Building Multi-Agent Environments with Theoretical Guarantees on the Learning of Ethical Policies[*]

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

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

ABSTRACT

This paper tackles the open problem of value alignment in multi-agent systems. In particular, we propose an approach to build an *ethical* environment that guarantees that all agents in the system learn to behave ethically while pursuing their individual objectives. Our contributions are founded in the framework of Multi-Objective Multi-Agent Reinforcement Learning. Firstly, we characterise a family of Multi-Objective Markov Games (MOMGs), the so-called *ethical* MOMGs, for which we can formally guarantee the learning of ethical behaviours. From these, we specify the process for building single-objective ethical environments that simplify the learning in the multi-agent system. Interestingly, our theoretical results for multi-agent environments generalise recent state-of-the-art results for single-agent environments.

KEYWORDS

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

1 INTRODUCTION

The challenge of guaranteeing that autonomous agents act *value-aligned* (in alignment with human values) [23, 27], is becoming critical as agents increasingly populate our society. Hence, it is of great concern to design ethically-aligned trustworthy AI [4] capable of respecting our values [7, 11]. Indeed, there has recently been a rising interest in both the Machine Ethics [22, 32] and AI Safety [1, 13] communities in applying Reinforcement Learning (RL) [28] to tackle the critical problem of *value alignment*.

These two communities typically deal with the value alignment problem by designing an environment endowed with incentives for behaving ethically. Thus, often in the literature a *single* agent receives incentives through an exogenous reward function (e.g., [2, 16, 31]). Firstly, this reward function is specified from some ethical knowledge. Afterwards, it is incorporated into an agent's learning environment through an *ethical embedding* process. Against this background, in this paper we tackle the novel problem of designing an ethical embedding process for multi-agent systems.

Specifically, we propose the creation of an *ethical environment* providing theoretical guarantees that all agents learn to behave ethically while pursuing their respective individual objectives. In particular, we focus on environments wherein agents share a *social* ethical objective that they should prioritise to their individual objectives (e.g., agents heading to their destinations over a pond should stop to rescue someone drowning in the waters below [26]). This consideration of ethical and individual objectives requires the learning agents to take a multi-objective approach that may be complex to handle. Hence, we propose to create a (simpler) single-objective ethical environment that embeds both ethical and individual objectives. For that purpose, we follow the prevailing approach of applying a linear scalarisation function that *weighs* individual and ethical rewards ([2, 31]). However, scalarisation has some known problems [9] and in fact some weightings may deviate the agents from behaving ethically and, thus, we must explicitly (and theoretically) ensure that this will not be the case. This has been addressed recently by [18] and yet, they just guarantee it for a single agent. Hence, guaranteeing that all agents in a multi-agent system learn to behave ethically remains an open problem. We tackle it by proposing the Multi-Agent Ethical Embedding (MAEE) process depicted in Figure 1, whose input is a multi-objective environment that is transformed into a single-objective ethical environment.

Our contribution is two-fold. Firstly, we formalise the MAEE Problem within the framework of Multi-Objective Markov Games (MOMG) [24, 25]. This formalisation allows us to characterise the so-called *Ethical* MOMGs, the family of MOMGs for which we can solve the problem. As Figure 1 shows, solving the problem amounts to transforming an ethical MOMG into an ethical Markov Game (MG), where agents do not need to apply *Multi-Objective* RL [20, 21], but just RL [12, 14]. Our formalisation involves the definition of *ethical policies* as well as *ethical equilibrium.* An ethical policy defines the behaviour of an agent prioritising the shared ethical objective over its individual objective. An ethical equilibrium is a joint policy composed of ethical policies that characterises the *target* equilibrium in the ethical environment.

Secondly, we propose a novel process to solve the MAEE problem that generalises the single-agent ethical embedding process in [18]. Our process involves two consecutive decompositions of the multi-agent problem into *n* single-agent problems: the first one (Figure 1 left) allows the computation of the ethical equilibrium (i.e., the target joint policy); whereas the second one (/right) computes the weight vector that solves our embedding problem.

Next, Section 2 presents our formalisation of the MAEE problem. Section 3 studies the multi-agent environments to which we can apply a MAEE process, and Section 4 details our process to build

[*]Work funded by projects AI4EU (H2020-825619), Crowd4SDG (H2020-872944), TAILOR (H2020-952215), COREDEM (H2020-785907), and 21S01802-001 from Barcelona City Council through the Fundació Solidaritat de la UB. Financial support was also received from grant PID2019-104156GB-I00 funded by MCIN/AEI/10.13039/501100011033. Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387)..

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

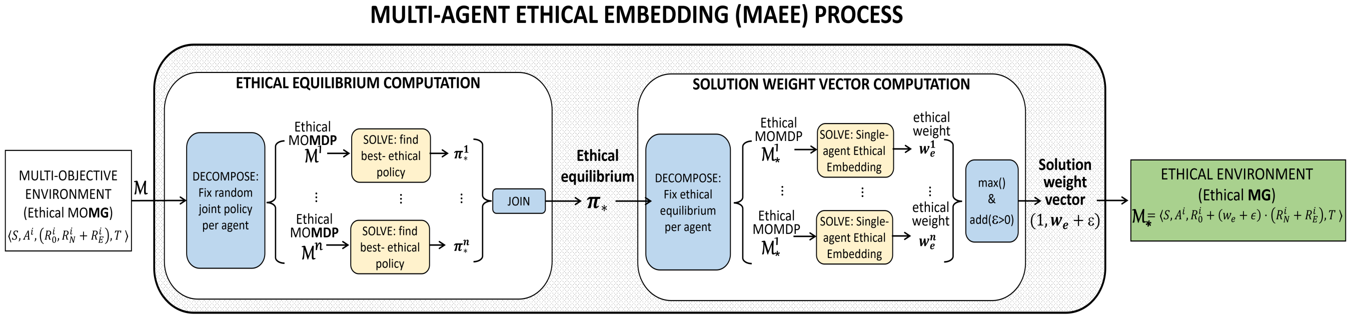
MULTI-AGENT ETHICAL EMBEDDING (MAEE) PROCESS
MULTI-OBJECTIVE ENVIRONMENT (Ethical MOMG)
$\langle S, A^i, (R_0^i, R_N^i + R_E^i), T \rangle$
$\mathcal{M}$
ETHICAL EQUILIBRIUM COMPUTATION
DECOMPOSE: Fix random joint policy per agent
Ethical MOMDP $\mathcal{M}^1$
Ethical MOMDP $\mathcal{M}^n$
SOLVE: find best- ethical policy
SOLVE: find best- ethical policy
π_*^1
π_*^n
JOIN
Ethical equilibrium π_*
SOLUTION WEIGHT VECTOR COMPUTATION
DECOMPOSE: Fix ethical equilibrium per agent
Ethical MOMDP $\mathcal{M}_*^1$
Ethical MOMDP $\mathcal{M}_*^1$
SOLVE: Single-agent Ethical Embedding
SOLVE: Single-agent Ethical Embedding
ethical weight w_e^1
ethical weight w_e^n
max() & add(ε>0)
Solution weight vector $(1, w_e + \varepsilon)$
ETHICAL ENVIRONMENT (Ethical MG)
$\mathcal{M}_* = \langle S, A^i, R_0^i + (w_e + \epsilon) \cdot (R_N^i + R_E^i), T \rangle$

Figure 1: Multi-Agent Ethical Embedding process for environment design. Rectangles stand for objects whereas rounded rectangles correspond to processes. Process steps (from left to right): Computation of the ethical equilibrium from the input multi-objective environment; and computation of the solution ethical weight that creates an output ethical (single-objective) environment.

ethical environments. Subsequently, Section 5 illustrates our approach with an example environment. Finally, Section 6 concludes and sets paths to future work.

2 FORMALISING THE MAEE PROBLEM

In Ethics, a moral value (or ethical principle) expresses a moral objective worth striving for [30]. Following [18], current approaches to align agents with a moral value propose: (i) the specification of rewards to actions aligned with a moral value, and (ii) an embedding that ensures that all agents learn to behave ethically (in alignment with the moral value). Since we tackle the specification process in [19], here we focus on the embedding and we assume that individual and ethical rewards are specified as a Multi-Objective Markov Game (MOMG) [1] [25] (see definition below). In particular, since the rewards in this given MOMG consider a social moral value, we refer to it as an *ethical MOMG*, and define it as a two-objective learning environment where rewards represent both the individual objective of each agent and the social[2] ethical objective (i.e., the moral value). Then, the purpose of the multi-agent ethical embedding problem, which we formalise below, is that of transforming an ethical MOMG into a single-objective MG wherein it is ensured that all agents learn to fulfil a social ethical objective while pursuing their individual objectives. We start by formalising an environment for n agents and m objectives as an MOMG.

Definition 1. *A (finite) m-objective Markov Game (MOMG) of n agents is defined as a tuple $\langle \mathcal{S}, \mathcal{A}^{i=1,\ldots,n}, \vec{R}^{i=1,\ldots,n}, T \rangle$ where: $\mathcal{S}$ is a (finite) set of states; $\mathcal{A}^i(s)$ is the set of actions available at state s for agent i; $\vec{R}^i = (R_1^i, \ldots, R_m^i)$ is a* vectorial *reward function with each R_j^i being the associated scalar reward function of agent i for objective $j \in \{1, \ldots, m\}$; and T is a transition function that, taking into account the current state s and the joint action of all the agents, returns a new state.*

Each agent i of an MOMG has its associated multi-dimensional state value function $\vec{V}^i = (V_1^i, \ldots, V_m^i)$, where each V_j^i is the expected sum of rewards for objective j of agent i. Moreover, given an MOMG $\mathcal{M}$, if we enforce all the agents but i to follow a fixed joint policy π^{-i}, we obtain an MOMDP[1] $\mathcal{M}^i$ for agent i.

Now, we define an *ethical MOMG* as a two-objective Markov game encoding the reward specification of both the agents' individual objectives and the social ethical objective (i.e., the moral value). Following the Ethics literature [5, 6] and recent work in single-agent ethical embedding processes [18], we define an ethical objective through two dimensions: (i) a *normative dimension*, which punishes the violation of moral requirements; and (ii) an *evaluative dimension*, which rewards morally praiseworthy actions. Formally:

Definition 2 (Ethical MOMG). *We define an* Ethical MOMG *as any n-agent MOMG*

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^{i=1,\ldots,n}, (R_0, R_\mathcal{N} + R_E)^{i=1,\ldots,n}, T \rangle, \qquad (1)$$

such that for each agent i: $R_\mathcal{N}^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^-$ penalises violating moral requirements; and $R_E^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^+$ positively rewards performing praiseworthy actions.

We define R_0^i, $R_\mathcal{N}^i$ and R_E^i as the individual, normative and evaluative reward functions of agent i, respectively. We refer to $R_e^i = R_\mathcal{N}^i + R_E^i$ as the ethical reward function and impose agents to be equally treated when assigning the (social) ethical rewards.

Although actions cannot be rewarded and punished simultaneously, having a two-fold ethical reward prevents agents from learning to disregard some of its normative requirements while learning to perform as many praiseworthy actions as possible. Moreover, the equal treatment homogenises what is considered as praiseworthy or blameworthy. Thus, it ensures that the ethical objective is indeed *social*. Finally, also notice that a single-agent Ethical MOMG corresponds to an Ethical MOMDP as defined in [18].

Within ethical MOMGs, we define the *ethical policy* π^i for an agent i as that maximising the ethical objective subject to the behaviour of the other agents (i.e., their joint policy π^{-i}). This maximisation is performed over the normative and evaluative components of agent i's value function:

[1] MOMGs generalise other widely used specifications [25]: an MOMG with a single objective corresponds to a Markov game (MG); and single-agent MOMG corresponds to a Multi-Objective Markov Decision Process (MOMDP).

[2] *Social* in the sense that it is shared by all agents in the system. We take this social stance because moral values are widely assumed to stem from the society as they are defined as "ideals shared by members of a culture about what is good or bad" [8].

Definition 3 (Ethical policy). *Let $\mathcal{M}$ be an ethical MOMG. A policy π^i of agent i is said to be ethical in $\mathcal{M}$ with respect to π^{-i} if and only if the value vector of agent i for the joint policy $\langle\pi^i, \pi^{-i}\rangle$, is optimal for its social ethical objective (i.e., its normative $V^i_{\mathcal{N}}$ and evaluative V^i_E components):*

$$V^i_{\mathcal{N}_{\langle\pi^i,\pi^{-i}\rangle}} = \max_{\rho^i} V^i_{\mathcal{N}_{\langle\rho^i,\pi^{-i}\rangle}},$$
$$V^i_{E_{\langle\pi^i,\pi^{-i}\rangle}} = \max_{\rho^i} V^i_{E_{\langle\rho^i,\pi^{-i}\rangle}}.$$

Ethical policies pave the way to characterise our target policies: *best-ethical* policies. These maximise pursuing the individual objective while ensuring (prioritising) the fulfilment of the ethical objective. Thus, from the set of ethical policies, we define as *best* those maximising the individual value function V^i_0 (i.e., the accumulation of rewards R^i_0):

Definition 4 (Best-ethical policy). *Let $\mathcal{M}$ be an Ethical MOMG. We say that a policy π^i of agent i is* best-ethical *with respect to π^{-i} if and only if it is maximal in V^i_0 in the set $\Pi_e(\pi^{-i})$ of ethical policies with respect to π^{-i}:*

$$V^i_{0_{\langle\pi^i,\pi^{-i}\rangle}} = \max_{\rho^i \in \Pi_e(\pi^{-i})} V^i_{0_{\langle\rho^i,\pi^{-i}\rangle}}.$$

In the MARL literature, if the policy π^i of each agent i is a best-response (i.e., optimal with respect to π^{-i}), then the joint policy $\pi = (\pi^1, \dots, \pi^n)$ that they form is called a **Nash equilibrium** [14]. Here we propose two similar equilibrium concepts for ethical MOMGs. First, if the policy π^i of each agent of an Ethical MOMG is ethical, we say that the joint policy π is an **ethical equilibrium**. Second, if each π^i is best-ethical, then we say that π is a **best-ethical equilibrium**.

Our approach consists in transforming the Ethical MOMG into a single-objective MG by means of what we call a multi-agent *embedding* function. In this way, the agents can learn by applying single-objective multi-agent reinforcement learning algorithms (MARL) [25]. In the Multi-objective literature, an embedding function receives the name of *scalarisation* function [25].

Therefore, our goal is to find an embedding function f_e that guarantees that it is only possible for the agents to learn ethical policies in the ethical environment (the single-objective Markov Game created from applying f_e). Formally, such f_e must ensure that best-ethical equilibria in the Ethical MOMG correspond with Nash equilibria in the single-objective MG created from f_e. For that reason, we refer to the MOMG scalarised by f_e as the *ethical MG*. In its simplest form, this embedding function f_e will have the form of a linear combination of individual and ethical objectives for each agent i:

$$f^i_e(\vec{V}^i) = \vec{w}^i \cdot \vec{V}^i = w^i_0 V^i_0 + w^i_e(V^i_{\mathcal{N}} + V^i_E), \tag{2}$$

where $\vec{w}^i \doteq (w^i_0, w^i_e)$ is a weight vector with all weights $w^i_0, w^i_e > 0$ to guarantee that each agent i takes into account all rewards (i.e., both objectives). Without loss of generality, hereafter we fix the individual weight of all agents to $w^i_0 = 1$. We refer to any linear f^i_e by its weight vector $\vec{w}^i$. Furthermore, since in an Ethical MOMG we have an equal treatment condition for all agents, we deem appropriate that all agents have the same scalarisation function with the same ethical weight, namely $w_e \doteq w^1_e = \cdots = w^n_e$.

Finally, we can formalise our multi-agent ethical embedding problem as that of computing a weight vector $\vec{w} \doteq (1, w_e)$ that incentivises all agents to behave ethically while still pursuing their respective individual objectives. Formally:

Problem 1 (MAEE: Multi-Agent Ethical Embedding). *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R^i_0, R^i_{\mathcal{N}} + R^i_E), T\rangle$ be an ethical MOMG. The multi-agent ethical embedding problem is that of computing the vector $\vec{w}$ of positive weights such that all Nash equilibria in the Markov Game $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, R_0 + w_e(R_{\mathcal{N}} + R_E), T\rangle$ are best-ethical equilibria in $\mathcal{M}$.*

Any weight vector $\vec{w}$ with positive weights that guarantees that all Nash equilibria (with respect to $\vec{w}$) are also best-ethical equilibria is a solution to Problem 1. Unfortunately, Problem 1 is not solvable for every Ethical MOMG. Therefore, in what follows we characterise the Ethical MOMGs for which a solution exists (and a way of computing it).

3 SOLVABILITY OF THE MAEE PROBLEM

This section is devoted to describing the minimal conditions under which there always exists a solution to Problem 1 for a given ethical MOMG, and to proving that such solution actually exists. This solution (a weight vector) will allow us to apply the ethical embedding process to the ethical MOMG at hand to produce an ethical environment (a single-objective MG) wherein agents learn to behave ethically (i.e., to reach a best-ethical equilibrium). In what follows, Subsection 3.1 characterises a family of ethical MOMGs for which Problem 1 can be solved, and Subsection 3.2 proves that the solution indeed exists for such family.

3.1 Characterising solvable Ethical MOMGs

We introduce below a new equilibrium concept for ethical MOMGs that is founded on the notion of dominance in game theory, the so-called *best-ethically-dominant* equilibrium. We find such equilibrium in environments where the best behaviour for each agent is to follow an ethical policy, provided that the ethical weight is properly set. The existence of such equilibria is important to characterise the ethical MOMGs for which we can solve the MAEE problem (Problem 1). Thus, as shown below in Section 3.2, we can only solve Problem 1 for Ethical MOMGs with a best-ethically-dominant equilibrium.

We extract our concept of dominance from the game theory literature. In a Markov game context, we say that a policy π^i of agent i is dominant if it yields the best outcome for agent i no matter the policies that the other agents follow. Then, we say that the dominant policy π^i *dominates* over all possible policies [15]:

Definition 5 (Dominant policy). *Given a Markov game $\mathcal{M}$, a policy π^i of agent i is a dominant policy if and only if for every joint policy $\langle\rho^i, \rho^{-i}\rangle$ and every state s in which $\rho^i(s) \neq \pi^i(s)$ it holds that:*

$$V^i_{\langle\pi^i,\rho^{-i}\rangle}(s) \geq V^i_{\langle\rho^i,\rho^{-i}\rangle}(s). \tag{3}$$

A policy is **strictly** dominant if we change ≥ to >. Moreover, if the policy π^i of each agent of an MG is a dominant policy, we say that the joint policy $\pi = (\pi^1, \dots, \pi^n)$ is a **dominant equilibrium**.

Now we adapt the concept of dominance in game theory for Ethical MOMGs. We start by defining policies that are dominant with respect to the ethical objective. We call these policies ethically-dominant policies. Formally:

Definition 6 (Ethically-dominant policy). *Let $\mathcal{M}$ be an ethical MOMG. We say that a policy π^i of agent i is an ethically-dominant policy in $\mathcal{M}$ if and only if the policy is dominant for its ethical objective (i.e., both its normative $V^i_{\mathcal{N}}$ and evaluative V^i_E components) for every joint policy $\langle \rho^i, \rho^{-i} \rangle$ and every state s in which $\rho^i(s) \neq \pi^i(s)$:*

$$V^i_{\mathcal{N}_{\langle \pi^i, \rho^{-i} \rangle}}(s) \geq V^i_{\mathcal{N}_{\langle \rho^i, \rho^{-i} \rangle}}(s),$$
$$V^i_{E_{\langle \pi^i, \rho^{-i} \rangle}}(s) \geq V^i_{E_{\langle \rho^i, \rho^{-i} \rangle}}(s).$$

Following Def. 3, every ethically-dominant policy π^i is an ethical policy with respect to any π^{-i}.

We also adapt the concept of dominance from game theory for best-ethical policies. Given an ethical MOMG, we say that a best-ethically-dominant policy is: (1) dominant with respect to the ethical objective among all policies; and (2) dominant with respect to the individual objective among ethical policies. Formally:

Definition 7 (Best-ethically-dominant policy). *Let $\mathcal{M}$ be an Ethical MOMG. A policy π^i of agent i is a* best-ethically-dominant *policy if and only if it is ethically-dominant and*

$$V^i_{0_{\langle \pi^i, \rho^{-i} \rangle}}(s) \geq V^i_{0_{\langle \rho^i, \rho^{-i} \rangle}}(s),$$

for every joint policy $\langle \rho^i, \rho^{-i} \rangle$, every state s in which $\rho^i(s) \neq \pi^i(s)$, and $\rho^i(s)$ is an ethical policy with respect to $\rho^{-i}(s)$.

Observe that by Def. 4, every best-ethically-dominant policy π^i is also a best-ethical policy with respect to any π^{-i}.

Next, we define the generalisation of the two previous dominance definitions considering the policies of all agents. This will lead to new equilibrium concepts. If the policy π_i of each agent of an Ethical MOMG is ethically-dominant, then the joint policy $\pi = (\pi^1, \ldots, \pi^n)$ is an **ethically-dominant (ED)** equilibrium. If the policies are best-ethically-dominant, then the joint policy π is a **best-ethically-dominant** (BED) equilibrium. Observe that every ethically-dominant equilibrium is an ethical equilibrium, and every best-ethically-dominant equilibrium is a best-ethical equilibrium.

Finally, observe that a joint policy $\pi = (\pi^1, \ldots, \pi^n)$ is a **strictly** best-ethically-dominant equilibrium if and only if every π^i is strictly dominant with respect to the individual objective among ethical policies (i.e., by changing $\geq$ with $>$ in Def. 7).

In the following subsection we prove that we can solve the multi-agent ethical embedding problem (Problem 1) for Ethical MOMGs with a best-ethically-dominant equilibrium.

3.2 On the existence of solutions

Next we prove that we can find a multi-agent ethical embedding function for ethical MOMGs with best-ethically-dominant equilibria. Henceforth, we shall refer to such ethical MOMGs as *solvable* ethical MOMGs, and if the BED equilibrium is strict, we will refer to such Ethical MOMGs as *strictly-solvable.* Below, we present Theorem 1 as our main result. The theorem states that given a solvable ethical MOMG, it is always possible to find an embedding function that transforms it into a (single-objective) MG where agents are guaranteed to learn to behave ethically. More in detail, following Theorem 1 guarantees that for the appropriate ethical weight in our embedding function, then the only Nash equilibria in the environment where the agents learn (the scalarised MOMG produced thanks to the embedding function) are best-ethical equilibria in the Ethical MOMG. In other words, such embedding function is the solution to Problem 1 we are looking for.

The proof of Theorem 1 requires the introduction of some propositions as intermediary results. The first proposition establishes the relationship between dominant and ethically-dominant policies.

Proposition 1. *Given an ethical MOMG $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R_0, R_{\mathcal{N}} + R_E)^i, T \rangle$ for which there exists ethically-dominant equilibria, there exists a weight vector $\vec{w} = (1, w_e)$ with $w_e > 0$ for which every dominant policy for an agent i in the MG $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}^i, w_0 R^i_0 + w_e(R^i_{\mathcal{N}} + R^i_E), T \rangle$ is also an ethically-dominant policy for agent i in $\mathcal{M}$.*

Proof. Without loss of generality we only consider deterministic policies, by the Indifference Principle [15].

Consider a weight vector $\vec{w} = (1, w_e)$ with $w_e \geq 0$. Suppose that for that weight vector, the only deterministic $\vec{w}$-dominant policies (i.e. policies that are dominant in the MOMG scalarised by $\vec{w}$) are ethically-dominant. Then we have finished.

Suppose now that it is not the case, and there is some $\vec{w}$-dominant policy ρ^i for some agent i that is not dominant ethically. This implies that for some state s' and for some joint policy ρ^{-i} we have that:

$$V^i_{\mathcal{N}_{\langle \rho^i, \rho^{-i} \rangle}}(s') + V^i_{E_{\langle \rho^i, \rho^{-i} \rangle}}(s') < V^i_{\mathcal{N}_{\langle \pi^i, \rho^{-i} \rangle}}(s') + V^i_{E_{\langle \pi^i, \rho^{-i} \rangle}}(s'),$$

for any ethically-dominant policy π^i for agent i.

For an $\epsilon > 0$ large enough and for the weight vector $\vec{w}' = (1, w_e + \epsilon)$, any ethically-dominant policy π^i will have a better value vector at that state s' than ρ^i against ρ^{-i}:

$$\vec{w}' \cdot \vec{V}^i_{\langle \rho^i, \rho^{-i} \rangle}(s') < \vec{w}' \cdot \vec{V}^i_{\langle \pi^i, \rho^{-i} \rangle}(s').$$

Therefore, ρ^i will not be a $\vec{w}'$-dominant policy. Notice that ρ will remain without being dominant even if we increase again the value of w_e by defining $\vec{w}'' = (1, w_e + \epsilon + \delta)$ with $\delta > 0$ as large as we wish.

Now consider the policy ρ^i not ethically-dominant that requires the maximum $\epsilon_* > 0$ in order to stop being $\vec{w}$-dominant. We can guarantee that this policy exists because there is a finite number of deterministic policies in a finite MOMG. Therefore, by selecting the weight vector $\vec{w}_* = (1, w_e + \epsilon_*)$, then only ethically-dominant policies can be $\vec{w}_*$-dominant for this ethical weight $w_e + \epsilon_*$. In other words, every $\vec{w}_*$-dominant policy is also ethically-dominant for this new weight vector. □

The former proposition helps us establish a formal relationship between dominant equilibria and ethically-dominant equilibria through the following proposition.

Proposition 2. *Given an ethical MOMG $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R_0, R_{\mathcal{N}} + R_E)^i, T \rangle$ for which there exists best-ethically-dominant equilibria, there exists a weight vector $\vec{w} = (1, w_e)$ with $w_e > 0$ for which every dominant policy for an agent i in the Markov Game $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}^i, R^i_0 + w_e(R^i_{\mathcal{N}} + R^i_E), T \rangle$ is also a best-ethically-dominant policy for agent i in the ethical MOMG $\mathcal{M}$ and vice-versa.*

Proof. By Proposition 1, there is an ethical weight for which every dominant policy in $\mathcal{M}'$ is ethically-dominant in $\mathcal{M}$.

Best-ethically-dominant policies dominate all ethically-dominant policies, and thus every dominant policy in $\mathcal{M}'$ is in fact a best-ethically-dominant in $\mathcal{M}$. In other words, at least one best-ethically-dominant policy is dominant for this ethical weight.

Finally, since every best-ethically-dominant policy π^i is best-ethical against any other joint policy ρ^{-i}, for a large enough ethical weight, it will be a best-response against any other joint policy ρ^{-i}, which is the definition of a dominant policy. □

Thanks to Proposition 2 we are ready to formulate and prove Theorem 1 as follows.

Theorem 1 (Multi-agent solution existence (dominance)). *Given an ethical MOMG $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R_0, R_{\mathcal{N}} + R_E)^i, T \rangle$ for which there exists at least one best-ethically-dominant equilibrium π_*, then: (1) there exists a weight vector $\vec{w} = (1, w_e)$ with $w_e > 0$ for which π_* is a dominant equilibrium in the scalarised MOMG $\mathcal{M}_*$ by $\vec{w}$; and (2) for the weight vector $\vec{w}' = (1, w_e + \epsilon)$ with $\epsilon > 0$ all Nash equilibria in the scalarised MOMG $\mathcal{M}'_*$ by $\vec{w}'$ are best-ethically-dominant equilibria in the ethical MOMG $\mathcal{M}$.*

Proof. By Proposition 2, there exists a weight vector $\vec{w} = (1, w_e)$ for which all best-ethically-dominant equilibria are dominant equilibria, and thus, Nash equilibria.

Let π_* be any best-ethically-dominant equilibrium. If for such ethical weight w_e there is another Nash equilibrium ρ that is not best-ethically-dominant, it cannot be either an ethically-dominant equilibrium because π_* would have more value for some state s' and some agent i (in which $\rho^i(s')$ is not best-ethical).

Thus, this other Nash equilibrium necessarily has less ethical value for some state s'' and for some agent j than π^i_* (in which $\rho^j(s'')$ is not an ethical). Therefore, ρ^j will stop being a best-response against ρ^{-j} as soon as we increase the ethical weight by any $\epsilon > 0$ (because π^j_* is already a best-response against ρ^{-j} for being dominant). This will cause that ρ will stop being a Nash equilibrium.

Once a joint policy ρ is not a Nash equilibrium for some ethical weight w_e, it is no longer a Nash equilibrium for any greater $w'_e > w_e$, so the only remaining Nash Equilibria are best-ethically-dominant equilibria. □

Theorem 1 guarantees that we can solve Problem 1 for any Ethical MOMG with at least one best-ethically-dominant equilibrium. Indeed, for that reason we call such family of Ethical MOMGs as *solvable*. In particular, we aim at finding solutions $\vec{w}$ that are as little intrusive with the agent's learning process as possible (i.e., the $\vec{w}$ that guarantees the learning of an ethical policy with the minimal ethical weight w_e). Every time we refer to the *minimal* ethical weight we do it in such sense.

4 SOLVING THE MAEE PROBLEM

This section details how to solve the Multi-Agent Ethical Embedding (MAEE) problem defined by Problem 1. This amounts to computing a solution weight vector $\vec{w}$ so that we can combine individual and ethical rewards into a single reward to yield a new, ethical environment.

Figure 1 illustrates our approach to solving a MAEE problem, which follows two main steps: (1) computation of a best-ethical equilibrium (the *target* joint policy), namely the joint policy that we expect the agents to converge to when learning in our ethical environment; and (2) computation of a solution weight vector $\vec{w}$ based on the target joint policy. Interestingly, both computations are based on *decomposing* the ethical MOMG (the input to the problem) into n ethical MOMDPs (one per agent), *solving* one local problem (MOMDP) per agent, and *aggregating* the resulting solutions.

In what follows we provide the theoretical grounds for computing a target joint policy and a solution weight vector. For the remainder of this Section we assume that there exists a *strictly best-ethically-dominant equilibrium* in the Ethical MOMG, that is, that the Ethical MOMG is *strictly-solvable.*

4.1 Computing the ethical equilibrium

As previously mentioned, we require to compute the best-ethical equilibrium π_* to which we want agents to converge when learning in our ethical environment. Figure 1 (Left) illustrates the three steps required to compute such joint policy.

In order to obtain the joint policy π_*, we can resort to decomposing the ethical MOMG $\mathcal{M}$, encoding the input multi-objective environment, into n ethical MOMDPs $\mathcal{M}^{i=1,\dots,n}$, one per agent. For each ethical MOMDP $\mathcal{M}^i$, we compute the individual contribution of agent i to the ethical equilibrium (π^i_*) with single-agent reinforcement learning.

Building the ethical equilibrium π_* via decomposition is possible because we assume that the ethical MOMG $\mathcal{M}$ is strictly-solvable, hence satisyfing the conditions of Theorem 1. This means that the best-ethical equilibrium π_* is also strictly dominant. Thus, each agent has one (and only one) strictly best-ethically-dominant policy (π^i_*), which by Def. 7 is the unique best-ethical policy against any other joint policy.

There is the issue of how to decompose the ethical MOMG $\mathcal{M}$. Since we know that each π^i_* is the only best-ethical policy against any other joint policy ρ^{-i}, we can select randomly ρ^{-i} to create an Ethical MOMDP $\mathcal{M}^i$ for each agent i.

After creating all Ethical MOMDPs $\mathcal{M}^{i=1,\dots,n}$, we must compute the policy π^i_* of each agent i as its best-ethical policy in the ethical MOMDP $\mathcal{M}^i$. We can do this using multi-objective single-agent RL. In particular, we apply the Value Iteration (VI) algorithm with a *lexicographic* ordering [29] (prioritising ethical rewards), since it is has the same computational cost as VI. Finally, we join all best-ethical policies π^i_* to yield the joint policy π_*.

4.2 Computing the solution weight vector

Once computed the ethical equilibrium π_*, we can proceed to compute the corresponding ethical weight w_e that guarantees that π_* is the only Nash equilibrium in the ethical environment (the scalarised MOMG) produced by our embedding. Figure 1 (Right) illustrates the steps required to compute it.

Similarly to Section 4.1, we compute w_e by decomposing the input environment (the ethical MOMG $\mathcal{M}$) into n ethical MOMDPs $\mathcal{M}^{i=1,\dots,n}_*$. Thereafter, we compute the individual ethical weight w^i_e for each ethical MOMDP. Finally, we aggregate the individual ethical weights to obtain w_e.

In this case, we create each Ethical MOMDP $\mathcal{M}^i_*$ by fixing the best-ethical equilibrium π^{-i}_* for all agents but i. Then, computing the ethical weight for each ethical MOMDP amounts to solving a Single-Agent Ethical Embedding (SAEE) problem as introduced in [18]. With this aim, we can employ the algorithm introduced in that work (details of SAEE are provided in the Supplementary Work) to solve all SAEE problems. Afterwards, we obtain an individual *ethical weight* w_e^i that ensures that each agent i will learn to behave ethically (following π_*^i) in the ethical MOMDP $\mathcal{M}_*^i$.

Finally, we select the greatest ethical weight $w_e = \max_i w_e^i$ to obtain the solution weight vector $(1, w_e + \epsilon)$ that solves our problem (where $\epsilon > 0$). The above-described procedure to produce an ethical environment (based on decomposing, individually solving single-agent embedding problems, and aggregating their results) does guarantee that agents will learn to behave ethically in such environment.

Notice that the cost of computing the solution weight vector mainly resides in applying n times the SAEE algorithm [18], once per agent. Following [18], the cost of such algorithm is largely dominated by the computational cost of the Convex Hull Value Iteration algorithm [3].

Our approach above requires that Ethical MOMGs fulfil the following condition: although the ethical objective is social, it is enough that a fraction of the agents (not all of them) intervene to completely fulfil it. To give an example inspired on the Ethics literature, consider a situation where several agents are moving towards their respective destination through a shallow pond and at some point a child that cannot swim falls into the water (similarly to the Drowning Child Scenario from [26]). To save the child, it is enough that one agent takes a dive to rescue them. Another example, from the AI literature, is the Cleanup Game from [10], in which a handful of agents need to stop collecting apples from time to time to repair the aquifer supplying water.

In terms of the ethical weight w_e, this assumption implies that the hardest situation to incentivise an agent i to follow an ethically-dominant policy π_*^i occurs when the rest of agents already behave ethically by following an ethical equilibrium π_*^{-i}. By definition, for such w_e the policy π_*^i will be a best-response against any joint policy, thus becoming a dominant policy. In other words, the ethical weight required to guarantee that π_*^i is dominant is the same as the ethical weight required to guarantee that π_*^i is a best-response against π_*^{-i}. This is formally captured by the next condition:

Condition 1. *Let $\mathcal{M}$ be an ethical MOMG $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R_0, R_{\mathcal{N}} + R_E)^i, T\rangle$ for which there exists at least one best-ethically-dominant equilibrium π_*. Consider the weight vector $\vec{w} = (1, w_e)$ and the scalarised MOMG $\mathcal{M}_* = \langle \mathcal{S}, \mathcal{A}^i, R_0 + w_e \cdot (R_{\mathcal{N}} + R_E)^i), T\rangle$. We require that if the best-ethically-dominant equilibrium π_* is a Nash equilibrium in $\mathcal{M}_*$, then π_* is also a dominant equilibrium in $\mathcal{M}_*$.*

Now we can proceed with proving the soundness of our method for computing a solution weight vector. First, we notice what implies for the weight vector to have the minimal ethical weight w_e necessary for π_* to be a Nash equilibrium. From an agent perspective, this implies that such ethical weight has to be the minimum one that guarantees that each π_*^i is a best-response. Formally:

Observation 1. *Given any joint policy π, then the minimum ethical weight w_e for which every policy π^i is a best-response against π^{-i} is also the minimum ethical weight w_e for which π is a Nash equilibrium.*

Second, the following Theorem tells us the minimal w_e necessary for π_* to be the *only* Nash equilibrium: any ethical weight slightly greater than the one that guarantees that π is a Nash equilibrium.

Theorem 2. *Given an ethical MOMG $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^i, (R_0, R_{\mathcal{N}} + R_E)^i, T\rangle$ for which there exists at least one best-ethically-dominant equilibrium π_* and for which Condition 1 holds, if for a weight vector $\vec{w} = (1, w_e)$ with $w_e > 0$ we have that π_* is a Nash equilibrium, then for $\vec{w}' = (1, w_e + \epsilon)$ it is the unique Nash equilibrium (except for other best-ethically-dominant equilibria).*

Proof. By Condition 1, we have that the ethically-dominant-Nash equilibrium π_* is also dominant for the weight vector $\vec{w} = (1, w_e)$. By Theorem 1, once we increase the ethical weight w_e by some $\epsilon > 0$ as small as we want, the only Nash equilibria will be those joint policies that are ethically-dominant-Nash equilibria as well. □

Finally, to compute the solution weight vector $(1, w_e)$ we resort to Observation 1 and Theorem 2. First, we consider the n different ethical weights $w_{e_1}, \ldots, w_{e_n}$ that we obtain by applying the single-agent ethical embedding algorithm for each agent individually (the second step of our solution weight vector computation in Figure 1), and we select the one with maximum value:

$$w_e = \max_i w_e^i. \tag{4}$$

Afterwards, we add a small $\epsilon > 0$ to its value $w_e' = w_e + \epsilon$ in order to guarantee that for w_e', the best-ethically-dominant equilibrium π_* is the only Nash equilibrium. The weight vector $\vec{w} = (1, w_e')$ is our desired ethical embedding that guarantees that the Markov Game $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}^i, R_0^i + (w_e + \epsilon)[R_{\mathcal{N}}^i + R_E^i], T\rangle$ is an ethical environment. In other words, $\vec{w}$ solves the multi-agent ethical embedding problem.

5 EXAMPLE: THE PUBLIC CIVILITY GAME

This section illustrates our process to design an ethical environment (outlined in Figure 1) through an example: the *Public Civility Game* (PCG) [16]. The PCG is a value alignment problem where two agents must learn to behave according to the moral value of civility. In [18], the authors solve a limited version of the problem where just one agent learns to behave ethically (the other agent follows a fixed policy). Instead, here we design an ethical environment for *all* the agents to learn.

Figure 2a depicts the PCG grid environment where agents (L)eft and (R)ight move from their initial positions to their goal destinations (GL and GR respectively). At the beginning of each episode, garbage (small purple square) can appear blocking the way of any of the two agents and we expect the agent to learn to handle it civically while moving towards its goal. The desired civic (ethical) behaviour is to push the garbage to the bin without throwing it at each other even though this requires additional effort. Next, we detail the processes to obtain an ethical environment for the Public Civility Game and the policies that the agents learn in such environment. For that, we borrow the action set (move and push) and the reward function specified in [18] to build the input environment (MOMG) for the multi-agent embedding: each agent receives an individual reward ($R_0^i = 20$) for reaching its destination (otherwise,

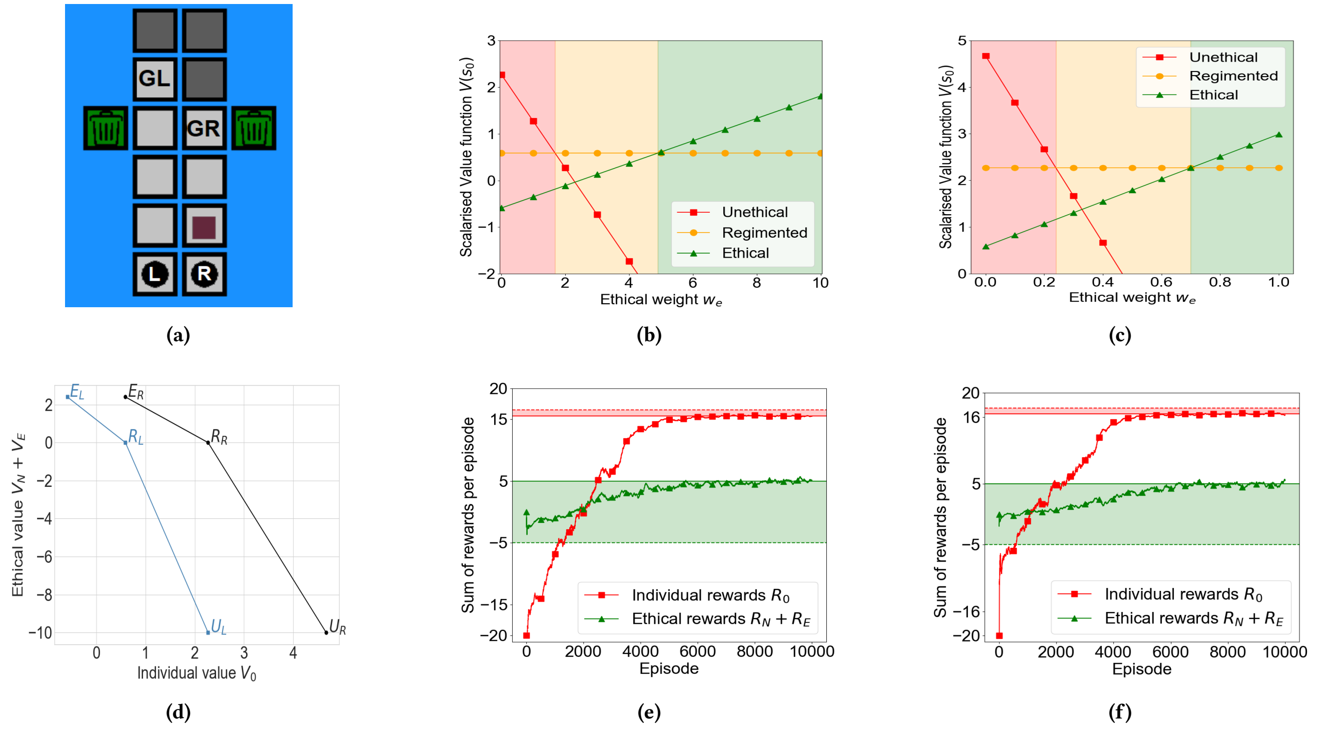

Figure 2: (a) A possible initial state of the public civility game. The agent on the right (R) has to deal with the garbage obstacle, which has been located in front of it. (b) Visualisation in Weight Space of the partial convex hull for agent L. (c) Visualisation in Weight Space of the partial convex hull for agent R. (d) Visualisation in Objective Space of the partial convex hull of each agent (L and R), composed by 3 policies per agent: E (Ethical), R (Regimented) and U (Unethical). (e) Evolution of the accumulated rewards per episode that the Left agent obtains in the ethical environment. Horizontal dotted lines mark convergence values for an unethical policy, and straight lines for an ethical policy (red for individual rewards, green for ethical rewards). (f) Evolution of the accumulated rewards per episode of the Right agent.

$R_0^i = -1$); a penalty ($R_N^i = -10$) for pushing the garbage to a position occupied by another agent; and an ethical incentive ($R_E^i = 10$) for pushing the garbage to a bin.

5.1 Building the ethical environment

5.1.1 Ethical equilibrium computation. Following Section 4.1 (and left side of Figure 1), we first decompose the input ethical MOMG into two ethical MOMDPs $\mathcal{M}^{i \in \{L,R\}}$. We build each $\mathcal{M}^i$ by randomly fixing the policy of agent $j \neq i$. Then, we let agent i learn its individual *ethical* policy π_*^i by applying Value Iteration with lexicographic ordering [29]. Policy π_*^i has agent i throw the garbage to a bin whenever found in front. It is worth noting that the vectorial values $(V_0^{\pi^i}, V_N^{\pi^i} + V_E^{\pi^i})$ of these policies are (-0.59, 0 + 2.4) for π_*^L and (0.59, 0 + 2.4) for π_*^R, which amount to an accumulation of rewards $\sum \vec{R}^L = (15.5, 0 + 5)$ and $\sum \vec{R}^R = (16.5, 0 + 5)$. Finally, we obtain the ethical equilibrium by joining the ethical policies of both agents: $\pi_* = (\pi_*^L, \pi_*^R)$.

5.1.2 Solution weight vector computation. Following Section 4.2 (and right side of Figure 1), the next step is to compute an ethical weight for which π_* is a Nash equilibrium. Recall that our approach is a three-step process based on (1) *decomposing* the ethical MOMG in several ethical MOMDPs, (2) *solving* one local ethical MOMDP per agent, and finally, (3) *aggregating* the resulting solutions. Specifically, we proceed as follows:

Step 1: Decompose. Following Observation 1, we need to apply the single-agent Ethical Embedding algorithm for each agent, by creating two Ethical MOMDPs. We first create the Ethical MOMDP $\mathcal{M}_*^L$ by fixing the policy of agent R as π_*^R. Then, we repeat the process for agent R by fixing the policy for agent L to create the ethical MOMDP $\mathcal{M}_*^R$.

Step 2: Solve. Thereafter, we apply single-agent Ethical Embedding for $\mathcal{M}_*^L$ and $\mathcal{M}_*^R$. The single-agent ethical embedding process has three phases [18]. First, it computes the partial *convex hull*[3] of the Ethical MOMDP. Afterwards, it extracts the two most ethical policies from the hull. Finally, it makes use of the two policies to compute an ethical weight that guarantees the learning of ethical policies. Next, we explain how to apply these three phases of the single-agent Ethical Embedding to the Public Civility Game.

[3] The partial convex hull of an ethical MOMDP, as defined in [18], consists on all the policies that are optimal for some weight vector of the form $(1, w_e)$ with an $w_e > 0$.

Partial convex hull computation: Without loss of generality we start with the ethical MOMDP $\mathcal{M}_*^L$, for which we only explain the results for the initial state s_L (garbage in front of agent L)[4].

Then, considering $\mathcal{M}_*^L$, we compute its partial convex hull $P \subseteq CH(\mathcal{M}_*^L)$. Figure 2d depicts, in blue dots, the resulting convex hull P for agent L in the initial state s_L. It is composed of 3 different policies named after the behaviour they encapsulate:

(1) An **U**nethical (uncivil) policy U_L, in which agent L moves towards the goal and throws away the garbage without caring about any ethical implication.
(2) A **R**egimented policy R_L, in which agent L complies with the norm of not throwing the garbage at the other agent.
(3) An **E**thical policy E_L, which we already found in the previous step and denoted as π_*^L. Following this policy, agent L always throws the garbage to the bin when found in front.

The same three policies appear in the partial convex hull of the ethical MOMDP $\mathcal{M}_*^R$ for agent R (depicted in Figure 2d with black dots). Likewise for agent L, we only explain the results for the initial state s_R for agent R. Notice that the difference between the partial convex hulls of the agents resides only in their respective individual objective values. This difference is to be expected since agent R has its goal position one cell nearer than agent L.

Extraction of the two value vectors with the greatest ethical value: For the two possible initial states (s_L and s_R) the **E**thical policies $\pi_*^L(s_L)$ and $\pi_*^R(s_R)$ have associated the ethical-optimal value vector since they are the policies with greatest ethical value within the partial hull of each agent. As previously mentioned in Section 5.1.1, the vectorial values $(V_0^{\pi^i}, V_{\mathcal{N}}^{\pi^i} + V_E^{\pi^i})$ of these policies are (-0.59, 0 + 2.4) for π_*^L and (0.59, 0 + 2.4) for π_*^R. In order to obtain the ethical weight for which the ethical-optimal policy of each agent is optimal, the single-agent ethical embedding requires another policy, which is the second most ethical one [17].

In this case, the second most ethical value vector for each agent corresponds to the value of their respective **R**egimented policies $\pi_r^L(s_L)$ and $\pi_r^R(s_R)$. As shown by Fig. 2d, the vectorial values $(V_0^{\pi^i}$, $V_{\mathcal{N}}^{\pi^i} + V_E^{\pi^i})$ of the **R**egimented policy are $(0.59, 0 + 0)$ for agent L, and $(2.27, 0 + 0)$ for agent R.

Computation of the ethical weight: Following [18], we need to find the ethical weight for which the **E**thical policy and the **R**egimented policy have the same scalarised value for each agent. For the left agent, its solution is to set the ethical weight as $w_e^L = 0.49$. We can check it: $-0.59 + 0.49 \cdot (0 + 2.4) = 0.59 = 0.59 + 0.49 \cdot (0 + 0)$.

Repeating this process for agent R gives us an ethical weight $w_e^R = 0.7$: $0.59 + 0.7 \cdot (0 + 2.4) = 2.27 = 2.27 + 0.7 \cdot (0 + 0)$.

Figure 2b illustrates the scalarised value of the 3 policies for varying values of w_e^i in (0,1) (for w_e^i>1 tendencies do not change) for agent L, and Figure 2c illustrates the same for agent R. In particular, we observe that the **E**thical policy indeed becomes the only optimal one when $w_e^L > 0.49$ for agent L, and $w_e^R > 0.7$ for agent R. Notice that $w_e^L < w_e^R$ because, since agent L needs to traverse a longer path to reach its destination than R, its accumulated individual reward is smaller, and thus, it also requires a smaller reward to behave ethically.

Step 3: Aggregate. Finally, we select the greatest ethical weight $w_e = \max_i w_e^i = 0.7$ to obtain the solution weight vector $\vec{w} = (1, w_e + \epsilon) = (1, 0.71)$, when setting $\epsilon = 0.01$. The resulting ethical environment $\mathcal{M}_*$ is thus an MG for the PCG with reward function $R_0^i + 0.71 \cdot (R_{\mathcal{N}}^i + R_E^i)$ where the agents are guaranteed to learn policies that jointly form a best-ethical equilibrium.

5.2 Learning in the ethical environment

Despite its simplicity, the PCG allows us to empirically check our formal results. We generate the ethical environment $\mathcal{M}_*$ with two initial states (garbage in front of L or R). As expected, when each agent learns with an independent Q-learner [12], both agents learn to bring the garbage (when found in front) to a bin while moving towards their goals. Figure 2e plots the rewards per episode that agent L accumulates while learning, and Figure 2f plots the rewards accumulated by agent R while learning. Now we briefly comment the accumulated rewards per agent, starting with agent L.

For agent L, the values stabilise at 15.5 for individual rewards and at 5 for ethical rewards so that they match precisely the values of the **E**thical policy $\sum \vec{R}^L$ shown in Section 5.1.1. For reference, we have added as horizontal dotted lines the values of the **U**nethical policy for agent L. The difference in the width of the coloured areas illustrate how, by behaving ethically, agent L increases by 200% its ethical rewards at the cost of decreasing by 6% its individual reward.

For agent R, the values stabilise at 16.5 for individual rewards and at 5 for ethical rewards so that they match precisely the values of the **E**thical policy $\sum \vec{R}^R$ shown in Section 5.1.1. Like for the previous agent, we have added as horizontal dotted lines the values of the **U**nethical policy for agent R. Again, the difference in the width of the coloured areas illustrate how, by behaving ethically, agent R increases by 200% its ethical rewards at the cost of decreasing by 6% its individual reward.

6 CONCLUSIONS AND FUTURE WORK

The literature in value alignment has largely focused on aligning a single agent with a moral value, and with the exception of [18], disregarding guarantees on an agent's ethical learning. Here we have tackled the open problem of building an *ethical* environment for multi-agent systems that guarantees that all the agents in the system learn to behave ethically while pursuing their individual objectives. Our novel contributions are founded in the framework of Multi-Objective Markov Games (MOMGs). First, we characterise a family of MOMGs, the so-called *ethical* MOMGs, for which we can formally guarantee the joint learning of *ethical* equilibria. For such family of MOMGs, we specify the process for building an ethical environment with a so-called *multi-agent ethical embedding* process. Interestingly, our embedding approach for multiple agents generalises that for a single agent in [18].

As future work, we plan to develop methods for testing if an MOMG has ethically-dominant equilibria or not. This is a challenging problem since testing the existence of dominant equilibria in Markov Games is still an open problem [33].

[4] If the garbage is not in front of the agent (as it is the case for agent L in the other initial state s_R), there is a unique policy in the agent's convex hull: to move forward until it gets to its destination.

REFERENCES

[1] 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).

[2] 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

[3] 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

[4] 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.

[5] R. M. Chisholm. 1963. Supererogation and Offence: A Conceptual Scheme for Ethics. *Ratio (Misc.)* 5, 1 (1963), 1.

[6] Amitai Etzioni and Oren Etzioni. 2016. Designing AI Systems That Obey Our Laws and Values. *Commun. ACM* 59, 9 (Aug. 2016), 29–31. https://doi.org/10.1145/2955091

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

[8] Jonathan Haidt. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage.

[9] Conor Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai Irissappane, Patrick Mannion, Ann Nowe, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik Roijers. 2021. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. (03 2021).

[10] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NeurIPS*.

[11] 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.

[12] B. De Schutter L. Busoniu, R. Babuska. 2010. Multi-agent reinforcement learning: An overview. *Innovations in Multi-Agent Systems and Applications – 1* (2010), 183–221.

[13] 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).

[14] Michael L. Littman. 1994. Markov Games As a Framework for Multi-agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (New Brunswick, NJ, USA) *(ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163. http://dl.acm.org/citation.cfm?id=3091574.3091594

[15] Michael Maschler, Eilon Solan, and Shmuel Zamir. 2013. *Game Theory, 2nd Edition*. Cambridge University Press.

[16] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodríguez-Aguilar. 2020. A Structural Solution to Sequential Moral Dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Aystems (AAMAS 2020)*.

[17] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodríguez-Aguilar. 2021. Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning. In *Adaptive and Learning Agents Workshop (AAMAS 2021)*.

[18] 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.

[19] 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

[20] 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.

[21] 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.

[22] 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

[23] 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

[24] Roxana Rădulescu. 2021. *Decision Making in Multi-Objective Multi-Agent Systems: A Utility-Based Perspective*. Ph.D. Dissertation. Vrije Universiteit Brussel.

[25] 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.

[26] Peter Singer. 1972. Famine, Affluence and Morality. *Philosophy and Public Affairs* (1972), 229–243.

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

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

[29] Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. 2021. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence* 100 (04 2021). https://doi.org/10.1016/j.engappai.2021.104186

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

[31] Yueh-Hua Wu and Shou-De Lin. 2017. A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. *arXiv* (12 2017).

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

[33] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. Springer International Publishing, Cham, 321–384. https://doi.org/10.1007/978-3-030-60990-0_12