pi} = (V_0^{\pi}, V_{\mathcal{N}}^{\pi}, V_E^{\pi})$. Weight ranges indicate the values of w_E for which each policy is optimal.

in which the agent moves towards the goal and throws away the garbage without caring about any ethical implication; (2) a **R**egimented policy, in which the agent complies with the norm of not throwing the garbage to the other agent; and finally, (3) an **E**thical policy, in which the agent behaves civically as desired. Table 1 provides the specific vectorial value $\vec{V}^{\pi} = (V_0^{\pi}, V_{\mathcal{N}}^{\pi}, V_E^{\pi})$ of each policy π and the range of values of the ethical weight w_E for which each policy is optimal.

Extraction of the ethical-optimal policies: In our case, the **E**thical policy π_E is the only ethical-optimal policy within the partial convex hull P. Indeed, π_E is the only policy that maximises both the normative and the evaluative components ($V_{\mathcal{N}}$ and V_E respectively). Last row in Table 1 shows the value of π_E for the initial state s_0: $\vec{V}^{\pi_E}(s_0) = (0.59, 0, 2.4)$.

Computation of the embedding function: Line 4 in Algorithm 1 computes the weight w_E in $\vec{w} = (1, w_E, w_E)$ for which π_E is the only optimal policy of P, by solving Eq. 5:

$$\vec{w} \cdot V^{\pi_E}(s_0) > \max_{\rho \in P \setminus \{\pi_E\}} [V_0^{\rho}(s_0) + w_E \cdot (V_{\mathcal{N}}^{\rho}(s_0) + V_E^{\rho}(s_0))].$$

By solving it, we find that if $w_E > 0.7$, then the **E**thical policy becomes the only optimal one. We can check it: $0.59 + 0.7 \cdot (0 + 2.4) = 2.27 \geq max((4.67 + 0.7 \cdot (-10 + 0)), (2.27 + 0.7 \cdot (0 + 0)) = max(-2.33, 2.27)$.

Figure 2 (right) illustrates the scalarised value of the 3 policies for varying values of w_E in [0,1] (for w_E>1 tendencies do not change). In particular, focusing on the green painted area, we can observe that the **E**thical policy becomes the only optimal one when $w_E > 0.7$.

Therefore, the last step in our algorithm returns an MDP whose reward comes from scalarising the MOMDP by $\vec{w} = (1, w_E, w_E)$, being w_E strictly greater than 0.7. Thus, adding any $\epsilon > 0$ will suffice. If, for instance, we set $\epsilon = 0.01$ then, the weight vector $(1, 0.7 + 0.01, 0.7 + 0.01) = (1, 0.71, 0.71)$ solves the Public Civility Game. More clearly, an MDP created from an embedding function with such w_E incentivises the agent to learn the **E**thical policy. Indeed, when we set up the agent L to learn with Q-Learning [Sutton and Barto, 1998] in the designed ethical environment, it learns to bring the garbage to the bin while moving towards its goal.

6 Conclusions and future work

Designing ethical environments for learning agents is a challenging problem. We make headway in tackling this problem by providing novel formal and algorithmic tools that build upon Multi-Objective Reinforcement Learning. In particular, our problem consists in ensuring that the agent wholly fulfils its ethical objective while pursuing its individual objective.

MORL is a valuable framework to handle multiple objectives. In order to ensure ethical learning (value-alignment), we formalise –within the MORL framework– *ethical-optimal* policies as those that prioritise their ethical objective. Overall, we design an ethical environment by considering a two-step process that first specifies rewards and second performs an ethical embedding. We formalise this last step as the ethical embedding problem and theoretically prove that it is always solvable. Our findings lead to an algorithm for automating the design of an ethical environment. Our algorithm ensures that, in this ethical environment, it will be in the best interest of the agent to behave ethically while still pursuing its individual objectives. We illustrate it with a simple example that embeds the moral value of civility.

As to future work, we would like to further examine empirically our algorithm in more complex environments.

Acknowledgments

Research supported by projects AI4EU (H2020-825619), LOGISTAR (H2020-769142), COREDEM (H2020-785907), Crowd4SDG (H2020-872944), CI-SUSTAIN (PID2019-104156GB-I00), COMRIDI18-1-0010-02, MISMIS PGC2018-096212B-C33, TAILOR (H2020-952215), 2017 SGR 172 and 2017 SGR 341. Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387).

References

[Abel *et al.*, 2016] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Work.: AI, Ethics, and Society*, volume 92, 2016.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *CoRR*, abs/1606.06565, 2016.

[Arnold *et al.*, 2017] T. Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment - what will keep systems accountable? In *AAAI Workshops*, 2017.

[Balakrishnan *et al.*, 2019] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3–11, 07 2019.

[Barcaro *et al.*, 2018] Rosangela Barcaro, M. Mazzoleni, and P. Virgili. Ethics of care and robot caregivers. *Prolegomena*, 17:71–80, 06 2018.

[Barrett and Narayanan, 2008] Leon Barrett and Srini Narayanan. Learning all optimal policies with multiple criteria. *Proceedings of the 25th International Conference on Machine Learning*, pages 41–47, 01 2008.

[Chisholm, 1963] R. M. Chisholm. Supererogation and offence: A conceptual scheme for ethics. *Ratio (Misc.)*, 5(1):1, 1963.

[Clarkson, 1988] K. L. Clarkson. Applications of random sampling in computational geometry, ii. In *Proceedings of the Fourth Annual Symposium on Computational Geometry*, SCG '88, page 1–11, New York, NY, USA, 1988. Association for Computing Machinery.

[Duignan, 2018] Brian Duignan. Ought implies can. https://www.britannica.com/topic/ought-implies-can, May 2018. Accessed: 2021-01-15.

[Etzioni and Etzioni, 2016] Amitai Etzioni and Oren Etzioni. Designing ai systems that obey our laws and values. *Commun. ACM*, 59(9):29–31, August 2016.

[Frankena, 1973] William K. Frankena. *Ethics, 2nd edition*. Englewood Cliffs, N.J. : Prentice-Hall,, 1973.

[Horgan and Timmons, 2010] Terry Horgan and Mark Timmons. Untying a knot from the inside out: Reflections on the "paradox" of supererogation. *Social Philosophy and Policy*, 27:29 – 63, 07 2010.

[Leike *et al.*, 2017] Jan Leike, Miljan Martic, Viktoriya Krakovna, Pedro Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *aRXiV 1711.09883*, 11 2017.

[Lin, 2015] Patrick Lin. *Why Ethics Matters for Autonomous Cars*, pages 69–85. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[Noothigattu *et al.*, 2019] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Ramazon Kush, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*, PP:6377–6381, 09 2019.

[Riedl and Harrison, 2016] Mark O. Riedl and B. Harrison. Using stories to teach human values to artificial agents. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[Rodriguez-Soto *et al.*, 2020] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodríguez-Aguilar. A structural solution to sequential moral dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Aystems (AAMAS 2020)*, 2020.

[Roijers *et al.*, 2013] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, October 2013.

[Rossi and Mattei, 2019] Francesca Rossi and Nicholas Mattei. Building ethically bounded ai. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9785–9789, 07 2019.

[Russell *et al.*, 2015] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36:105–114, 12 2015.

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

[Sutton and Barto, 1998] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998.

[Vamplew *et al.*, 2018] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20, 03 2018.

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

[Wu and Lin, 2017] Yueh-Hua Wu and Shou-De Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. *arXiv*, 12 2017.

[Wynsberghe, 2016] A. Wynsberghe. Service robots, care ethics, and design. *Ethics and Inf. Technol.*, 18(4):311–321, December 2016.

[Yu *et al.*, 2018] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building ethics into artificial intelligence. In *IJCAI*, page 5527–5533, 2018.