A Structural Solution to Sequential Moral Dilemmas*

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

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

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

ABSTRACT

Social interactions are key in multi-agent systems. Social dilemmas have been widely studied to model specific problems in social interactions. However, state-of-the-art social dilemmas have disregarded specific ethical aspects affecting interactions. Here we propose a novel model for social dilemmas, the so-called *Sequential Moral Dilemmas*, that do capture the notion of moral value. First, we provide a formal definition of sequential moral dilemmas as Markov Games. Thereafter, we formally characterise the necessary and sufficient conditions for agents to learn to behave ethically, so that they are aligned with the moral value. Moreover, we exploit our theoretical characterisation to provide a *structural solution* to a sequential moral dilemma, namely how to configure the Markov game to solve the dilemma. Finally, we illustrate our proposal through the so-called *public civility game*, an example of a sequential moral dilemma considering the *civility* value. We show the social benefits obtained when the agents learn to adhere to the moral value.

CCS CONCEPTS

• **Theory of computation → Multi-agent reinforcement learning**; • **Computing methodologies** → *Cooperation and coordination*;

ACM Reference Format:
Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez-Aguilar. 2020. A Structural Solution to Sequential Moral Dilemmas. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

The increasing presence of intelligent systems in human societies has emphasised the need to consider numerous ethical questions such as how to ensure that artificial intelligences are trustworthy and do not pose any risk to humans [3, 4, 29, 39, 42]. It is of utter importance to develop algorithms so that autonomous agents learn to behave ethically, that is, in alignment with the ethical criteria established in the societies where they are meant to operate. Value alignment is of the utmost importance because Artificial Intelligence (AI) applications in all areas could be seriously discredited if ethical considerations are not taken into consideration. For example, a cleaning robot could do more harm than good if it decided to knock over a vase because it was the fastest way to clean a room [1]. Thus, the question being raised is: how can we instruct an agent to act responsibly so that it can be integrated into our societies? [12]

Social dilemmas, such as *the tragedy of the commons* [14], represent conflicts between individual and collective interests [21]. They present situations where if every agent tries to maximise only its own benefit, the final outcome is worse for everybody. Recently, social dilemmas have been studied in the context of temporally extended scenarios in the so-called *sequential social dilemmas* (SSD) [23, 40]. The *cleanup game* [19] constitutes an example of SSD where agents aim to collect apples from a field while also needing to occasionally clean the aquifer that supplies water to the apples. SSDs are a particular case of Markov games (MG), the formal framework of multi-agent reinforcement learning (MARL) [22, 24].

The formalism of SSDs serves as an effective way of modelling classical social problems where our only goal is to make agents learn to cooperate, that is, to maximise the outcome for every agent [6]. However, actual-world social dilemmas can be much more complex [5, 21]. Hence, here we argue that SSDs lack an ethical dimension:

(1) Actions can be as important as outcomes themselves. Agents' behaviours may be constrained by norms they must obey.
(2) Actual-world agents pursue outcomes aligned with the moral values of the society they live in, even if they are not the best outcomes for them.

Against this background, the purpose of this paper is twofold: (1) to tackle the aforementioned issues via creating a model for social dilemmas that includes a moral perspective; (2) and to develop a solution for such social dilemmas that makes agents act ethically.

Firstly, we introduce the so-called Sequential *Moral* Dilemma (SMD), an extension of Markov games where agents need to choose between behaving ethically or pursuing their individual goals.

Secondly, considering that solutions to social dilemmas can be strategic, motivational, or structural[1] [21], we present a structural solution for SMDs that changes the rules of the agent society. In particular, we assume that agents learn to behave by applying a classical MARL method, and thus, we modify agents' rewards so that they learn to behave ethically. Specifically, we propose an ethical function that rewards alignment with a moral value and that penalises non-compliance with established regulations.

Moreover, we provide theoretical results of the necessary and sufficient conditions for an agent to learn to act ethically. We show how to extend the rewards of an agent so that its behaviour becomes ethically-aligned. With this characterisation we also provide a formal definition of a policy ethically-aligned to a moral value.

*Research supported by projects AI4EU (H2020-825619), LOGISTAR (H2020-769142) and PGC2018-096212-B-C33. Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387).

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

[1] According to [21], motivational solutions assume that agents are not completely egoistic, strategic solutions assume egoistic actors, and structural solutions change the rules of the game.

Finally, we present an example of a sequential moral dilemma – the so-called *public civility game*, which is related to keeping streets clean – that illustrates our structural solution. After applying our structural solution, we empirically show that agents are capable of learning an ethically-aligned equilibrium with a simple Q-learning algorithm. Furthermore, we evaluate the effects of the learnt behaviour with several social behaviour metrics [23] that quantify the benefits of behaving ethically.

The remainder of the article is structured as follows. Section 2 presents some background. Section 3 introduces SMDs and Section 4 describes our structural solution for SMDs. Section 5 presents an example of SMD, the public civility game, which is evaluated in Section 6. Finally, Section 7 draws conclusions and outlines possible lines of future work.

2 BACKGROUND

Definition 1 (Markov game). *A (finite) Markov game (MG) [22, 24, 28] of m agents is the multi-agent extension of Markov decision processes. It is defined as a tuple containing a (finite) set $\mathcal{S}$ of the possible states of the environment, and a (finite) set $\mathcal{A}^i$ of actions for every agent i. Actions upon the environment change the state according to the transition function $T : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^m \times \mathcal{S} \to [0, 1]$. After every transition, each agent i receives a reward based on function $R^i : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^m \times \mathcal{S} \to \mathbb{R}$.*

Each agent i decides which action to perform according to its policy $\pi^i : \mathcal{S} \times \mathcal{A}^i \to [0, 1]$ and we call joint policy $\pi = \prod_{i=1}^{m} \pi^i$ to the union of all agents' policies. The agents learn their respective policies with the goal of maximising their expected sum of rewards

$$V^i_\pi(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R^i_{t+k+1} \mid \pi, S_t = s] \tag{1}$$

for every state s, where $\gamma \in [0, 1]$ is called the discount factor and is problem-dependant. Notice that V^i_π depends on the joint policy.

When an agent i tries to maximise its V^i with respect to all the policies of the other agents (assuming the rest have fixed policies), we refer to such policy as the best-response. When all agents reach a situation such that all have a best-response policy, we say that we have a Nash equilibrium (NE). NEs are stable points where no agent would benefit from deviating from its current policy. Formally:

Definition 2 (Nash equilibrium). *Given a Markov game, we define a Nash equilibrium (NE) [18] as a joint policy π_* such that for every agent, $V^i{}_{\langle \pi^i_*, \pi^{-i}_* \rangle}(s) = \max_{\pi^i} V^i{}_{\langle \pi^i, \pi^{-i}_* \rangle}(s)$ for every state s.*

Here, π^{-i} refers to the joint policy of all the agents except agent i.

3 SEQUENTIAL MORAL DILEMMAS

In this section, we model sequential moral dilemmas (SMD) as a particular kind of Markov games where each agent is intended to learn a policy aligned with a given moral value. We gradually introduce the SMD concept. First, we propose a definition of the so-called moral value signature in subsection 3.1 to build our model on top of it. Then, in subsection 3.2, we show how this concept can be introduced in Markov games. This allows us to formalise, in subsection 3.3, what it means for a policy to be ethically-aligned with respect to a moral value. After introducing all these concepts, we can finally define sequential moral dilemmas in subsection 3.4.

3.1 Considering moral values

When considering a moral value, we propose to take into account two main dimensions: (1) a normative dimension regulating those actions that agents are obliged or forbidden to perform in order to support a given moral value, and (2) an evaluative dimension that considers praiseworthiness (with respect to the same moral value) of actions performed by agents. Indeed, norms have been extensively related to the values that they support [13, 33, 34, 38], though they can also be related to legality [2]. Praiseworthy actions follow a purely ethical perspective [17].

We call our model the *signature* of a moral value to emphasise that we do not try to capture all the complexity and richness of moral values, which is beyond the scope of this work. Instead, we only aim at creating a workable model towards learning value-aligned behaviours.

However, before defining the signature of a moral value, we need to introduce the concept of norm. Norms are coordination mechanisms that regulate (and constrain) the behaviour of agents within a society. They have been extensively studied [8, 9, 27] and are usually expressed in the form of prohibitions (*Prh*), permissions (*Per*) or obligations (*Obl*) over given actions. Most often norms are enforced in societies by means of punishments that are applied to non-compliant agents. There is a myriad of norm definitions in the normative multi-agent systems literature [8, 35]. The norm definition that we consider in this work is based on [26]. In our model we expand their definition by including the concept of associated penalty of a norm. *Penalties* or punishments have also been long studied in the norm research community [32].

Definition 3 (Norm). *A norm is a tuple $\langle c, \theta(a), p \rangle$, where c is a condition for norm application, $\theta \in \{Obl, Per, Prh\}$ is a deontic operator regulating action $a \in \mathcal{A}$, and p is a positive value representing the punishment for non-compliance.*

Note 1. *Notice that the condition c of a norm is a set of first-order predicates $p(\vec{\tau})$, where each p is a k_p-arity predicate symbol and $\vec{\tau} \in \mathcal{T}_1 \times \cdots \times \mathcal{T}_{k_p}$ is a vector of terms, and each $\mathcal{T}_i$ is a set of terms of a first-order language $\mathcal{L}$.*

Punishment p is considered to be a positive penalty, as for specifying the quantity that will be discounted from an agent's outcome upon non-compliance.

Example 1. *In the* public civility game *(further detailed in Section 5), two agents walking in the street come across a piece of garbage. In this context, we can think of a norm n_1 that prohibits to throw garbage at another agent to avoid aggressive behaviours and agents being hurt. Following Def. 3, we define n_1 as:*

$$n_1 = \langle (adj\_agent, front\_garbage), Prh(throw\_to\_agent), \mathsf{p}_1 \rangle \tag{2}$$

As previously mentioned, we consider norms promote (or support) moral values. Moral values are the object of study of moral philosophy or *ethics* [11]. In particular, one of the main questions of relevance to ethics is how we ought to resolve a moral dilemma [5, 16]. Moral theories (such as Kantian or utilitarian ethics) provide guidelines to accomplish ethically-aligned behaviours. These guidelines contain norms and also recommendations [37]. Recommendations are actions that are *good to do but not bad not to do*[2]. They are

[2] https://plato.stanford.edu/entries/supererogation/

strongly related with praiseworthiness, since recommended actions are also worthy of praise, a status that normative actions lack (since the latter ones are the minimum expected for everybody). Hence, recommendations can be regarded as praiseworthy actions.

Therefore, with the aim of giving the agents a framework to learn to behave ethically, we propose that a moral value signature is composed by: *normative* component containing a set of norms that promote the value; and an *evaluative* component defined as an evaluation function that signals how good (praiseworthy) are actions according to the moral value:

Definition 4 (Moral value signature). *The signature of a moral value sgn_v is a pair $\langle \mathcal{N}_v, E_v \rangle$ such that:*

- *$\mathcal{N}_v$ is a finite set of norms promoting the value.*
- *E_v is an action evaluation function that, for a condition c (expressed in a first-order language $\mathcal{L}$) and an action 'a', returns a number in $\mathbb{R}$ meaning the degree of praiseworthiness of that action to the moral value. Thus, given condition c, the bigger $E_v(c, a) > 0$, the more praiseworthy an action 'a' is according to v. Conversely, if $E_v(c, a) < 0$, it means 'a' is considered a blameworthy action, whereas $E_v(c, a) = 0$ represents a neutral action with respect to v.*

Here, $\mathcal{N}_v$ and E_v satisfy the following consistency constraint:

- *Given a norm $n = \langle c, \theta(a), p \rangle \in \mathcal{N}_v$, if n is such that $\theta = Prh$, then $E_v(c, a) < 0$. Otherwise, if $\theta = Per$ or Obl, then $E_v(c, a) \geq 0$.*

To simplify the notation, where there is no confusion, we will write the signature of a moral value v as $sgn = \langle \mathcal{N}, E \rangle$, without sub-indices.

Example 2. *Back to our previous example, in the context of our* public civility game*, we can consider the moral value signature of* civility *sgn_{civ} that: promotes the action of throwing the garbage into the wastebasket and considers that throwing it at other agents is inadmissible. Thus, we include norm n_1 into sgn_{civ} so to formalise civility following Definition 4 as*

$$sgn_{civ} = \langle \{n_1\}, E_{civ} \rangle, \tag{3}$$

where E_{civ} is an evaluation function for the civility moral value defined as: E_{civ}(front_garbage, garbage_to_wastebasket) = $\texttt{eval}_{\texttt{civ}}$, E_{civ}((adj_agent, front_garbage),throw_garbage) < 0 and 0 otherwise (i.e., for any other action and condition), being $\texttt{eval}_{\texttt{civ}} > 0$ positive.

3.2 Extending Markov games with a moral value signature

The next step is to introduce our formalisation of moral value signatures inside the framework of Markov games. The most direct way to do so is to extend the reward function of agents so they take moral values into account. In this subsection we construct this extension step by step.

We first need to define a couple of auxiliary functions to translate the conditions of norms and moral values in terms of states. We begin with the condition function, which describes the states in which the deontic part of a norm holds, that is, where the conditions of the norm hold.

Definition 5 (Condition function). *Given a Markov game with a set of states $\mathcal{S}$ and a first-order language $\mathcal{L}$, with its associated set of predicates $\mathcal{P}(\mathcal{L})$, we define the Condition function $C : \mathcal{S} \to 2^{\mathcal{P}(\mathcal{L})}$ that maps every state to the set of predicates describing the state.*

Next, we proceed with the penalty function, which tells us in which states s an agent would receive a penalty for violating a norm that is enforced (i.e., performing action a when forbidden or failing to perform it when obliged) and what is the value of such penalty.

Definition 6 (Penalty function). *Given a norm $n = \langle c, \theta(x), p \rangle$, and a Markov game with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}^i$ for every agent i, we define the penalty function $P^i_n : \mathcal{S} \times \mathcal{A}^i \to \{0, p\}$ of every agent i as*

$$P^i_n(s, a^i) \doteq \begin{cases} p & \text{if } c \in C(s), \theta = Prh \text{ and } a^i = x \\ & \text{or if } c \in C(s), \theta = Obl \text{ and } a^i \neq x, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where s is a state of $\mathcal{S}$ and a^i is an action of $\mathcal{A}^i$.

With the introduction of the penalty function we can now extend the reward function of a Markov game with a normative component, ensuring that violating norms is penalised.

Definition 7 (Normative extension of a Markov game). *Given a set of norms $\mathcal{N}$ and a Markov game of m agents with reward functions $R_0^{i=1,\ldots,m}$, we define its normative extension as another Markov game such that the reward function R^i for each agent i is defined as $R^i = R^i_0 + R^i_{\mathcal{N}}$, where $R^i_{\mathcal{N}} : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^-$ corresponds to the normative reward function and is defined as*

$$R^i_{\mathcal{N}}(s, a^i) \doteq -\sum_{n \in \mathcal{N}} P^i_n(s, a^i). \tag{5}$$

The normative reward function $R^i_{\mathcal{N}}$ accumulates the penalties (see Eq. 4) of all the norms in $\mathcal{N}$ enforced in a given state-action pair $\langle s, a^i \rangle$.

Now that we have a method for incorporating norms in Markov games, we can introduce the signature of a moral value in the same vein. Thus, following Definition 4, our ethical extension of Markov games has: i) a normative component identical to the one in Definition 7, and ii) an evaluative component that rewards praiseworthy actions.

Definition 8 (Ethical extension of a Markov game). *Given a moral value signature $sgn = \langle \mathcal{N}, E \rangle$ and a Markov game of m agents with reward functions $R_0^{i=1,\ldots,m}$, we define its ethical extension as another Markov game such that the reward function R^i of each agent i is defined as $R^i = R^i_0 + R^i_{\mathcal{N}} + R^i_E$, where $R^i_{\mathcal{N}} : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^-$ is the normative reward function of norm set $\mathcal{N}$ applied to agent i and $R^i_E : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^+$ is is a function of the form*

$$R^i_E(s, a^i) = \max(0, E(C(s), a^i)). \tag{6}$$

We will refer to R^i_E as the evaluative reward function of a moral value signature, which rewards praiseworthy actions performed under certain conditions.

Notice that the evaluative reward function R^i_E from Eq. 6 is just an adaptation of the action evaluation function E from Def. 4 so it can be used in Markov games, that have states instead of predicates.

3.3 Defining ethically-aligned policies

Thanks to Definition 8, we can extend the agents' rewards in a Markov game to incorporate moral values. Thereafter, we move a step further and define an ethically-aligned policy as one such that the agent minimises the accumulation of normative punishments and maximises the accumulation of evaluative rewards coming from performing praiseworthy actions.

Likewise in previous subsections, we create the concept of an ethically-aligned policy gradually. We start by defining norm compliant policies as those that accumulate no normative penalty, and then we expand this concept to define ethically-aligned policies as policies that are norm-compliant and also accumulate the maximum possible evaluative reward.

Prior to these definitions, it would be useful to count on functions that measure the accumulation of normative and evaluative rewards respectively. As explained in the background section above, Markov games already have a function for the accumulation of reward for each agent i: the state value function V^i. Furthermore, since, according to Def. 8, in an ethically-extended Markov game the reward can always be divided in three components ($R^i = R^i_0 + R^i_{\mathcal{N}} + R^i_E$), we will also divide the state value function V^i in three components ($V^i = V^i_0 + V^i_{\mathcal{N}} + V^i_E$) in order to obtain our desired functions. Formally:

Definition 9 (Normative and evaluative state value functions). *Given a Markov game with state value functions V^i_0, and a moral value signature $sgn = \langle \mathcal{N}, E \rangle$, we define the random variables $R^i_{\mathcal{N}_t}$ and $R^i_{E_t}$ such that they re-express the normative reward function $R^i_{\mathcal{N}}$ and the evaluative reward function R^i_E in the ethical extension in the following way:*

$$R^i_{\mathcal{N}}(s, a^i) = \mathbb{E}[R^i_{\mathcal{N}_{t+1}} \mid S_t = s, A^i_t = a^i], \tag{7}$$

$$R^i_E(s, a^i) = \mathbb{E}[R^i_{E_{t+1}} \mid S_t = s, A^i_t = a^i], \tag{8}$$

where S_t and A_t are random variables. Moreover, we can respectively define the normative state value function $V^i_{\mathcal{N}}$ and the evaluative state value function V^i_E of an agent i as:

$$V^i_{\mathcal{N}_\pi}(s) \doteq \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{\mathcal{N}_{t+k+1}} \mid \pi, S_t = s], \tag{9}$$

$$V^i_{E_\pi}(s) \doteq \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{E_{t+k+1}} \mid \pi, S_t = s]. \tag{10}$$

Note that a policy π^i that never violates any norm in a set $\mathcal{N}$ will not receive a penalisation for its behaviour. Consequently, it will generate no accumulated normative reward $V^i_{\mathcal{N}_{\langle \pi^i, \pi^{-i} \rangle}}$. We will refer to such policies as norm-compliant.

Definition 10 (Norm-compliant policy). *Given a Markov game $\mathcal{M}$ and a set of norms $\mathcal{N}$, we say that π^i is a norm-compliant policy with respect to $\mathcal{N}$ if and only if for every state s of the normative extension of $\mathcal{M}$:*

$$V^i_{\mathcal{N}_{\langle \pi^i, \pi^{-i} \rangle}}(s) = 0. \tag{11}$$

We can make a similar observation for a policy π^i that acts on the most praiseworthy way possible according to an evaluation function E of some moral value signature $\langle \mathcal{N}, E \rangle$. Such policy will have the maximum possible accumulated evaluative reward $V^i_{E_{\langle \pi^i, \pi^{-i} \rangle}}$ that can be obtained. We will refer to those policies as praiseworthy.

Definition 11 (Praiseworthy policy). *Given a Markov game $\mathcal{M}$ and a moral value signature $sgn = \langle \mathcal{N}, E \rangle$, we say that π^i is a praiseworthy policy with respect to E if and only if for every state s of the ethical extension of $\mathcal{M}$:*

$$V^i_{E_{\langle \pi^i, \pi^{-i} \rangle}}(s) = \max_{\rho^i} V^i_{E_{\langle \rho^i, \pi^{-i} \rangle}}(s). \tag{12}$$

With these two definitions we can conclude this subsection enunciating that a policy is ethically-aligned if it is both norm-compliant and praiseworthy.

Definition 12 (Ethically-aligned policy). *Given a Markov game $\mathcal{M}$ and a moral value signature $sgn = \langle \mathcal{N}, E \rangle$, a policy π^i is ethically-aligned with respect to sgn if and only if it is norm-compliant with respect to $\mathcal{N}$ and praiseworthy with respect to E.*

We will also use the term *ethically-aligned joint policy* when every agent follows an ethically-aligned policy with respect to a moral value signature *sgn*.

Notice that ethically-aligned policies with respect to a given *sgn* do not necessarily exist. The trivial example would be a Markov game with one state s and only one action a that violates some norm n of a moral value signature. For that reason, we need to differentiate between two kinds of Markov games: those for which an ethically-aligned policy is attainable and those for which it is not.

Definition 13 (Ethically-attainable Markov game). *Given a Markov game $\mathcal{M}$ and a moral value signature sgn, then $\mathcal{M}$ is ethically-attainable with respect to sgn if and only if there is at least one joint policy π ethically-aligned to sgn in $\mathcal{M}$.*

3.4 Characterising sequential moral dilemmas

With ethically-aligned policies characterised by Definition 12, we are finally prepared to define sequential moral dilemmas as Markov games such that, if every agent just follows its individual interests (i.e. by maximising its V^i), then, the result is an equilibrium joint policy that is not ethically-aligned. In game-theoretical terms [21], we will also refer to such equilibria as *ethically deficient.*

Definition 14 (Sequential moral dilemma). *Let $\mathcal{M}$ be a Markov game, sgn_v the signature of a moral value v, Π_* the set of all Nash equilibria, and Π_v the set of all ethically-aligned joint policies with respect to sgn_v. Then $\mathcal{M}$ is a sequential moral dilemma with respect to sgn_v if and only if*

- *there is at least one Nash equilibrium that is not ethically-aligned with respect to sgn_v (i.e., $\Pi_* \nsubseteq \Pi_v$); and*
- *the Markov game $\mathcal{M}$ is ethically-attainable with respect to sgn_v (i.e., $\Pi_v \neq \varnothing$).*

In a SMD, we want the agents to avoid ethically-deficient NE. For that reason we consider that a SMD is solved when agents learn an ethically-aligned Nash Equilibrium. Next section details how we propose to solve them.

4 A STRUCTURAL SOLUTION FOR SEQUENTIAL MORAL DILEMMAS

As mentioned above, SMDs are Markov games in which agents may learn to behave unethically if they solely follow their individual goals. Hence, in SMDs there are NE not ethically-aligned and we aim at solving them by avoiding those ethically-deficient NE.

The game theory community has long studied problems where there exist deficient NE under the label of social dilemmas. They have proposed three alternative solutions: strategic, motivational, and structural [21]. Strategic solutions assume egoistic actors, motivational solutions assume that agents are not completely egoistic, and structural solutions change the rules of the game.

As a starting point in the study of SMDs, this paper proposes a structural solution ensuring that agents learn to pursue an ethically-aligned policy. Specifically, this solution extends the Markov game of a SMD into a new one that is no longer a dilemma. More formally, if the problem of SMDs is that the set of NE Π_* is not a subset of the set of ethically-aligned joint policies Π_v, we will transform the game to ensure that Π_* is indeed a subset of Π_v.

As explained in the previous section, the natural way to create such extension will be to reshape the reward functions of the game through an ethical extension following Def. 8.

In a Markov game, there always exists at least one NE [10]. Hence, our structural solution will extend the rewards so that no *ethically-deficient* joint policy can be a NE in the extended Markov game. By elimination, any remaining Nash equilibrium will be ethically-aligned. The only condition for application of this approach is that ethically-aligned policies do exist in the original Markov game in the first place (i.e., it is ethically-attainable).

Likewise in previous sections, we present our structural solution step by step. First we characterise the properties that any structural solution extending the rewards must fulfil and then we offer our particular solution. We start with an initial result observing that in a Markov game, every NE is ethically-aligned if and only if an ethical policy is always the best response. Or, in other words, that an unethical policy is never the best response. That is formally captured by the following lemma:

Lemma 1. *Given a Markov game, every Nash equilibrium joint policy is ethically-aligned if for every joint policy π with at least one agent i such that π^i is not ethically-aligned, there is at least one state s such that $V^i_{\langle \pi^i_*, \pi^{-i} \rangle}(s) > V^i_{\langle \pi^i, \pi^{-i} \rangle}(s)$ for some other ethically-aligned policy π^i_* in $\langle \pi^i_*, \pi^{-i} \rangle$.*

Proof. Apply the contrapositive of Def. 2. □

From this lemma we know that any structural solution must extend the Markov game so that being ethical is the best response in the extended Markov game. With that, we are ready to characterise the conditions that must hold for a SMD so that its ethical extension is not a SMD. In other words, the conditions that guarantee that in its extension agents always decide to behave ethically. For that, we just need to impose that the conditions of Lemma 1 hold for the extended Markov game.

Theorem 1 (Structural solutions characterisation). *Given a sequential moral dilemma $\mathcal{M}_0$ with respect to sgn_v, the ethical extension $\mathcal{M}$ of $\mathcal{M}_0$ is not a sequential moral dilemma if for every joint policy π with at least one agent i such that π^i is not ethically-aligned, there is at least one state s such that*

$$V^i_{\langle \pi^i_*, \pi^{-i} \rangle}(s) > V^i_{\langle \pi^i, \pi^{-i} \rangle}(s) \tag{13}$$

for some other ethically-aligned policy π^i_* in $\langle \pi^i_*, \pi^{-i} \rangle$.

Proof. Extension $\mathcal{M}$ is not a SMD if every NE is ethically-aligned. Use Lemma 1 to reword the relation as in Theorem 1. □

Theorem 1 is telling us that an ethical extension will solve the dilemma if and only if there is a reward surplus from being ethical.

Since Theorem 1 does not specify for which states inequation 13 must hold for every Nash equilibrium to be ethically-aligned, we can assume that, in particular, it must hold at the initial state. For Markov games that have more than one initial state, we can simply divide them in several sub-Markov games with a different unique initial state each. Therefore, without loss of generality, we are going to assume from this point onwards that a Markov game has only one initial state s_0.

Corollary 1. *Given a sequential moral dilemma $\mathcal{M}_0$ with respect to a moral value signature sgn_v, the ethical extension $\mathcal{M}$ of $\mathcal{M}_0$ is not a sequential moral dilemma if for every joint policy π with at least one agent i such that π^i is not ethically-aligned*

$$V^i_{\langle \pi^i_*, \pi^{-i} \rangle}(s_0) > V^i_{\langle \pi^i, \pi^{-i} \rangle}(s_0) \tag{14}$$

at the initial state s_0 for some other ethically-aligned policy π^i_* in $\langle \pi^i_*, \pi^{-i} \rangle$.

Proof. An initial state $s = s_0$ is still a state, so by Theorem 1 the implication is true. □

In the particular case of a Markov game $\mathcal{M}_0$ with only one initial state s_0, Corollary 1 tells us exactly where we need to check the inequality. This corollary tells us that by conveniently setting the values for penalties for violating norms and rewards for praiseworthy actions, no unethical policy will be a best response because we will always have a better alternative (that is also ethically-aligned). And in order to find these values, it will suffice to check the inequalities at the initial state.

Corollary 1 presents the minimal conditions that any structural solution affecting the initial state s_0 must fulfil. In particular, the solution here presented requires a more demanding condition so we can detect if we have chosen the correct sets of penalties and ethical rewards via checking only one inequality. Our solution demands that, for every agent, even the best non-ethically-aligned policy provides a worse payoff than the ethically-aligned best-response policy in the worst situation for being ethically-aligned. Without further ado, we present our formula to solve a SMD:

Corollary 2 (Structural solution). *Given a sequential moral dilemma $\mathcal{M}_0$ with respect to a moral value signature sgn_v, the ethical extension $\mathcal{M}$ of $\mathcal{M}_0$ is not a sequential moral dilemma if for every agent i:*

$$\min_{\pi^{-i}} V^i_{\langle BR^i_v(\pi^{-i}), \pi^{-i} \rangle}(s_0) > \max_{\rho \notin \Pi^i_v} V^i_{\langle \rho^i, \rho^{-i} \rangle}(s_0) \tag{15}$$

at the initial state s_0. Here, Π^i_v is the subset of joint policies where at least the agent i is ethically-aligned, and BR^i_v is a function that

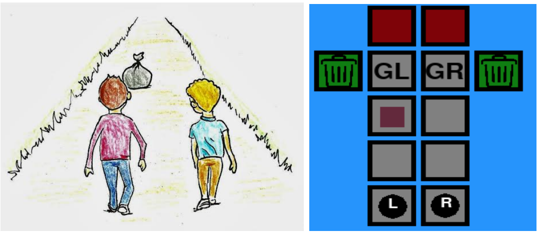

Figure 1: Left: garbage blocking the path of the agent at the left. Right: Our simulation representing the same state.

returns, for any joint policy π^{-i}, the best-response policy π^i subject to being ethically-aligned with respect to sgn_v.

Proof. Cor. 2 is a particular case of Cor. 1. □

Corollary 2 proves that any SMD can be solved. We only need to select the values to set normative penalties and evaluative rewards so inequality 15 holds for every agent. However, while checking the inequation is a simple calculation from a mathematical point of view, it can be computationally expensive for MG's relatively big.

In order to illustrate how our structural solution can be applied in a small SMD, we present in next section the *public civility game*.

5 AN EXAMPLE OF SMD: THE PUBLIC CIVILITY GAME

The *public civility game* is a SMD in which two agents move every day from their initial position to their destinations. At some point, they find a garbage obstacle blocking the way of one agent, who may decide how to deal with it by considering (or not) the moral value of civility. This value demotes the violence of throwing the garbage to other agents and praises throwing the garbage to a wastebasket. Left-hand-side of Figure 1 illustrates the game.

The right image in Figure 1[3] depicts how we model our case study as a multi-agent system consisting on a 2-dimensional grid, where two agents traverse grey cells in their way towards their destination. For illustrative purposes, we represent agents as black circles –labelled as L (Left) and R (Right)– whose starting positions are the ones depicted in the figure and their destination (Goal) cells appear marked as GL and GR respectively. Moreover, two agents cannot populate the same cell simultaneously. Initially, the garbage –which is depicted as a purple square– is randomly located at any of the grey cells except for the initial positions of the agents.

Time is discrete and measured in time-ticks. An episode or day (which lasts for Max_t ticks at most) corresponds to the period of time both agents need to reach their destinations. Every tick agents are allowed to perform a single action: moving to an adjacent cell or pushing the garbage if it is located in front.

As for the pure Markov game setting, we consider a *state* $s \in S$ to be defined as $s = \langle cell^L, cell^R, cell^G \rangle$ where $cell^L$ and $cell^R$ correspond to the position (cell) of agents L and R respectively and $cell^G$ identifies the position of the garbage.

The set of *actions* each agent can perform in every scenario is $\mathcal{A} = \{mF, mR, mL, pF, pR, pL\}$, where m stands for **m**ovement, p for **p**ush, F=Forward, R=Right, and L=Left. Actions mF, mR, and mL imply a change (if possible) in the agent position ($s.cell^L$ or $s.cell^R$), whereas actions pF, pR, and pL will change the garbage's position ($s.cell^G$) whenever the garbage is in front of the agent.

As for the *reward functions*, considering $s \in S$ to be the current state, $a^L \in \mathcal{A}$ the action agent L performs, $a^R \in \mathcal{A}$ the action agent R performs, and $s' \in \mathcal{S}$ such that $\langle s, a^L, a^R, s' \rangle$ is a transition, we define a deterministic reward function $R^i(s, a^L, a^R, s')$ for each agent, with $i \in \{L, R\}$ to identify the agent that it is associated with.

Each agent's individual goal is to reach its respective destination Gi (GR or GL) as fast as possible while avoiding getting hurt, thus

$$R_0^i(s, a^L, a^R, s') \doteq \begin{cases} Max_t & \text{if } s'.cell^i\text{=Gi and } s'.cell^i \neq s'.cell^G, \\ Max_t - h & \text{if } s'.cell^i\text{=Gi and } s'.cell^i = s'.cell^G, \\ -h - 1 & \text{otherwise if } s'.cell^i = s'.cell^G, \\ -1 & \text{otherwise.} \end{cases} \tag{16}$$

By penalising the agent with a reward of -1 for being in any position except its goal, we are encouraging it to never stop until it gets to its goal. We also penalise getting hurt with a detrimental reward of h so agents try to avoid it. It is important to remark that other formulations may be perfectly valid as well.

Finally, we describe three possible policies that an agent might choose from upon encountering the garbage in front of it:

(1) **Unethical policy:** push the garbage away to reach the goal as fast as possible.
(2) **Regimented policy:** wait until the other agent is not nearby in order to push it away without hurting anybody. This policy is compliant with norm n_1 defined in Eq. 2.
(3) **Ethical policy:** push it all the way to the nearest wastebasket. This policy is ethically-aligned with *civility* as defined in Eq. 3. Hence, this is the policy that we would want the

6 SOLVING THE PUBLIC CIVILITY GAME

We now apply our structural solution to the public civility game to extend it to a new game where agents learn to behave civilly. Afterwards, we let the agents choose their policy using Q-learning, a classical reinforcement learning algorithm. Once they have finished learning, we evaluate the behaviour of our agents through several experiments. Specifically, we ascertain whether the agents learn an ethically-aligned NE: we check that each agent manages to find a balance between pursuing its individual interests (reach the goal as fast as possible) and societal ones (promote civility). Moreover, we use several social behaviour measures to also assess if the multi-agent society improves (as a whole) when they perform ethically.

Results illustrate (and corroborate) our theoretical findings and show that agents can readily learn to behave ethically using a simple RL algorithm if the environment structure is properly shaped.

6.1 Simulation Settings

In our experiments, we consider the following settings. The maximum amount of time-ticks per episode is set to $Max_t = 20$, likewise the reward function in Eq. 16 considers $Max_t = 20$, The damage for being hurt is $h = 3$. The discount factor is set to $\gamma = 0.7$.

[3] Drawing courtesy of Jordi Reyes Iso.

		Agent R	
		E	U
Agent L	E	5.30 5.30	6.38 4.37
	U	4.37 6.38	**5.45** **5.45**

Table 1: Payoff matrix of the public civility game. Agent actions correspond to an unethical policy (U) and an ethically-aligned policy (E). NE (in bold) is ethically-deficient.

With these settings, Table 1 shows the expected return $V_{\pi}^{i}(s_0)$ (i.e., expected accumulated rewards per episode averaged for the different initial states[4] s_0) for the different joint policies. Notice that the public civility game corresponds to a sequential moral dilemma with the NE in the U-U (non-ethically-aligned) joint policy.

6.2 Solution

In order to ensure that agents learn an ethically-aligned policy, we use our structural solution as explained in section 4. We do so by extending the reward function of the Markov game defined in subsection 5 in a way that shapes agents' policies with ethical components $R^i = R_0^i + R_{\mathcal{N}}^i + R_E^i$ following Definition 8.

More in detail, we define the normative reward function $R_{\mathcal{N}}^i$ instantiating Eq. 5:

$$R_{\mathcal{N}}^i(s, a^i) = -P_{n_1}^i(s, a^i), \tag{17}$$

and following Eq. 6, the evaluative reward function R_E^i becomes:

$$R_E^i(s, a^i) = \max(0, E_{civ}(C(s), a^i)). \tag{18}$$

where $E_{civ}(C(s), a^i)$ only returns $eval_{civ}$ from Eq. 3 if agent i performs any garbage pushing action (pF, pR or pL) that will put the garbage into a wastebasket, and returns 0 or less otherwise.

Using our structural solution defined in Corollary 2, we have to set p_1 and $eval_{civ}$ so even the ethically-aligned best-response in the worst case (which from the point of view of agent L corresponds to the case E-U from Table 1) is better than the best possible non-ethically-aligned policy (which from the point of view of agent L corresponds to the case U-E from Table 1).

To ensure that inequality 15 holds, we set a punishment of $\mathrm{p}_1 = 10$ for not complying with norm n_1 (see equation 2) and a reward of 10 for behaving civilly $\texttt{eval}_{\texttt{civ}} = 10$ in equation 3. Other settings might be valid as well, since the inequality has infinite solutions.

6.3 Social behaviour metrics

It may seem reasonable to think of a society composed by ethical agents as a good one. In order to assess it, we can compare the payoffs obtained in an ethical scenario versus an unethical one, as we actually do in subsection 6.5. However, there are some global aspects that can improve in an ethical scenario that are hard to study by merely focusing on the rewards that individual agents receive. For that reason, we have defined four *social behaviour metrics* [23] for our *public civility game.*

These four metrics measure the accomplishment of the societal goals of the game: that agents reach their goals in a reasonable time, that agents do not get hurt, and that streets are kept clean:

- `Time:` measures the average time-ticks each agent needs to get to its goal.
- `Violence:` measures the degree of harmfulness of the society as the ratio of episodes where an agent is hurt.
- `Semi-civility:` measures the number of episodes in which the garbage ends up being on a side place without obstructing agents' way (i.e., red cells in Figure 1) divided by the total number of testing episodes.
- `Civility:` measures the number of episodes in which the garbage ends up being on a wastebasket (i.e., green cells in Figure 1) divided by the total number of testing episodes.

6.4 Experiments

We compare the aforementioned social behaviour metrics and also study the evolution of the obtained rewards in three scenarios.

First, an **unethical scenario** that corresponds to the original Markov game. It represents an unregulated society where agents only act on behalf of their own interests. This kind of amoral societies has been long studied by moral philosophy and moral politics under the name *state of nature* [7, 15, 25].

A second, **ethical scenario** that corresponds to our ethically-extended Markov game with respect to *civility.* It is a more sophisticated scenario that represents the interactions of agents that have internalised the moral value of civility. Moral philosophers have also been interested in these proper –civil– societies that they study under the name of *social contract* [30, 31].

A third, **regimented scenario**, that corresponds to a normative extension of the Markov game with respect to norm n_1. To complete the picture, we also study this intermediate scenario, that represents a society where agents have not fully internalised the moral value of civility but only its minimal, normative part. Similar scenarios have been studied in moral philosophy and psychology, being the closest example the *intermediate stages of moral reasoning* of Kohlberg's theory of moral development [20].

In each scenario, we use *reinforcement learning* (RL) in order to let agents select the policy they want to achieve. We consider this a natural solution for our problem if we take into account that we have framed the *public civility game* as a Markov game.

In particular, agents use **Q-learning** [41] to learn their policies. It is both easy to implement –since it is a model-free off-policy algorithm– and capable of finding an optimal solution under the right conditions. However, we consider it as an initial attempt to tackle our problem, prior to trying more sophisticated algorithms in further research. As for the training policy for Q-learning prior to agents switching to their learnt policies, we use the well-known ϵ-greedy policy [36] with a learning rate $\alpha = 0.5$.

In order to minimise the effects of randomness in the evaluation, we repeat training-testing experiences (where each experience lasts for 3000+1000 = 4000 episodes) 300 times per scenario.

6.5 Results

The reported results show the average metrics of the $3 \cdot 300 = 900$ experiments. The social behaviour metrics are measured after the

[4] There are 6 initial states corresponding to the random initial positions of the garbage.

Scenario	`Time`	`Viol.`	`Semi-civ.`	`Civility`
Unethical	3.68 ± 0.1	0.63	0.13 ± 0.0	0.13 ± 0.0
Regimented	4.05 ± 0.1	0.0	0.45 ± 0.1	0.45 ± 0.1
Ethical	4.08 ± 0.1	0.0	0.0 ± 0.0	1.0 ± 0.0

Table 2: Results in terms of our performance measures.

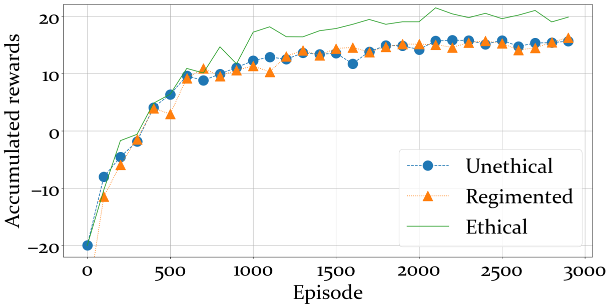

Figure 2: Evolution of the accumulated rewards per episode in the three scenarios: unethical, regimented, and ethical.

agents finish training, whereas the reward analysis is measured while the agents are learning.

6.5.1 Social behaviour metrics. Table 2 shows the results in terms of our social behaviour metrics. The first row shows that in the base-line *unethical* scenario agents take an average time of 3.68 ticks per trip, which represents a 23% of increment compared to the 3 ticks required for reaching the goal position without the garbage blocking the way. The level of `Violence` is 63%, which indicates this is a wild, aggressive scenario. As for `Civility`, both agents learn to behave civilly only 13% of times because the garbage ends up on a grey cell (i.e., blocking the way) 74% of the times, and the remaining 26% is equally distributed among red and green (wastebasket) cells.

The *regimented* scenario (see second raw in Table 2) tackles the undesirably high aggressiveness in the unethical scenario by enacting norm n_1. Thus, agents learn this norm-compliant behaviour in order to avoid the associated punishment. The effects of reducing `Violence` down to 0 are two-fold. First, `Time` increases a 10%. Second, the garbage ends up blocking the way far less times (10%) and `Civility` and `Semi-Civility` increase because agents distribute the garbage equally between red and green cells (45% each).

As for our *ethical* scenario (see third raw in Table 2), it does not only keep `Violence` down to 0, but also increases `Civility` up to 1 by always throwing the garbage to the wastebasket. Obviously, there is a price to pay related to the extra `Time` agents take to tidy up the street. Thus, agents learn to sacrifice part of their individual goal of reaching their goal as fast as possible to avoid violence and to have clean streets, showing a praiseworthy behaviour.

6.5.2 Reward analysis. Figure 2 shows the averaged accumulated reward that the agent obtains per episode[5], which is the sum of all the rewards the agent obtains during an episode[6].

The unethical (blue) curve serves as the baseline curve. We can appreciate that it starts at less than -20 (meaning that the agent cannot even get to the goal position) and quickly this value rises in less than 500 episodes up to 10. We observe that in 2000 episodes it finally stabilises at around 15. This seems reasonable if we consider that the maximum possible accumulated reward (when no garbage blocks the way) is $Max_t - 3 = 17$, where $Max_t = 20$ and 3 comes from the 3 cells that an agent has to cross to get to its goal position.

The regimented (orange) curve in Fig. 2 is almost equal to the unethical one, except that it sometimes has a lower value due to norm violations. We can see that at the end this difference is hard to detect, which means that the agent has learnt to comply with n_1 (see Eq. 2), the norm in place.

The ethical (green) curve is always the one that grows the most (getting to up to 21), which was to be expected since only in the ethical scenario the reward function gives an extra positive reward associated with throwing the garbage to the wastebasket. Specifically, the maximum reward it can get is $(Max_t + eval_{civ}) - (3+d) = 27 - d$, where the $3 + d$ comes from considering that the agent will need to move itself thrice and also push the garbage d times. Considering that on average d will have a value of 2, and that the agent only gets the $eval_{civ}$ surplus half of the times (when the wastebasket is on its side) its reward should stabilise at around $(25 + 17)/2 = 21$ which is exactly what it does. This indicates us that the agent has both learnt to throw the garbage to the wastebasket (to behave ethically) and also an optimal policy from its point of view.

After studying analytically all these curves (and particularly the one from the ethical scenario) we can claim that both agents always manage to learn the best possible policy (since all the curves stabilise at the highest possible reward values), and therefore we obtain a Nash Equilibrium joint policy (that is also ethically-aligned in the ethical scenario). In case you are interested, we have made available some videos showing the learnt behaviours of agents in all three scenarios [7].

We finish this subsection by remarking that these empirical results are just a consequence of what was already asseverated by Theorem 1: with the proper setting of our moral value signature, every Nash equilibrium becomes ethically-aligned.

7 CONCLUSIONS

This paper proposes the inclusion of ethical aspects into Markov game settings. In particular, we study value-alignment and propose the so-called *Sequential Moral Dilemma* (SMD), which considers the signature of a moral value. Subsequently, we characterise ethically-aligned agent policies and discuss how to obtain them. Our solution consists on extending the rewards of the Markov game with an ethical component that ensures all NE become ethically-aligned.

We illustrate our proposal with the *Public Civility game* and solve it with the tools herein presented. We empirically show that the multi-agent society improves its overall performance in terms of street cleanness and agents' aggressiveness reduction.

As future work, we would like to further explore the formal relationship between SSDs and SMDs, as well as the algorithmic complexity of our structural solution.

[5] Without lose of generality all results here only refer to the L agent, which are extremely similar to the results for agent R.

[6] For the sake of reducing the noise produced by the randomness while training, we average these accumulated rewards considering a sliding window of last 100 episodes.

[7] Unethical policy: https://youtu.be/20W3rAEpgJY. Regimented policy: https://youtu.be/ICjrCNCCjcQ. Ethical policy: https://youtu.be/ZgM0vmlRvCU

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] Guido Boella, Leendert W.N. van der Torre, and Harko Verhagen. 2007. Introduction to Normative Multiagent Systems. In *Normative Multi-agent Systems.*

[3] Nick Bostrom and Eliezer Yudkowsky. 2011. Ethics of Artificial Intelligence. *Cambridge Handbook of Artificial Intelligence* (2011).

[4] Ryan Calo. 2017. Artificial Intelligence Policy: A Primer and Roadmap. https://doi.org/10.2139/ssrn.3015350

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

[6] R M Dawes. 1980. Social Dilemmas. *Annual Review of Psychology* 31, 1 (1980), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125 arXiv:https://doi.org/10.1146/annurev.ps.31.020180.001125

[7] Benedictus de Spinoza. 1883. *A Theologico-Political Treatise.* Dover Publications.

[8] Frank Dignum. 1999. Autonomous Agents with Norms. *Artif. Intell. Law* 7, 1 (1999), 69–79.

[9] F. Dignum. 1999. Autonomous Agents with Norms. *Artificial Intelligence and Law, 7: 69* (1999). https://doi.org/10.1023/A:1008315530323

[10] A. M. Fink. 1964. Equilibrium in a stochastic n-person game. *J. Sci. Hiroshima Univ. Ser. A-I Math.* 28, 1 (1964), 89–93. https://doi.org/10.32917/hmj/1206139508

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

[12] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Venable, and Brian Williams. 2016. Embedding Ethical Principles in Collective Decision Support Systems. (2016).

[13] Sven Ove Hansson. 2001. *The structure of values and norms.* Cambridge University Press.

[14] Garrett Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (1968), 1243–1248. https://doi.org/10.1126/science.162.3859.1243 arXiv:https://science.sciencemag.org/content/162/3859/1243.full.pdf

[15] Thomas Hobbes. 1651. *Leviathan, 1651.* Menston, Scolar P.

[16] Robert L. Holmes. 1990. The Limited Relevance of Analytical Ethics to the Problems of Bioethics. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 15, 2 (04 1990), 143–159. https://doi.org/10.1093/jmp/15.2.143 arXiv:http://oup.prod.sis.lan/jmp/article-pdf/15/2/143/2681996/15-2-143.pdf

[17] Terry Horgan and Mark Timmons. 2010. Untying a knot from the inside out: Reflections on the "paradox" of supererogation. *Social Philosophy and Policy* 27 (07 2010), 29 – 63. https://doi.org/10.1017/S026505250999015X

[18] Junling Hu and Michael P. Wellman. 2003. Nash Q-learning for General-sum Stochastic Games. *J. Mach. Learn. Res.* 4 (Dec. 2003), 1039–1069. http://dl.acm.org/citation.cfm?id=945365.964288

[19] 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.*

[20] Lawrence Kohlberg, Charles Levine, and A. Hewer. 1983. Moral Stages: a Current Formulation and a Response to Critics.

[21] Peter Kollock. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183 arXiv:https://doi.org/10.1146/annurev.soc.24.1.183

[22] 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.

[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). arXiv:1702.03037 http://arxiv.org/abs/1702.03037

[24] 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 (ICML'94).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163. http://dl.acm.org/citation.cfm?id=3091574.3091594

[25] John Locke. 1967. *Two Treatises of Government.* Cambridge: Cambridge University Press.

[26] Javier Morales, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Wamberto Vasconcelos, and Michael Wooldridge. 2015. Online automated synthesis of compact normative systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 10, 1 (2015), 33.

[27] Javier Morales, Maite López-Sánchez, Juan Antonio Rodríguez-Aguilar, Michael Wooldridge, and Wamberto W. Vasconcelos. 2015. Synthesising Liberal Normative Systems. *Proceedings of the fourteenth International Conference on Autonomous Agents and Multiagent Systems, Wiley* (2015).

[28] Gonçalo Neto. 2005. From Single-Agent to Multi-Agent Reinforcement Learning: Foundational Concepts and Methods. (2005).

[29] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. *Version 2. IEEE* (2017).

[30] John Rawls. 1958. Justice as Fairness. *Philosophical Review* 67, 2 (1958), 164–194. https://doi.org/10.2307/2182612

[31] Jean-Jacques Rousseau. 1950. *The Social Contract.* New York: Harmondsworth, Penguin.

[32] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7 (2011), 21–54.

[33] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Javier Morales, Michael Wooldridge, and Carlos Ansotegui. 2018. Exploiting moral values to choose the right norms. In *Proceedings of the 1st Conference on artificial intelligence, ethics and society (AIES'18).* 1–7. https://doi.org/10.1145/3278721.3278735

[34] Marc Serramia, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, and Carlos Ansotegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18).* International Foundation for Autonomous Agents and Multiagent Systems, 1294–1302.

[35] Yoav Shoham and Kevin Leyton-Brown. 2009. *Multiagent Systems - Algorithmic, Game-Theoretic, and Logical Foundations.* Cambridge University Press.

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

[37] James O. Urmson. 1958. Saints and Heroes. In *Essays in Moral Philosophy*, A. I. Melden (Ed.). University of Washington Press.

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

[39] Wendell Wallach. 2008. Implementing Moral Decision Making Faculties in Computers and Robots. *AI and Society* 22, 4 (2008), 463–475. https://doi.org/10.1007/s00146-007-0093-6

[40] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 683–692. http://dl.acm.org/citation.cfm?id=3306127.3331756

[41] Christopher J. C. H. Watkins and Peter Dayan. 1992. Technical Note Q-Learning. *Machine Learning* 8 (1992), 279–292. https://doi.org/10.1007/BF00992698

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