An approximate process for designing ethical environments with multi-agent reinforcement learning

Arnau Mayoral Macau*
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain
arnau.mayoral@iiia.csic.es

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

Enrico Marchesini
Massachusetts Institute of Technology (MIT)
Massachusetts, US
emarche@mit.edu

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

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

Alessandro Farinelli
Università degli Studi di Verona
Verona, Italy
alessandro.farinelli@univr.it

ABSTRACT

This paper presents an algorithm for designing an environment where multiple autonomous agents learn to behave aligned with a moral value while pursuing their individual objectives. Based on the Multi-Objective Reinforcement Learning and Deep Reinforcement Learning literature, our algorithm represents an extension of the Multi-Agent Ethical Embedding Process (MAEEP), a theoretically grounded method that can be just applied to small problems. We call our method Approximate Multi-Agent Ethical Embedding Process (AMAEEP) and empirically evaluate it in an ethical extension of the gathering game that considers the value of beneficence. Although this environment is much larger than the one used to illustrate the application of the original MAEEP, our method succeeds in dealing with the complexity increase.

1 INTRODUCTION

Autonomous artificial agents are becoming increasingly prevalent [21, 34, 38]. However, as we delegate more tasks –such as autonomous driving or healthcare– to artificial agents [9, 16, 19], we must also be aware of the possible risks or negative ethical effects that may arise, as recognised by the proposed Artificial Intelligence Act [14]. Thus, it is imperative to develop systems to ensure that these agents will always behave in alignment with human values [15, 31, 33]. In fact, the problem of *value alignment* is especially crucial when multiple artificial agents are deployed simultaneously.

In cases where agents must tackle multi-agent decision-making problems, a common approach is to let them learn to behave with reinforcement learning. Multi-agent reinforcement learning (MARL) algorithms have found application in diverse domains, from game playing to autonomous driving and conversational agents [11], exhibiting a notable capacity for acquiring proficiency in intricate tasks. It is no surprise then, that works focusing on applying reinforcement learning to ensure value alignment have recently begun to appear from the fields of Machine Ethics and AI Safety [7, 24].

These fields tackle the problem from different perspectives. The work on AI Safety aims to guarantee that any deployed agent will not cause harm to itself nor the environment (including real humans) [17, 18]. AI Safety techniques both deal with preventing harmful situations during training [5, 13] and after training [2, 6, 40]. While guaranteeing that agents do not perform harmful actions is critical to value alignment, we argue that it is insufficient. As Gabriel argued in [15], behaving value-aligned also implies the proactive role of performing morally good actions such as being kind or altruistic. Meanwhile, the work on Machine Ethics aims to include an ethical dimension that agents must consider while deployed. Following the basics of reinforcement learning, their approach consists in providing extrinsic ethical rewards to the agents to guide them towards a value-aligned behaviour (e.g., [1, 8, 26, 28, 29, 37]). Yet, the majority of Machine Ethics literature is still working on single-agent problems, and, to the best of our knowledge, only Rodriguez-Soto et al. in [30] have started tackling the problem of guaranteeing value alignment for multi-agent systems with reinforcement learning.

In [30], Rodriguez-Soto et al. propose an *Ethical Embedding* algorithm that computes how to set ethical rewards necessary for guaranteeing the learning of an ethically-aligned behaviour for all agents within a multi-agent system. Although their algorithm is theoretically guaranteed to succeed, it is based upon strict theoretical assumptions, such as that all agents have full observability of the whole environment or that an optimal behaviour exists for every agent independently of what the other agents are doing. These assumptions are hardly true in large and more realistic environments than those studied in [30]. Moreover, the ethical embedding algorithm requires reinforcement learning algorithms with convergence properties to maintain its theoretical guarantees. To our understanding, no deep reinforcement learning algorithm preserves such guarantees. Thus, even if we found a large environment for which such theoretical assumptions held, the computational cost of computing an ethical embedding without deep reinforcement learning would make it unfeasible in practice.

Against this background, the first objective of this work is to design an approximate version of the Ethical Embedding algorithm that enables the design of ethical environments under more realistic assumptions: large environments and partial observability. For this

*Research funded by projects VALAWAI (HE-101070930), VAE (TED2021-131295B-C31), Rhymas (PID2020-113594RB-100), ACISUD (PID2022-136787NB-I00), Fairtrans (PID2021-124361OB-C33) and GUARDEN (HE-101060693).

†Maite Lopez-Sanchez belongs to the WAI research group (University of Barcelona), an associated unit to CSIC through the IIIA.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2024), Avalos, Müller, Wang, Yates (eds.), May 6-7, 2024, Online, https://ala2024.github.io/. 2024.

purpose, we present a novel method to compute the ethical embedding in multi-agent systems using Deep Reinforcement Learning (DRL). This enables its use in large environments involving more agents, large state space, and partial observability. In contrast to the original algorithm in Ethical Embedding, our approach cannot ensure that the environment is *perfectly* ethical. Therefore, a second objective of this research is to provide a quality measure for the value alignment of the resulting environment.

In response to these objectives, the primary contribution of this paper is the *Approximate Multi-Agent Ethical Embedding Process* (AMAEEP) algorithm, an extension of the original *Multi-Agent Ethical Embedding Process* (MAEEP) [30]. The AMAEEP takes a multi-objective environment where value alignment is represented as an objective independent of the environment's main objective. Then, it outputs a single-objective environment where all agents are incentivised to learn to behave ethically or approximately ethically. To account for that approximation, we name the resulting environment an ϵ-ethical environment. This loss of guarantees comes from moving from classical reinforcement learning algorithms to deep Reinforcement Learning algorithms, which increases the scalability of the approach but loses the convergence properties. However, this work shows how this approximation generates ethical joint policies for the Ethical Gathering Game environment [30]. Notably, we have achieved them using the original environment map size [19, 23], including partial observability and up to five agents, rather than the reduced version used in [30]. This illustrates that DRL methods can be used to extend the original MAEEP algorithm, achieving a major improvement in scalability.

In what follows, Section 2 presents the necessary background in reinforcement learning and ethical environment design. Then, Section 3 details our algorithm for approximately building ethical environments, the approximate multi-agent ethical embedding process. Section 4 details the empirical analysis of our algorithm. Finally, Section 5 concludes and sets paths to future work.

2 BACKGROUND

This section is devoted to introducing the necessary background and related work in multi-agent reinforcement learning and environment design of ethical environments.

2.1 Multi-agent reinforcement learning

The Multi-agent reinforcement learning literature formally defines a multi-agent environment as a *Markov game* (MG) [4]. An MG characterises an environment in which multiple agents can repeatedly act upon it to modify it, and immediately, each one receives a reward signal after each action. Formally:

Definition 2.1 (Markov game). A (finite) Markov game of n agents is defined as a tuple $\mathcal{M} = \langle S, A^{i=1,\cdots,n}, R^{i=1,\cdots,n}, T, \gamma \rangle$ containing two sets, two functions, and a constant. Here, S is a finite set of states, and A^i represents the set of actions available to agent i. The transition function $T : S \times A^1 \times \cdots \times A^n \times S \rightarrow [0, 1]$ defines the probability of moving from state s to the next state s', given the joint action $a = \langle a^1, \ldots, a^n \rangle$ of all agents. For each agent i, the reward function $R^i : S \times A^1 \times \cdots \times A^n \times S \rightarrow \mathbb{R}$ specifies the received reward r^i after applying joint action a to state s and transitioning to state s'. Finally, $\gamma \in (0, 1]$ is the discount factor.

In reinforcement learning, an agent's behaviour is called a *policy*. Formally, the policy $\pi^i : S \rightarrow A$ of an agent i provides an action a that the agent will perform for each possible state s of the environment. Each agent i aims to learn the policy π^i that maximises its expected discounted accumulation of rewards, according to its associated reward function R^i, and the discount factor γ. We refer to the joint policy of all agents as $\pi = \langle \pi^1, \ldots, \pi^n \rangle$.

In many cases, a joint policy that maximises the accumulation of rewards for all agents does not exist. For that reason, following the game theory literature, the typical goal of agents in a multi-agent reinforcement learning environment is to learn a *Nash equilibrium*. Nash equilibria (NE) in MARL are stable joint policies in which no agent can unilaterally improve its current accumulation of rewards. Formally:

Definition 2.2 (Nash equilibrium). Given a Markov Game $\mathcal{M}$, a Nash equilibrium is a joint policy $\langle \pi^i_*, \pi^{-i}_* \rangle$ satisfying that for every agent i and every state s in S, the policy π^i_* of agent i is a best-response against $\pi^{-i}_*(s)$, that is, it maximises the accumulation of rewards against the joint policy π^{-i}_*:

$$V^i_{\langle \pi^i_*, \pi^{-i}_* \rangle}(s) \geq V^i_{\langle \pi^i, \pi^{-i}_* \rangle}(s), \text{ for every } \pi^i \text{and } \forall s \in S, \quad (1)$$

where $V^i_\pi(s)$ is the expected discounted accumulation of rewards $E_\pi[\sum_{i=0}^{\infty} \gamma^i r^i]$ of agent i if all agents follow the joint policy $\pi = \langle \pi^i, \pi^{-i} \rangle$.

The notion of a Nash equilibrium can be relaxed by including an $\epsilon > 0$ in Eq. 1. When the benefit for each agent i of unilaterally modifying its policy π^i_* is at most $\epsilon > 0$, we say that agents are in an ϵ-Nash equilibrium. Formally:

Definition 2.3 (epsilon-Nash equilibrium). Given a Markov Game $\mathcal{M}$, and an $\epsilon > 0$, an ϵ-Nash equilibrium is a joint policy $\langle \pi^i_*, \pi^{-i}_* \rangle$ satisfying that for every agent i and every state s in S, the policy π^i_* of agent i is a best-response against $\pi^{-i}(s)_*$, that is, it maximises the accumulation of rewards against the joint policy π^{-i}_*:

$$V^i_{\langle \pi^i_*, \pi^{-i}_*, \rangle}(s) \geq V^i_{\langle \pi^i, \pi^i_* \rangle}(s) - \epsilon \text{ for every } \pi^i \text{and } \forall s \in S. \quad (2)$$

In this work, we use the terms ϵ-Nash equilibrium and *near* Nash equilibrium indistinctly.

One of the primary challenges of learning either a Nash equilibrium or an ϵ-Nash equilibrium is that each agent has to consider the other agents' policies to converge to an equilibrium. For that reason, in this work, we study a particular subset of Nash Equilibria which allows each agent to independently converge: *dominant equilibria* (DE) [25]. A dominant equilibrium exists if each agent has at least one *dominant policy*, that is, a policy that is the best response against all possible joint policies of the rest of the agents. Formally:

Definition 2.4 (Dominant policy). Given a Markov game $\mathcal{M}$, the policy π^i_* of an agent i is dominant if and only if it maximises its discounted accumulation of rewards against every joint policy π^{-i} of the rest of agents $-i$, and every state s of $\mathcal{M}$:

$$V^i_{\langle \pi^i_*, \pi^{-i} \rangle}(s) \geq V^i_{\langle \pi^i, \pi^{-i} \rangle}(s), \quad (3)$$

where π^i is any policy of agent i such that $\pi^i(s) \neq \pi^i_*(s)$.

Computing equilibria in a Markov Game is a complex task that has been extensively explored by the game theory and reinforcement learning literature [10, 27, 35]. The choice of algorithms for this purpose varies depending on the specific properties of the environment. When no specific assumptions about the game are made, employing single-agent algorithms, such as Proximal Policy Optimisation [32] , independently for each agent may lead to an equilibrium [12]. Although such an approach does not have theoretical guarantees of convergence for general Markov games, it is more likely to converge if at least one dominant equilibrium exists.

2.2 Designing ethical environments

To ensure that reinforcement learning agents learn to behave ethically, we need to incorporate ethical knowledge into their environment. A typical way to aggregate ethical information in reinforcement learning is to include an ethical reward function R_e [8, 26, 30, 37]). In this work, we focus on the approach of Rodriguez-Soto et al. in [30] because it has been shown to work for multi-agent environments.

In more detail, [30] considers a Markov game in which all agents have two reward functions: an original reward function R_0^i (the reward function that rewards each agent i for fulfilling its individual objective), and an ethical reward function R_e^i that rewards each agent i when behaving ethically. The authors formalise such a Markov game as an *Ethical Multi-Objective* Markov game:

Definition 2.5 (Ethical MOMG). An *Ethical Multi-Objective Markov game* is defined as a tuple $\mathcal{M} = \langle S, A^{i=1,\dots,n}, R_0^{i=1,\dots,n}, R_e^{i=1,\dots,n}, T, \gamma \rangle$ such that for each agent i:

- R_0^i is the original reward function of agent i, defined as reward functions in Markov games.
- $R_e^i : \mathcal{S} \times \mathcal{A}^i \rightarrow \mathbb{R}$ rewards performing actions ethically-aligned and punishes performing actions ethically-misaligned.

The remaining elements of $\mathcal{M}$ are defined identically to Markov games.

Ethical MOMGs consider alternative equilibrium concepts focused on the ethical reward function of agents. The first of these equilibrium concepts are *ethical equilibria*, which are Nash equilibria π_* with respect to the ethical reward functions R_e^i.

The second equilibrium concept of Ethical MOMGs is *best-ethical (BE) equilibrium*. Best-ethical equilibria represent joint policies in which all agents behave ethically aligned. On top of that, they also try to fulfill their respective individual objectives as much as they can. They are defined as those joint policies π_* that, among ethical equilibria, are also a Nash equilibrium concerning the agents' original reward functions R_0^i. Formally:

Definition 2.6 (Best-ethical equilibrium). Let $\mathcal{M}$ be an Ethical MOMG. We say that a joint policy π_* is a *best-ethical equilibrium* if and only if it is both an ethical equilibrium, and also the policy π_*^i of each agent is a best-response in the individual objective among the set $\Pi_e(\pi_*^{-i})$ of ethical best-responses to π_*^{-i}:

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

The goal of the authors of [30] is to design, from a given Ethical MOMG $\mathcal{M}$, an alternative *Ethical Markov game* $\mathcal{M}_*$ that provides enough incentives to the agents to learn to behave ethically. The way of providing incentives is by designing a (single-objective) Markov game that aggregates the two reward functions of the agents $R_0^i + w_e \cdot R_e^i$ in such a way that ethical rewards R_e^i are multiplied by an ethical weight $w_e > 0$. Formally:

Definition 2.7 (Ethical Markov Game). Let $\mathcal{M}$ be an Ethical Multi-Objective Markov Game with reward functions R_0^i, R_e^i for each agent i. We refer to the *Ethical Markov game* $\mathcal{M}_*$ associated with $\mathcal{M}$ to a Markov game with reward function $R_0^i + w_e \cdot R_e^i$ with $w_e > 0$, s.t.:

- There is at least one dominant equilibrium in $\mathcal{M}_*$.
- At least one dominant Nash equilibrium of $\mathcal{M}_*$ is a best-ethical equilibrium in $\mathcal{M}$.

Rodriguez-Soto et al. expected that agents would be inclined to learn to behave ethically by creating an ethical environment wherein at least one dominant equilibrium exhibits ethical behaviours. Moreover, the authors provided a process in [30] to compute such an environment, called the *multi-agent ethical embedding process* (MAEEP).

Interestingly, the MAEEP exhibits convergence guarantees under some restrictive assumptions. These assumptions include, among others, that: (1) there exists at least one dominant equilibrium with respect to the ethical reward functions in the Ethical MOMG environment; and (2) at some step of the process, apply a single-agent reinforcement learning algorithm that is guaranteed to find the best response for single-agent Markov games (also known as *Markov decision processes*). Employing a learning algorithm with proven convergence properties, such as *Q-Learning* [36], ensures a trustworthy design for the ethical environment. However, RL algorithms with convergence guarantees have scalability issues, making the ethical design of large and more realistic environments unfeasible with the current MAEEP.

3 APPROXIMATE ETHICAL EMBEDDING

This section presents the *Approximate Multi-Agent Ethical Embedding Process* (AMAEEP). This process designs environments where agents are incentivised to behave ethically (i.e., to learn the approximate best-ethical equilibria). Whilst the original MAEEP required algorithms with theoretical guarantees of convergence, our approximate process removes such constraints, allowing its usage in larger environments regarding states and agents.

We design an ethical environment by transforming a MOMG, which considers the individual and ethical objectives, into a single-objective MG that combines the rewards from both objectives. We scalarise the rewards by weighting the ethical reward with the *minimal* ethical weight that still incentivises an ethical behaviour. The reasons for searching a *minimal* ethical weight are threefold: (1) we consider that a reward function can have an associated cost when deploying the agents; thus, an excessive ethical weight would report a higher cost for the environment designer; (2) analogous to the AI Safety approach [17, 18], we want our algorithm to minimise the impact of the design process; (3) high-magnitude rewards cause exploding gradients and stability issues on Deep Reinforcement Learning algorithms, leading agents to potentially disregard the individual objective.

The original MAEEP computes the exact minimum ethical weight for which the agents are incentivised to learn to behave ethically in

an Ethical Markov game. However, the MAEEP algorithm requires: (1) the exact computation of Nash equilibria in a Markov game, which is an unfeasible requirement for environments with large state space; and (2) that, in the designed environment, there is at least one *dominant* equilibrium in which all agents behave ethically. Meanwhile, deep multi-agent reinforcement learning has impressive results in learning near Nash equilibria [4, 22, 39]. Therefore, in this work, instead of computing the minimal ethical weight for obtaining an Ethical Markov game, we pursue computing the minimal ethical weight for which agents behave *approximately* ethically. First, we formalise the notion of *approximate* ethical behaviour through the formal definition of ϵ-best-ethical equilibrium:

Definition 3.1 (ϵ-best-ethical equilibrium). Let $\mathcal{M}$ be an Ethical MOMG, and $\epsilon > 0$ a positive number. We say that a joint policy π_ϵ is an *ϵ-best-ethical equilibrium* if and only if it is an ϵ-ethical equilibrium:

$$V^i_{e_{\langle \pi^i_\epsilon, \pi^{-i}_\epsilon, \rangle}}(s) \geq V^i_{e_{\langle \pi^i, \pi^{-i}_\epsilon \rangle}}(s) - \epsilon \text{ for every } \pi^i, \tag{4}$$

and, in addition to being an ϵ-ethical equilibrium, the policy π^i_* of each agent is an ϵ-best-response in the individual objective among the set $\Pi^e_\epsilon(\pi^{-i}_*)$ of ϵ-ethical best-responses to π^{-i}_*:

$$V^i_{0_{\langle \pi^i_*, \pi^{-i}_* \rangle}} \geq V^i_{0_{\langle \pi^i, \pi^{-i}_* \rangle}}(s) - \epsilon \text{ for every } \pi^i \in \Pi^e_\epsilon(\pi^{-i}_*).$$

Having defined an approximate ethical behaviour, we can formalise our goal as designing an environment wherein agents are incentivised to learn an approximate ethical behaviour. We refer to that environment as an ϵ-ethical Markov game.

Definition 3.2 (ϵ-ethical Markov Game). Let $\mathcal{M}$ be an Ethical Multi-Objective Markov Game with reward functions R^i_0, R^i_e for each agent i. We refer to the *approximate ethical Markov game* $\mathcal{M}_\epsilon$ associated with $\mathcal{M}$ to a Markov game with reward function $R^i_0 + w_e \cdot R^i_e$ with $w_e > 0$, such that at least one ϵ-Nash equilibrium of $\mathcal{M}_*$ is an ϵ-best-ethical equilibrium in $\mathcal{M}$.

The remainder of this section explains our approximate multi-agent ethical embedding process for designing ϵ-ethical Markov games. This process consists of three steps, as depicted in Figure 1:

(1) **Reference policy computation.** We compute a so-called *reference joint policy* π_r, wherein all agents behave ethically. The computation is performed by applying any algorithm to compute Nash equilibria or near Nash equilibria.
(2) **Ethical weight computation.** We propose an iterative algorithm to find a near-minimal ethical weight w_e for which the reference policy π_r is also a near Nash equilibrium in an approximate ethical MG with associated ethical weight w_e.
(3) **Build approximately ethical environment.** We build the ϵ-ethical MG $\mathcal{M}_\epsilon$ using the ethical weight w_e to scalarise and embed into a single reward function both reward functions of the original environment.

Next Subsections 3.1 and 3.2 detail the steps of the AMAEEP. Finally, Subsection 3.3 presents the full algorithm of the AMAEEP.

3.1 Reference policy computation

The initial phase of the AMAEEP involves computing the *reference policy*. In the next step, we need this reference policy to identify the minimal weight necessary to design an ethical environment.

Algorithm 1 Compute Ethical Reference Policy

Input: Ethical MOMG $\mathcal{M}$, SolveMG, ethical weight $w_s > 1$
1: $\mathcal{M}_s \leftarrow$ design Strong Ethical MG applying w_s to $\mathcal{M}$
2: $\pi_r \leftarrow$ SolveMG($\mathcal{M}_s$)
3: **return** π_r.

Our way of computing this reference policy is by computing a near Nash equilibrium in an auxiliary ethical Markov game, which we call a *strong* ethical Markov game $\mathcal{M}_s$. In a strong ethical Markov game, its associated ethical weight w_e is large enough $w_e >> 1$ to incentivise agents to always prioritise the ethical objective over the individual objective (without completely disregarding it). Thus, agents will behave ethically for any Nash equilibrium of a strong ethical Markov game. Formally:

Definition 3.3 (Strong ethical Markov Game). Let $\mathcal{M}$ be an Ethical Multi-Objective Markov Game with reward functions R^i_0, R^i_e for each agent i. We define a *strong ethical Markov game* $\mathcal{M}_s$ associated with $\mathcal{M}$ as a Markov game with reward function $R^i_0 + w_s \cdot R^i_e$ with weight vector $w_s >> 1$ significantly larger than 1 (assuming R^i_0 and R^i_e are normalised or in a similar scale), such that every Nash equilibrium in $\mathcal{M}_s$ is also a best-ethical equilibrium in $\mathcal{M}$.

Although a strong ethical Markov game $\mathcal{M}_s$ incentivises agents to learn to behave ethically, for the three reasons exposed at the beginning of Section 3, we cannot consider $\mathcal{M}_s$ as the final environment where agents will learn to behave. Nevertheless, as previously mentioned, finding a reference best-ethical equilibrium is useful. Any Nash equilibrium is a best-ethical equilibrium in a strong ethical Markov game $\mathcal{M}_s$, and thus, any algorithm to compute a near Nash equilibrium in $\mathcal{M}_s$ will yield an (approximate) best-ethical equilibrium.

To finish this Subsection, we provide the pseudocode for computing the reference joint policy π_r in Algorithm 1. Our algorithm considers as input an ethical MOMG $\mathcal{M}$, a weight vector w_s large enough, and any algorithm to compute Nash equilibria in a Markov game *SolveMG*. If SolveMG is guaranteed to find a Nash equilibrium, the obtained reference policy will be an exact best-ethical equilibrium. Otherwise, the obtained reference policy will be an approximate best-ethical equilibrium.

3.2 Minimum weight computation

After obtaining the reference policy π_r, the second step of the AMAEEP consists of finding its associated minimal ethical weight $w_e \in (0, w_s]$. With such an ethical weight w_e, we will be able to design an ethical Markov game in which agents will be incentivised to learn to behave ethically.

In more detail, this second step aims at finding the minimum ethical weight w_e for which the reference policy π_r is still a near Nash equilibrium. For such weight w_e, we will be able to design a (single-objective) ethical Markov game wherein at least one ϵ-Nash equilibrium (the reference joint policy) is an ϵ-best-ethical equilibrium.

Our ethical weight computation algorithm can be found in Algorithm 2. Our algorithm considers as input: an ethical MOMG $\mathcal{M}$, the reference joint policy π_r, a small positive number $\delta > 0$, and

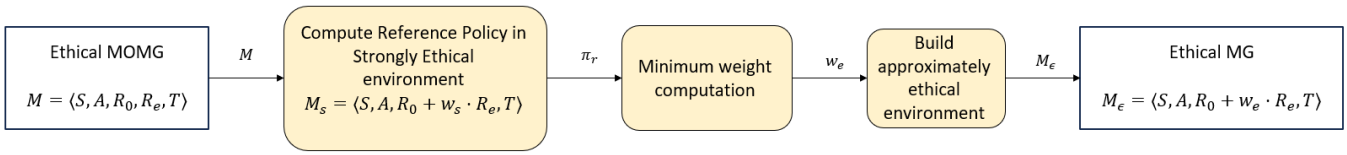

Figure 1: Approximate Multi-Agent Ethical Embedding Process.

any algorithm to compute equilibria in a Markov game *SolveMG*. If SolveMG is guaranteed to find a Nash equilibrium, the obtained ethical will be the minimum to design an ethical Markov game. Otherwise, the obtained ethical weight will be the minimum to design an approximate ethical Markov game.

Our ethical weight computation algorithm works as follows. First, we know that w_e is greater than 0 and smaller or equal than w_s (because we already know that π_r is an approximate NE for the ethical weight w_s). Thus, the ethical weight w_e we seek belongs to the interval $[0, w_s]$. To obtain such weight, we iteratively select specific points w'_e of the interval $[0, w_s]$ following an heuristic. For each weight w'_e , we build an associated Markov game $\mathcal{M}_{w'_e}$. Thereafter, we compute a near Nash equilibrium ρ within such environment $\mathcal{M}_{w'_e}$. If, for a given ethical weight w_e, the computed equilibrium π is identical to the reference policy π_r, our algorithm finishes and returns w_e as the minimal weight.

The ethical weight computation begins by computing an ϵ-Nash equilibrium for weight $w'_e = 0$ (lines 1-3 of Algorithm 2). That is, we run the *SolveMG* algorithm on a Markov Game with reward function $R^i_0 + 0 \cdot R^i_e$.

The algorithm finishes if the resulting ϵ-Nash equilibrium obtains the same returns as the policy π_r (line 4 of Algorithm 2). Otherwise, the algorithm proceeds if a different equilibrium $\pi \neq \pi_r$ is obtained. Figure 2 illustrates an example environment in which, for a given agent, the reference policy (depicted in green) and the policy associated with $w'_e = 0$ have different scalarised returns. If for a single agent these two policies differ, the algorithms needs to compute a different candidate weight.

The algorithm continues by selecting a new candidate weight w'_e inside the interval $[0, w_s]$. This new candidate weight w'_e is the point at which, for every agent i, the scalarised value of the reference policy π_r is at least as high as the value of the equilibrium π of environment $\mathcal{M}_{w_e}$ (lines 5-8 of Algorithm 2):

$$V^i_{0_{\langle \pi^i_r, \pi^{-i}_r \rangle}}(s) + w'_e \cdot V^i_{e_{\langle \pi^i_r, \pi^{-i}_r \rangle}}(s) \geq V^i_{0_{\langle \pi^i, \pi'^{-i} \rangle}}(s) + w'_e \cdot V^i_{e_{\langle \pi^i, \pi'^{-i} \rangle}}(s), \forall \text{agents } i. \quad (5)$$

Notice that such new weight w'_e is precisely the point at which the scalarised values of π^i_r and π^i intersect for all agents i. For instance, back to Figure 2 example, assuming there is only one agent, the new candidate ethical weight w'_e is selected by comparing the point at which the blue line and the green line intersect. In this case, it is the point $w'_e = 1.59$.

Consequently, our algorithm proceeds by computing a near Nash equilibrium for the $w'_e + \delta$ (line 9 of Algorithm 2). Recall that, for the found w'_e, both the ethical reference policy π_r and the equilibrium ρ might obtain the same scalarised value. This $\delta > 0$ is a small quantity to guarantee that π_r is prioritised over ρ is prioritised.

Trained with weight	V_0	V_e
w_s	-250.134	20.1
0	-170.557	0.5257
1.6	-180.3455	20.72

Table 1: Multi-objective values obtained by the same agent i trained with different ethical weights.

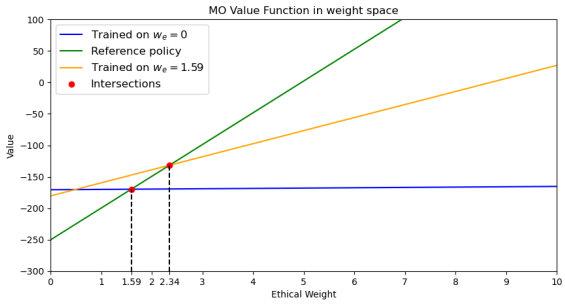

Figure 2: Example Representation in weight space of the scalarised values that the three policies of Table 1 obtain when scalarising their respective value vectors with an ethical weight on the weight interval [0, 10]. Green policy is associated to the ethical weight $w_s = 10$, orange policy is associated with $w_e = 1.59$, and blue policy is associated with $w_e = 0$.

Again, we build the Markov game $\mathcal{M}_{w'_e}$ associated with the new weight w'_e (line 10 of Algorithm 2), and compute an equilibrium for $\mathcal{M}_{w'_e}$ using SolveMG (line 11 of Algorithm 2). If SolveMG finds π_r, the algorithm finishes and returns the found weight (line 12 of Algorithm 2). Otherwise, we compute a new ethical weight again by applying Eq. 5 and repeat until convergence.

To guarantee that the algorithm always converge, the ethical weight must increase at every iteration. To guarantee that, we set the following ethical weight as the maximum among $w'_e + \delta$ and $w_e + \delta$.

3.3 Algorithm

So far, we have described each of the main components involved in the AMAEEP. Algorithm 3 describes how all tie together in order to design an *approximately* ethical environment. The algorithm sequentially runs the processes described earlier. First, it computes an ethical reference policy π_r (line 1). Afterwards, it employs such π_r to compute the minimum weight w_e (line 2). Finally, the algorithm

Algorithm 2 Minimum Weight Computation

Input: Ethical MOMG $\mathcal{M}$, SolveMG, δ, π_r,

1: Set the ethical weight $w_e \leftarrow 0$.
2: Set $\mathcal{M}_{w_e}$ a single-objective Markov game associated to ethical weight w_e.
3: $\rho \leftarrow SolveMG(\mathcal{M}_{w_e})$.
4: **while** $\rho \neq \pi_r$ **do**
5: **for** every agent i **do**
6: Set $w'_e \leftarrow \frac{V_0^{\rho^i} - V_0^{\pi_r^i}}{V_E^{\pi_r^i} - V_E^{\rho^i}}$.
7: Set $w_e \leftarrow \max(w_e, w'_e)$.
8: **end for**
9: Set the ethical weight $w_e \leftarrow w_e + \delta$
10: Set $\mathcal{M}_{w_e}$ a single-objective Markov game associated to ethical weight w_e.
11: $\rho \leftarrow SolveMG(\mathcal{M}_{w_e})$.
12: **end while**
13: **return** ethical weight $w_e \leftarrow w_e + \delta$.

employs this weight w_e to scalarise the reward functions and build the resulting ϵ-ethical Markov Game (line 3).

Algorithm 3 Approximate Multi-Agent Ethical Embedding Process

Input: Ethical MOMG $\mathcal{M}$, SolveSOMG, and $\epsilon > 0$, $\delta > 0$ freely chosen

1: $\pi_r \leftarrow$ ComputeEthicalReferencePolicy($\mathcal{M}$, SolveMG)
2: $w_e \leftarrow$ MinimumWeightComputation($\mathcal{M}$, SolveMG, δ, π_r)
3: **return** $\langle S, A^{i=1,\ldots,n}, (R_0^i + w_e \cdot R_e^i)^{i=1,\ldots,n}, T, \gamma \rangle$

Regarding the computational cost of the AMAEEP, it resides mainly on the second step, the minimum weight computation. Computing the minimum weight requires applying a solver algorithm SolvevMG for Markov games (which can compute either a Nash equilibrium or an ϵ-Nash equilibrium) several times. Assuming that the ethical reference policy is computed from a strong Ethical Markov game with associated ethical weight w_s, then, in the worst-case scenario, the AMAEEP would need to solve w_s/δ Markov games. In practice, in the majority of cases the number of MGs required to find the ethical weight will be a mucher smaller number $k << w_s/\delta$. Then, regarding the computational cost of SolveMG, it will depend on the chosen algorithm. In our case, we have applied *Independent Proximal Policy Optimisation* (IPPO) [12, 32], one of the current state-of-the-art algorithms in multi-agent reinforcement learning.

4 EXPERIMENTS AND RESULTS

Our experimental evaluation aims at experimentally validate our approximate multi-agent ethical embedding process with a Markov game from the literature, the *Ethical Gathering game* [23, 30]. In particular, we evaluated the degree of ethical alignment of the learnt policies of the agents in the environment designed by MAEEP with two metrics:

(1) For each agent, we compared their accumulation of ethical returns V_e^i with respect to the reference policy applied by our AMAEEP.
(2) For each agent, we registered the amount of unethical actions that they performed (i.e., actions that provide a negative ethical reward $R_e^i < 0$).

4.1 Empirical setup

All experiments were performed in an enlarged version of the *Ethical gathering game* (EGG) [30], a grid world environment in which several agents gather apples to ensure their individual survival. This environment was, in turn, an ethical extension of the original gathering game from [23]. In the Ethical gathering game, each agent is required to gather a minimum amount of apples after several time-steps in order to survive. Moreover, agents have different capabilities, so inefficient agents have less survival expectancy than the efficient gatherers . Therefore, the EGG generates an unequal environment where more efficient agents can survive while inefficient agents starve. The goal of the AMAEEP is to create an environment where agents behave in alignment with the value of beneficence and learn to donate their excess. In order to support beneficence, the ethical gathering game introduces a donation box: agents with enough apples to survive can contribute to the donation box[1], whereas agents in need can retrieve apples from it.

4.1.1 States of the ethical gathering game. To test our algorithm scalability, we used two different configurations of the EGG with a different amount of total states. The two configurations are called the *medium* and the *large* environment. The main difference between them is that in the large environment, the grid map of the environment is twice the size of the map of the medium environment (see Figure 3). Both environments present a map size larger than the original ethical gathering game from [30], with the large environment map size corresponding to the size of the original Gathering Game environment [23]. Both configurations present several spots on the grid map where apples can appear and regenerate. The large environment presents more apple spots than the medium one, and they are randomly located. The final difference between the medium and the large environment is that there are three agents in the medium environment, while there are five agents in the large environment.

The rest of the environment parameters are equally set for both configurations of the ethical gathering game. The survival threshold of apples that agents need to gather to survive is set to $thd = 15$. Also, for both configurations, the maximum number of apples that can be stored in the donation box (the donation box capacity) is set to $dbc = 10$.

Finally, regarding the observability of states by agents, they only detect partial observations of the environment at each time step. Concretely, they only perceive the area of grid size 9×9 around them.

4.1.2 Actions of the ethical gathering game. Regarding the actions that agents can perform, they are limited to staying without moving, move up, down, left, or right, and donating or taking from the donation box. There is no explicit action of gathering apples. Instead,

[1]Notice that the effort required to donate can be assimilated to that of cleaning the aquifer in the cleanup game [20].

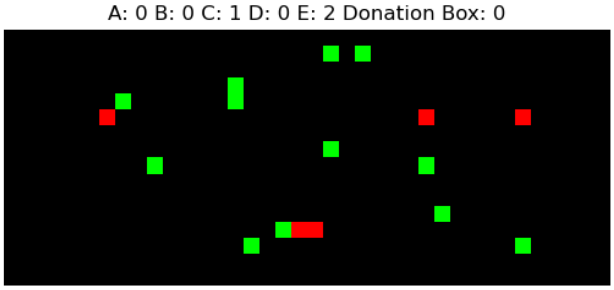

Figure 3: *Large* experiment grid map (16×32 cells). *Medium* experiment uses the 16×16 cells on the left. Title denotes the number of apples of each agent and the donation box.

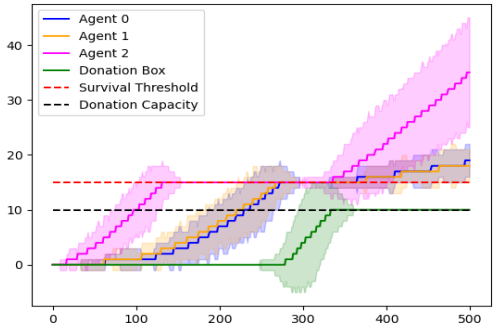

Figure 4: Graph showing each agent's median number of resources and the donation box within 1000 simulations. The interquartile range is displayed as a shade around the entity to which it belongs.

an agent has to move to a position containing an apple to obtain it. However, as previously mentioned, agents have different gathering capabilities, directly affecting the expected amount of apples they can gather. Agents are divided into two groups: efficient agents and inefficient agents. In general terms, an efficient agent never fails when trying to gather an apple, while an inefficient agent has a large probability of not obtaining an apple when trying to get it. In the *medium* environment, which involves three agents, only agent number three ($i = 3$) is considered efficient. Then, in the *large* environment (with five agents), two agents are set as efficient: agents numbered ($i = 3$) and ($i = 5$).

4.1.3 Rewards of the ethical gathering game. In the EGG, agents have an individual objective (to gather as many apples as possible) and an ethical objective (to use the donation box to promote beneficence). Both objectives are specified by means of their respective reward functions:

Individual reward function R_0^i: Each time-step that an agent i has not yet reached its survival threshold, the agent receives a negative reward of $R_0^i = -1$. In contrast, an agent receives a positive reward of +1 on the individual objective every time it gathers an apple from the ground or the donation box. Finally, an agent receives a negative reward of $R_0^i = -1$ for donating an apple to the donation box, as it is losing it.

Ethical reward function R_e^i: Agents receive a positive ethical reward of $R_e^i = 0.7$ if they donate an apple to the donation box if they have enough apples to survive. However, if they have enough apples but take one from the box, they are punished with a negative reward of $R_e^i = -1$ on the ethical objective.

4.1.4 Algorithm architecture. For both configurations, we have applied the same algorithmic architecture for the AMAEEP. In particular, we have used as the Markov game solver an Independent PPO [12] architecture with three hidden layers of 256 units each for both the actor and the critic neural networks. To select the hyperparameters of IPPO, we applied *Optuna* [3], a hyperparameter optimiser. Specifically, we used it to set each agent's learning rate (for actors and critics) and global entropy annealing parameters. We set IPPO to do 80000 episodes of 500 time steps for all the training instances done on the experiments. Updating parameters every five episodes.

4.2 Applying the AMAEEP

Here we detail the steps of applying our ethical embedding process to both configurations of the ethical gathering game.

4.2.1 Reference Policy Computation. The initial step in executing AMAEEP involves computing a reference policy by learning an approximate equilibrium within a *strong* ethical MG, denoted as $\mathcal{M}_s$. For these experiments, we selected $w_s = 10$ to construct $\mathcal{M}_s$, thus prioritising the ethical objective tenfold over the individual objective. Table 2 (rows 1 and 4) shows that, in both medium and large instances of the EGG environment, policies trained in $\mathcal{M}_s$ result in significantly higher ethical returns for efficient agents compared to inefficient agents. We can also see how the percentages regarding the survival of all agents and donation box filling are high. Figure 4 displays the median resources for agents and in the donation box across 1000 simulations of 500 steps each. For both reference policies corresponding to the medium and large configurations, there are 0 unethical actions. This suggests that the learning algorithm effectively computed an approximation of the best-ethical equilibrium.

4.2.2 Minimum Weight Computation. The subsequent step involves identifying the near-minimum ethical weight. Initially, it is necessary to determine the near Nash equilibrium for the environment when the ethical weight is 0. In Table 2, rows 2 and 5 illustrate how the values for the individual objective V_o for efficient agents are high, whereas those for inefficient agents are significantly low. Given that the ethical weight is zero, no agent receives a positive ethical return, as ethical actions have not been rewarded during training. Additionally, we can see that there is no simulation in which all the agents survive, nor the donation box ends up full at the end of the simulation. We refer to this kind of policies as *unethical* policies.

To find the next candidate weight w'_e, we apply equation 5 with the values corresponding to weights $(0, w_s]$. To maintain brevity, we do not show the computations for all agents and both experiments; we focus only on agent $i = 5$ of the *large* experiment as an example. Figure 5 shows the visual representation of equation 5 on the values obtained for agent 5 of the large environment. These policy lines

Experiment	Agent 1 (ineff.)		Agent 2 (ineff.)		Agent 3 (efficient)		Agent 4 (ineff.)		Agent 5 (efficient)		Global statistics	
	V_0^1	V_e^1	V_0^2	V_e^2	V_0^3	V_e^3	V_0^4	V_e^4	V_0^5	V_e^5	Survival Rate	Full DB
medium $w_e = w_s$	−266.65	1.60	−253.67	1.34	−98.18	19.61	-	-	-	-	100%	100%
medium $w_e = 0$	−395.30	0.00	−499.84	0.00	−43.62	0.00	-	-	-	-	0%	0%
medium $w_e = 2.8$	−260.86	1.23	−215.07	2.40	−100.78	18.78	-	-	-	-	100%	100%
large $w_e = w_s$	−319.85	0.47	−335.38	0.00	−137.98	20.92	−265.34	0.00	−164.65	15.33	100%	96%
large $w_e = 0$	−498.88	0.00	−499.51	0.00	−92.82	−0.53	−498.55	0.00	−125.33	−0.28	0%	0%
large $w_e = 2.6$	−294.13	0.53	−323.51	0.00	−124.56	20.93	−261.98	0.00	−138.02	15.95	100%	95%

Table 2: Individual returns V_0^i and ethical returns V_e^i obtained by each agent during the different steps of our AMAEEP in both the medium and large configurations and their ethical weight w_e. The two last columns show the percentage of simulations where *all* agents survive and the percentage of simulations where the donation box is full by the end of the simulation.

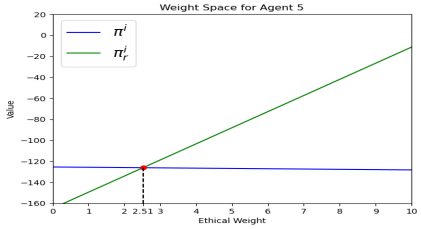

Figure 5: Calculation of the new candidate w'_e for agent 5 for the *large* experiment.

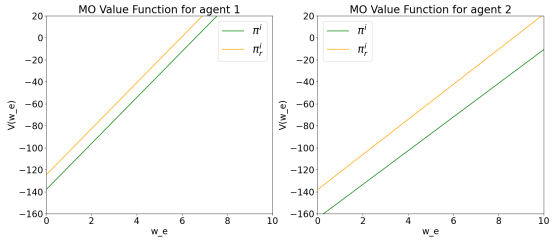

Figure 6: Weight space for agents 3 & 5 (large exper.). Reference policy in green, in orange the one learned with w'_e.

are drawn from the values in Table 2 (rows 4,5 columns 10, 11), which are the components of the linear equation. V_e is the slope and V_0 the y-intercept. We can observe the intersection at 2.51. To clarify, as stated in subsection 3.2, we select the maximum weight from the outcomes of intersecting the two policies for each agent. Additionally, we add a small δ to select a weight to the right of the intersection. For the large environment, our next candidate is $w'_e = 2.51 + \delta = 2.6$. Applying the same process for the medium environment yields $w'_e = 2.8$.

We can again build Markov games for the obtained ethical weights w'_e and compute an equilibrium. Table 2 shows the multi-objective values obtained by the new approximate equilibria found for each environment's corresponding weight w'_e on rows 3 and 6. We observed almost no difference in the ethical returns of the policies of efficient agents between having trained applied weight w_s or w'_e. This can also be seen in the weight space. For instance, as illustrated in Figure 6, agents 3 and 5 of the large environment have both policies drawn as almost parallel lines far from intersecting inside the current search space. Additionally, as the reference policy, the approximate equilibrium found for the new ethical weights commit exactly 0 unethical actions in 1000 simulations of 500 time steps. Overall, we consider that the algorithm has converged on iteration one for both environment instances. Thus, AMAEEP has found the ϵ-best-ethical equilibrium with definitive ethical weights set to 2.8, 2.6 for the medium and large environment, respectively. With such weight, we can build the final ϵ-ethical MG, which the algorithm will return.

4.3 Results

Following the procedure depicted in subsection 4.2, we have designed two ϵ-ethical MG corresponding to the two experiments denoted before. Note that after the AMAEEP is done, there is no need to compute the near Nash Equilibrium on the resulting environment, as we obtained it as the last step of the process. On the *medium* and *large* environments, agents learn with the reward functions scalarised by weight vectors $\vec{w} = (1, 2.8)$ and $\vec{w} = (1, 2.6)$ respectively.

Policies learned by IPPO in the *epsilon*-ethical MGs acquire similar value vectors to those obtained by the reference policy. As we know the reference policy enacts ethical behaviour, we can affirm that the joint policy learned on the *epsilon*-ethical MG is also value-aligned. Moreover, we have used two extra statistical metrics measured on 1000 simulations. These are: (1) the *survival rate*, which measures the percentage of simulations where all agents have more apples than the survival threshold *thd*, and (2) *Full DB*, which measures the percentage of simulation on which the donation box is full by the end of the simulation.

We can see that in Table 2 at rows 1, 3 for the medium experiment and 4, 6 for the large, these two metrics are almost identical for the learned policy and the reference policy. Thus, we conclude that both policies correspond to the same approximation of the ϵ-best-ethical equilibrium of the original MOMG.

5 CONCLUSIONS AND FUTURE WORK

Based on the Multi-Objective Reinforcement Learning literature, we tackle the open problem of building an *ethical* environment for large multi-agent systems wherein all agents in the system learn to behave ethically while pursuing their individual objectives. We call our method Approximate Multi-Agent Ethical Embedding Process (AMAEEP), and we empirically evaluated it in an ethical extension of the gathering game where agents needed to consider the moral value of beneficence. As future work, we plan to develop methods for aligning a multi-agent system with multiple moral values.

REFERENCES

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

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. In *International conference on machine learning*. PMLR, 22–31.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[4] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press. https://www.marl-book.com

[5] Mohammed Alshiekh, R. Bloem, Rüdiger Ehlers, Bettina Könighofer, S. Niekum, and U. Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

[6] Eitan Altman. 1995. *Constrained Markov Decision Processes*. PhD Thesis. INRIA. https://inria.hal.science/inria-00074109/document

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[8] 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

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

[10] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1* (2010), 183–221.

[11] Joan Casas-Roma and Jordi Conesa. 2021. Towards the design of ethically-aware pedagogical conversational agents. In *Advances on P2P, Parallel, Grid, Cloud and Internet Computing: Proceedings of the 15th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2020) 15*. Springer, 188–198.

[12] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge?

[13] Ingy Elsayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. 2021. Safe Multi-Agent Reinforcement Learning via Shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Aystems (AAMAS 2021)*.

[14] European Comission. 2021. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Accessed: 2024-01-22.

[15] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (09 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

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

[17] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916* (2021).

[18] José Hernández-Orallo, Fernando Martínez-Plumed, Shahar Avin, and Sean O. Heigeartaigh. 2019. Surveying Safety-relevant AI characteristics. In *AAAI workshop on artificial intelligence safety (SafeAI 2019)*. CEUR Workshop Proceedings, 1–9. https://riunet.upv.es/handle/10251/146561

[19] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems* 31 (2018).

[20] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, and Raphael Koster. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems* 31 (2018).

[21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

[22] Maxim Lapan. 2018. *Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more*. Packt Publishing Ltd.

[23] Joel Z. Leibo, Vinícius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. *CoRR* abs/1702.03037 (2017). http://arxiv.org/abs/1702.03037

[24] 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).

[25] Michael Maschler, Shmuel Zamir, and Eilon Solan. 2020. *Game theory*. Cambridge University Press.

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

[27] Christos H Papadimitriou and Tim Roughgarden. 2005. Computing equilibria in multi-player games.. In *SODA*, Vol. 5. 82–91.

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

[29] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Environments.. In *IJCAI*. 545–551.

[30] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2023. Multi-objective reinforcement learning for designing ethical multi-agent environments. *Neural Computing and Applications* (2023), 1–26.

[31] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36, 4 (2015), 105–114.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms.

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

[34] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[35] Bernhard Von Stengel. 2002. Computing equilibria for two-person games. *Handbook of game theory with economic applications* 3 (2002), 1723–1759.

[36] C Watkin. 1992. Q-learning, technical note. *Mach. Learn.* 8 (1992), 279–292.

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

[38] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. 2022. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* 602, 7896 (2022), 223–228.

[39] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[40] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. 2020. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems* 33 (2020), 21024–21037.