Ethics and Information Technology (2022) 24:9
https://doi.org/10.1007/s10676-022-09635-0

ORIGINAL PAPER

Instilling moral value alignment by means of multi-objective reinforcement learning

Manel Rodriguez-Soto[1] · Marc Serramia[1] · Maite Lopez-Sanchez[2] · Juan Antonio Rodriguez-Aguilar[1]

Accepted: 8 January 2022 / Published online: 24 January 2022

Abstract

AI research is being challenged with ensuring that autonomous agents learn to behave ethically, namely in alignment with moral values. Here, we propose a novel way of tackling the value alignment problem as a two-step process. The first step consists on formalising moral values and value aligned behaviour based on philosophical foundations. Our formalisation is compatible with the framework of (Multi-Objective) Reinforcement Learning, to ease the handling of an agent's individual and ethical objectives. The second step consists in designing an environment wherein an agent learns to behave ethically while pursuing its individual objective. We leverage on our theoretical results to introduce an algorithm that automates our two-step approach. In the cases where value-aligned behaviour is possible, our algorithm produces a learning environment for the agent wherein it will learn a value-aligned behaviour.

Keywords Value alignment · Reinforcement learning · Multi-objective reinforcement learning · Ethics

Introduction

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

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

The literature is populated with reward specification approaches that encode ethical knowledge directly from observing human behaviour, which is presumed to be ethical (e.g. Hadfield-Menell et al., 2016; Noothigattu et al., 2019; Riedl and Harrison, 2016), or from a human that directly gives ethical feedback to the agent in form of rewards (e.g. Balakrishnan et al., 2019). These approaches are convenient because they relieve the agent designer from the burden of defining the expected ethical behaviour of the agent for every possible situation. However, these approaches also suffer from well-known shortcomings, as discussed in Arnold et al. (2017), Tolmeijer et al. (2021), Gabriel (2020): (1)

✉ Manel Rodriguez-Soto
manel.rodriguez@iiia.csic.es

Marc Serramia
marcserr@iiia.csic.es

Maite Lopez-Sanchez
maite_lopez@ub.edu

Juan Antonio Rodriguez-Aguilar
jar@iiia.csic.es

[1] Artificial intelligence research institute (IIIA-CSIC), Carrer de Can Planas, Campus de la UAB, 08193 Bellaterra, Spain

[2] Department of Mathematics and Computer Science, University of Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain

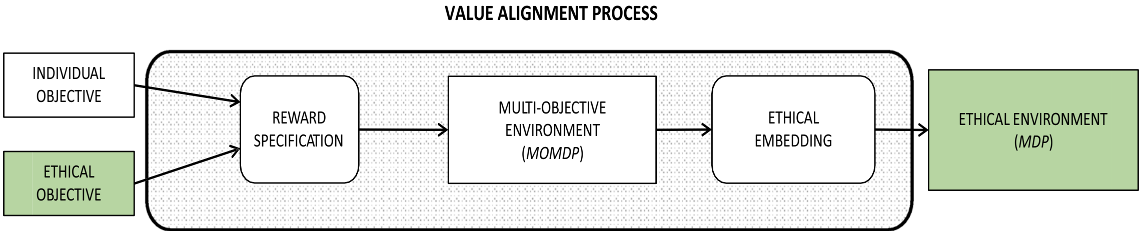


Fig. 1 The value alignment process is performed in two steps: a reward specification and an ethical embedding. Rectangles stand for objects whereas rounded rectangles correspond to processes

observing (learning from) human behaviour may ensure alignment with human habits but does not guarantee the learnt behaviour to be ethical; (2) the knowledge acquired by the agent through learning condenses experience in a manner that lacks of explicit representation (and reasoning) of the moral considerations that need to be taken into account (such as moral norms).

All the above-mentioned shortcomings are specially relevant when there are some moral norms that must be wholly fulfilled (e.g., a robot in charge of buying an object should never decide to steal it Arnold et al. (2017)). For those cases, we argue that reward specification cannot be done by only observing human behaviour, and thus, we instead require an approach that is also rooted in solid philosophical foundations.

Against this background, the objective of this work is to design a value alignment process that produces a learning environment for the agent, in which the agent will learn to behave value-aligned while pursuing its individual objective. We consider that a value-aligned agent is one that behaves ethically, following a *moral* value by acting in the most praiseworthy way possible and always respecting moral norms. Furthermore, we also assume in this work that it is possible for the agent to behave ethically as we have defined it. These are the necessary assumptions for all our subsequent contributions.

We address our goal by proposing our view of the value alignment process, which is outlined in Fig. 1. According to such view, a reward specification step combines the individual and ethical objectives to yield a multi-objective environment. Thereafter, an ethical embedding step transforms the multi-objective environment into a single-objective *ethical* environment, which is the one wherein an agent learns. Within the framework of such value alignment process, we address the goal above, focusing on the reward specification and the ethical embedding steps separately. In particular, we address our goal by means of the following main contribution: a novel well-founded approach based on philosophical foundations for automating the whole value alignment process. Our approach tailors current developments in the Multi-Objective Reinforcement Learning literature to build an ethical environment in which the agent learns to behave ethically. Specifically, we construct our approach by means of the following four novel contributions.

1. We provide philosophical foundations that serve as a basis for formalising the notion of moral value and subsequently the notion of ethical behaviour, which together allow us to characterise the concept of ethical objective of Fig. 1.
2. Based on such formalisations, we also characterise the particular ethical behaviours we want an agent to learn: those that prioritise ethical objectives over individual objectives.
3. We offer a solution to the reward specification problem that takes as an input the ethical and individual objectives of the agent, as shown in Fig. 1, and creates a so-called *ethical* reward function such that any agent trying to maximise it will be value-aligned.
4. We present a solution to the ethical embedding problem that, making use of our reward specification, creates a so-called *ethical* environment (shown as the output of Fig. 1), in which an agent learns to behave ethically while pursuing its individual objective.

In what follows, 'Dealing with the value alignment problem' introduces the value alignment problem as a two-step problem. Thereafter, 'Case study: the public civility problem' presents our running example of value alignment problem: the Public Civility Game. Then, 'The reward specification problem' presents our formalisation of the first step: the reward specification problem, and our solution to it. Subsequently, 'The ethical embedding problem' presents our formalisation of the second one: the ethical embedding problem, and our solution to it. Next, 'An algorithm for designing ethical environments' introduces our algorithm to implement our solution to the value alignment problem. Subsequently, 'Related work', summarises the related work in the value alignment literature. Finally, 'Conclusions and future work' concludes and sets paths to future work.

Dealing with the value alignment problem

We devote this section to explaining what the value alignment problem is and to outlining our approach for tackling it.

Problem description

The *value alignment* problem is defined as the problem of ensuring that artificial intelligent agents are aligned with human values (Soares & Fallenstein, 2014; Russell et al., 2015). Thus, a value-aligned agent should pursue goals and objectives that are beneficial to humans, as stated by Soares, Fallenstein, Russell, Arnold, and Sutrop, among others (Soares & Fallenstein, 2014; Russell et al., 2015; Arnold et al., 2017; Sutrop, 2020).

There is an ongoing debate in the literature about what the exact meaning of a human *value* is when referring to the value alignment problem. We follow the philosophical stance of Arnold et al. (2017), Gabriel (2020), Sutrop (2020), and consider that values are: *natural or non-natural facts about what is good or bad, and about what kinds of things ought to be promoted, from an ethical point of view*. Hence, moral values state, for instance, that inequity is bad, and that civility and beneficence are good. In other words, we consider that values are more than simple preferences over actions, and that the objective of value alignment is to guarantee that agents behave ethically. For that reason, henceforward and by abuse of language, we will be using the terms *ethical* and *value-aligned* interchangeably.

The value alignment problem, as an ethical-technical problem, can be subdivided in two challenges, as observed by Gabriel (2020). The first one, the *ethical* one, is the challenge of deciding what moral theory (or a mixture of them) we ought to encode in artificial agents. The second one, the *technical* one, is then how to actually encode the chosen moral theory into the agents in a way that guarantees ethical behaviour. In this paper we will focus on the technical challenge.

Outline of our Reinforcement-Learning approach

In order to tackle the technical challenge of value alignment, there has recently been a growing interest in the use of Reinforcement Learning. In reinforcement learning, an agent learns to behave by a trial-and-error-fashion: it can freely act upon its environment, but each action will have a corresponding reward or punishment (Littman, 2015). The agent learns to behave through a sequence of actions that maximises its obtainment of rewards. These rewards and punishments are defined by specifying what is called a *reward function* (R) (Kaelbling et al., 1996; Sutton & Barto, 1998).

Hence, the technical challenge of value alignment is dealt with by the RL framework as a two-step process: the ethical knowledge is first encoded into a reward function (*reward specification*); and then, this reward function is incorporated into the agent's learning environment (*ethical embedding*). If both processes are performed correctly, the agent then will behave ethically, that is, value-aligned.

Behaviours are typically formalised as *policies* in Reinforcement Learning (Kaelbling et al., 1996). A policy dictates what action to perform in each possible state of the environment. In Reinforcement Learning, agents' rationality is tightly bounded to maximise the accumulated reward, and the policy that maximises the accumulation of rewards is called the *optimal policy* (Kaelbling et al., 1996). Hence, the reward function can be interpreted as expressing the agent's *objective* (Sutton & Barto, 1998; Roijers & Whiteson, 2017).

In reinforcement learning, it is also possible to consider several objectives within the same environment. In such case, we model the environment as a *Multi-Objective Markov Decision Process* (MOMDP) (Roijers & Whiteson, 2017). Multiple (n) objectives are characterised trough n separate reward functions $R_1, \dots, R_n$.

In this paper we will show that Multi-Objective MDPs constitute a useful tool for guaranteeing that agents learn to behave value-aligned. Specifically, we will consider environments in which the agent receives two sources of reward:

1. An individual reward R_0 that only considers the agent's performance according to its original design objective (that is, without ethical considerations).
2. An ethical reward R_v that considers how ethical are the agent's actions. This is the reward that needs to be specified in order to guarantee value alignment.

Figure 1 depicts the overall value alignment process. Firstly, the reward specification process on the left takes, as input, both the individual and ethical objectives. The ethical objective encapsulates the ethical knowledge needed to produce the corresponding reward ethical function R_v. Similarly, the R_0 is naturally derived from the individual objective. Both reward functions R_v and R_0 are then embedded into a resulting Multi-Objective MDP.

Secondly, the ethical embedding process on the right of Fig. 1 will transform this MOMDP into a single-objective MDP by combining these two reward functions into a single one. We will do this process in such a way that ensures that an agent will learn to behave ethically while pursuing its individual objective. Reducing a multi-objective MDP into a single-objective MDP eases the agent's learning because it allows it to use a handful of single-objective RL algorithms such as Q-learning (Watkins & Dayan, 1992). Thus, we refer to this resulting MDP as *ethical environment*, and consider

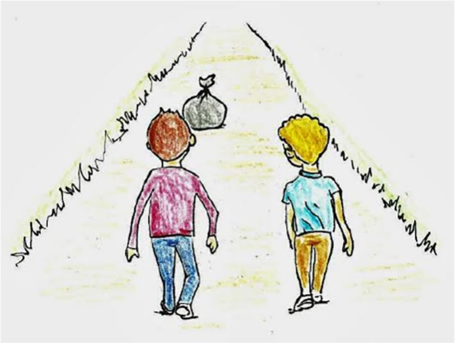


Fig. 2 Possible initial state of a public civility game. The agent on the left must deal with a garbage obstacle ahead

it to be the solution to the value aligned problem as stated above.

Our proposed value alignment process is a refinement from the approach presented by Rodriguez-Soto *et al.* in Rodriguez-Soto et al. (2020), because it allows us to capture the specification into an MOMDP as we have mentioned, instead of directly into an single-objective MDP (as it was done in Rodriguez-Soto et al. (2020)). While their approach was meant for value-alignment in multi-agent system, here we make use of their reward specification for our single-agent value-alignment process. We also provide philosophical foundations and theoretical guarantees for our reward specification process. Furthermore, we also provide an ethical embedding process with algorithmic tools to implement it, unlike in Rodriguez-Soto et al. (2020) in which there was no ethical embedding process nor any novel algorithm presented.

The subsequent sections are devoted to detail how we undertake these two processes (i.e., the reward specification and the ethical embedding). However, we first introduce the running example that we will use along the paper.

Case study: the public civility problem

To illustrate the concepts that will be introduced along this paper we use a single-agent version the *public civility game*. Initially introduced in Rodriguez-Soto et al. (2020) to explore moral dilemmas, we adapt it here to induce ethical behaviour. In short, the game represents a situation wherein two agents move daily from their initial positions (which can be their homes) to their respective target destinations (their workplaces, for instance). Along their journey, the agent on the left finds garbage on the floor that prevents it from progressing. Figure 2 represents this game scenario where the left agent can deal with the garbage in different ways:

- By throwing the garbage aside to unblock his way. However, if the agent throws the garbage at the location where the right agent is, it will hurt the other agent.
- By taking the garbage to the bin. This option is safe for all agents. However, it will delay the agent performing the action.

As for the agent on the right, it is endowed with a fixed behaviour for reaching its goal. Specifically, the right agent moves forward most of the time, just at the beginning it has a 50% chance of being still, to induce some randomness in the scenario.

In this scenario we aim at inducing the moral value of *civility* so that the left agent learns to pick the garbage and to bring it to a bin without throwing it to other agent. In the following sections we will refer back to the public civility game to illustrate how we can induce the agents to learn to behave aligned with the civility value.

The reward specification problem

In this section we focus on the formalisation of the notion of moral value and how it can be translated to rewards in a Reinforcement Learning scenario. First, in 'Philosophical foundations' we dive into the philosophy literature to identify the fundamental components of a moral value. Based on such findings, in Moral value specification' we propose a novel formalisation of the notion of moral value as our approach to tackle the aforementioned ethical challenge of the value alignment problem. Then, we proceed to tackle the technical challenge of the value alignment problem, and in 'From values to rewards' we detail how to derive rewards from this definition. Finally, 'Formal discussion on the soundness of the proposed solution' is devoted to prove that our specification of rewards is sound, that is, they indeed translate our moral value formalisation.

Philosophical foundations

Ethics or moral philosophy is the branch of philosophy that studies goodness and right action (Audi, 1999; Cooper, 1993; Fieser & Dowden, 2000; Frankena, 1973). Citing (Audi, 1999): *Correlatively, its principal substantive questions are what ends we ought, as fully rational human beings, to choose and pursue.* Thus, right action becomes closely related to the the core concept of moral value, which expresses the moral objectives *worth striving for* (van de Poel & Royakkers, 2011).

Prescribing how people ought to act is the subject of study of prescriptive ethics. *Prescriptive* ethics (also known as *normative* ethics), constitutes one of the main areas of research in ethics. Three of the most well-known types of prescriptive ethical theories are: virtue ethics, consequentialist ethics, and duty ethics.

- Virtue ethics (developed by Socrates, Plato and Aristotle among other ancient Greek philosophers) states that by honing virtuous[1] habits –such as being honest, just, or generous– people will likely make the right choice when faced with ethical challenges (van de Poel & Royakkers, 2011).
- Consequentialist ethics holds that actions must be morally judged depending on their consequences. For example, in utilitarianism (developed by Jeremy Bentham and John Stuart Mill in its classical form), actions are judged in function of how much *pleasure* (utility) or pain they cause. To act ethically is to act in a way that maximises the amount of goodness for the largest number of people (van de Poel & Royakkers, 2011).
- Duty ethics (or deontology, from the Greek deon, which means duty) states that an action is good if it is in agreement with a moral duty[2] that is applicable in itself, regardless of its consequences (van de Poel & Royakkers, 2011). Examples of duty ethics include Immanuel Kant's theory or the Divine Commands theory, (in which for instance we find the moral norm of "thou shalt not kill", under any circumstance).

It is important to remark that all these ethical theories are not opposing theories we need to choose from. They are all complementary and must be all taken into account (Camps, 2013). For that reason, in this paper we aim at a formal definition of moral value that can be compatible with any of these ethical theories.

What all these prescriptive ethical theories share in common is that they were developed in historical contexts in which all actions were assumed to fall in either one of the following three categories (Heyd, 2016):

1. Actions morally obliged because they are good to do.
2. Actions morally prohibited because they are bad to do.
3. Actions permitted because they are neither good nor bad to do.

That is, these theories translated *evaluative* notions (an action is either good, bad, or neutral) into *normative* notions (an action is either obliged, prohibited or permitted). However, in the last century, an ethical discussion has developed around the existence of a fourth category (Chisholm, 1963; Urmson, 1958):

4. Actions that are good to do, but *not* morally obligatory.

These are actions that go *beyond the call of duty* (Urmson, 1958), such as beneficence or charity, are termed *supererogatory* actions.

This fourth category implies that the normative dimension alone is not enough to categorise actions morally. Thus, in order to fully judge an action morally, it is required to look at it from these two *dimensions*, as argued by Chisholm (1963), Frankena (1973), Etzioni and Etzioni (2016): (1) a *deontic or normative* dimension, considering whether it should be morally obliged, permitted, or prohibitted; and (2) an *axiological or evaluative* dimension, that considers how praiseworthy or blameworthy it is.

Therefore, as argued by Heyd (2016), the deontic dimension deals with the minimal conditions for morality, while the axiological dimension aims at higher (ethical) ideals which can only be commended and recommended but not strictly required.

In conclusion, we consider moral values as principles for discerning between right and wrong actions, and, moreover, we argue that they must be endowed with a normative and an evaluative dimension. Any action will thus need to be considered from these two ethical dimensions, in order to fully consider the four action categories identified above.

Moral value specification

As we just mentioned, we formalise moral values with two dimensions: a normative one and an evaluative one.

In the normative dimension, we formalise the moral norms that promote "good" actions and forbid "bad" actions (for example: "it is morally prohibited to kill others"[3]). These moral norms constitute the minimum that an agent should align with in order to co-inhabit with humans, as explained in Amodei et al. (2016), Leike et al. (2017).

Conversely, in the evaluative dimension we formalise how good or bad each action is. These two dimensions may not always apply to the same set of possible actions, since some actions may be evaluated as good without being obligatory

[1] The concepts of virtues and values may seem very similar at first. Indeed, many virtues such as honesty and generosity are also moral values. The difference strikes in that a virtue refers to the character traits of an agent that is truly realising this moral value (van de Poel & Royakkers, 2011).

[2] Some theories consider that there is a unique supreme duty that needs to be followed, such as Kantian's categorial imperative. Other theories argue that ther are several duties, for instance in Ross's ethics, in which we have the duties of beneficence, gratitude and justice among others (van de Poel and Royakkers, 2011).

[3] Notice that although moral norms are the basis for legal norms (Audi, 1999; Cooper, 1993), they encompass a larger set of norms than what is legally obliged or prohibited. We use legal norms as examples because they are widely known, and hence easy to understand.

(and this is specially the case for supererogatory actions)[4]. In this paper we consider that an agent that performs those actions as value-aligned, following the same direction that Gabriel and Sutrop (Gabriel, 2020; Sutrop, 2020).

Notice that, since we will ethically evaluate actions, it is important to also consider the context where they are performed when doing so. For instance, consider the action of performing an abortion to a woman that has already agreed to abort. The context where it takes place dictates how blameworthy or praiseworthy it is: performing it in many Western European countries is not seen as blameworthy, whereas in many other countries it is seen even as very blameworthy and even morally (and legally) prohibited. In the next subsection we will see that this connection between contexts and actions is especially relevant in Reinforcement Learning, for which contexts receive the name of states.

In summary, in addition to the normative dimension –by which each value is defined in terms of the norms that promote good actions with respect to the value–, we will also include in our moral value definition an action evaluation function that enriches our ethical system with an evaluative perspective.

Therefore, we next introduce our formal definition of value, which includes these two dimensions as two value components (i.e., norms promoting the value and an action evaluation function). We adopt our definition of moral value from Rodriguez-Soto et al. (2020).

Definition 1 *(Moral value)* Given a set of actions $\mathcal{A}$, we define a moral value v as a tuple $\langle \mathcal{N}_v, E_v \rangle$ such that:

- $\mathcal{N}_v$ is a finite set of norms promoting good actions with respect to the value. We succinctly represent norms as $n = \theta(a)$, where $\theta \in \{Prh, Per, Obl\}$ is a deontic operator with the semantics of Prohibiting, Permitting or Obliging the performance of action a respectively.
- $E_v : \mathcal{A} \rightarrow [-1, 1]$ is an action evaluation function that measures the degree of value promotion/demotion of an action $a \in \mathcal{A}$. Specifically, $E_v(a) = 1$ means that the performance of the action a strongly promotes the moral value; whereas $E_v(a) = -1$ stands for strong demotion.

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

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

Observe that a moral value contains those norms that promote it, but our definition goes beyond norms, since the action evaluation function encapsulates knowledge about actions morally good but not obligatory. Moreover, it is worth noticing that we assume the moral value is defined so that it does not contain mutually exclusive (contradictory) norms. If that was the case, it would mean that the moral value encompasses genuine (unsolvable) moral dilemmas (for more information on moral dilemmas, see for instance (Conee, 1982; Zimmerman, 1987)). Moreover, paraphrasing Russell in Russell (2019), if for a given situation there is a true moral dilemma, then there are good arguments for all the possible solutions to it, and therefore artificial agents cannot cause more harm than humans even if they take a wrong decision. Hence, here we adhere to Russell's reasoning and disregard moral dilemmas.

Example 1 Considering the scenario of the public civility game introduced in 'Case study: the public civility problem', we focus on two actions: *bin*, which corresponds to the action of throwing the garbage to a bin when having run into it (i.e., if the agent had previously found the garbage in front); and *hit*, which represents throwing garbage nearby and hitting the other agent when having run into it.

Then, we can define a norm $n \in N$ prohibiting to perform action *hit* ($n = Prh(hit)$). Since this norm is aligned with the *civility* moral value, we include it in the definition of such value together with an action evaluation function E_v. In this manner, $civility = \langle \{n\}, E_v \rangle$ where $E_v(bin) = 1$ since, in terms of civility, the action of bringing garbage to a bin is highly praiseworthy to perform; and, finally, $E_v(hit) = -1$ since it is very blameworthy to perform (and even prohibited by the norm n).

Notice that what is morally prohibited according to the moral value of *civility* is to hit another agent with a piece of garbage, hence hurting it. Nevertheless, it is still permitted for the agent to throw the garbage aside if no other agent is harmed.

Since one of our objectives was the characterisation of ethical behaviour, we can now do so from the definition of moral value v. We expect an ethical agent to abide by all the norms of v while also behaving as praiseworthily as possible[5] according to v. Formally:

[4] One may argue those actions are indeed permitted, but we prefer not to abuse the semantics of permissions.

[5] It might be worth noticing that although our definition of ethical behaviour seems too restrictive, we encourage the reader to interpret it as a necessary requirement for providing the theoretical guarantees that the value alignment problem needs. Notice that our requirement is keen to the ones in other areas such as game theory, in which it is assumed that any rational agent tries to always maximise its utility function, and this assumption serves as the basis of its most important theoretical results.

Definition 2 (*Ethical behaviour*) Given a moral value v, an agent's behaviour (the sequence of actions that it will perform) is ethical with respect to v if and only if: (1) it complies with all the norms in $\mathcal{N}_v$; and also (2) it acts in the most praiseworthy way according to E_v.

Example 2 In the context of the public civility game, the only ethical behaviour is to bring the garbage to the bin (which implies to never throw it to the other agent).

From values to rewards

We now proceed to explain our approach for the first step of the value alignment process: the reward specification. Specifically, we detail how to adapt our formal definition of a moral value into a reward function of a Reinforcement Learning environment. Our approach consists on presenting the individual and the ethical objectives of the agent as two separate reward functions of a Multi-Objective MDP, as Fig. 1 illustrates.

As previously mentioned in 'Dealing with the value alignment problem', we formalise the agent learning environment as a Markov Decision Process (MDP) $\mathcal{M}$, which can have one or multiple objectives (MOMDP). States of such environment $\mathcal{M}$ are defined as a set S. Moreover, for each state $s \in \mathcal{S}$, we consider $\mathcal{A}(s)$ to be the set of actions that the agent can perform in s. Then, the performance of a specific action a in a state s is rewarded according to each objective in $\mathcal{M}$. We notate this by means of the reward function $R_i(s, a)$, which returns a real number –either positive or negative– with respect to the i-th objective in $\mathcal{M}$.

This way, we associate how praiseworthy or blameworthy an action is with a reward from a so-called *ethical* reward function. Therefore, we can formalise the ethical reward specification problem as that of computing a reward function R_v that, if the agent learns to maximise it, the learnt behaviour is aligned with the moral value v. Formally:

Problem 1 (*Ethical reward specification*) Given a moral value v, and an MDP $\mathcal{M}$ with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$, compute an ethical reward function R_v such that an optimal policy for $\mathcal{M}$ with respect to R_v is value-aligned with respect to v.

We solve this problem by mapping the two components of a moral value ($\mathcal{N}_v$ and E_v) into two different reward components ($R_\mathcal{N}$ and R_E, respectively) that we combine to obtain the ethical reward function $R_v = R_\mathcal{N} + R_E$.

On the one hand, we create the normative component $R_\mathcal{N}$ through two main steps: firstly, we identify which action-state pairs do represent violations of the norms in $\mathcal{N}_v$, and define the corresponding penalties; and, secondly, we aggregate all these penalties into the normative reward function.

Thus, we first formalise the *Penalty* function for a norm n as the function P_n that returns -1 whenever performing action a in state s represents a violation of the norm. Therefore, in fact, non-compliance stems from either performing a forbidden action or from failing to perform an obliged action. Our definition of the Penalty function is based on the one present in Rodriguez-Soto et al. (2020), adapted here for contextualised actions.

Definition 3 (*Penalty function*) Given a norm $n = \theta(k)$, and an MDP with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$, we define the penalty function $P_n : \mathcal{S} \times \mathcal{A} \rightarrow \{-1, 0\}$ as

$$P_n(s,a) \doteq \begin{cases} -1 & \text{if } a = k, \theta = Prh \text{ and } k \in \mathcal{A}(s), \\ & \text{or if } a \neq k, \theta = Obl \text{ and } k \in \mathcal{A}(s), \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where s is a state of $\mathcal{S}$ and k, a are actions of $\mathcal{A}(s)$.

Second, we consider all norms in $\mathcal{N}_v$ and aggregate their penalties into a normative reward function $R_\mathcal{N}$ that adds these penalties for each state-action pair. Formally:

Definition 4 (*Normative reward function*) Given a set of norms $\mathcal{N}$ and an MDP, we define the reward function of a set of norms $\mathcal{N}$ as a reward function $R_\mathcal{N} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^-$, defined as

$$R_\mathcal{N}(s,a) \doteq \sum_{n \in \mathcal{N}} P_n(s,a). \tag{2}$$

The reward function $R_\mathcal{N}$ aggregates the punishments from all those norms that are violated (see Eq. 1) in a given state-action pair $\langle s, a \rangle$.

The Normative reward function here present is a direct adaptation for MDPs of the one present in Rodriguez-Soto et al. (2020), which was designed for Markov games.

On the other hand, we translate the action evaluation function E_v in the moral value (see Definition 1) into the evaluative component R_E in R_v by (positively) rewarding praiseworthy actions. Formally:

Definition 5 (*Evaluative reward function*) Given an action evaluation function E_v of a moral value v, and an MDP, we define the reward function of E_v as a reward function $R_E : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^+$, defined as

$$R_E(s,a) = \begin{cases} \max(0, E_v(a)) & \text{if } a \in \mathcal{A}(s), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The reward function R_E rewards praiseworthy actions performed under certain contexts (i.e., those states in the MDP where the action can be done).

The Evaluative reward function here present is an adaptation for MDPs of the one present in Rodriguez-Soto et al. (2020), which was designed for Markov games.

Notice that our evaluative reward function definition implies that E_v need not be defined for all the actions of an MDP. The environment designer just needs to define it for those that they explicitly consider praiseworthy to perform. Thus, from a pragmatic perspective, the environment designer must only focus on specifying R_E for a limited subset of state-action pairs out of all the possible ones in the MDP.

Moreover, it is worth mentioning that we set a reward of 0 to any action that is not praiseworthy to perform –including those that are blameworthy but still permitted– not to further restrict the choices of the learning agent.

We are now capable of formally defining the ethical reward function R_v in terms of previous definitions of $R_{\mathcal{N}}$ and R_E. Following the Ethics literature (Chisholm, 1963; Etzioni & Etzioni, 2016; Frankena, 1973; van de Poel & Royakkers, 2011), we consider $R_{\mathcal{N}}$ and R_E of equal importance, and, therefore, we simply define R_v as an addition of the normative reward function $R_{\mathcal{N}}$ and the evaluative reward function R_E. Formally:

Definition 6 (*Ethical reward function*) Given a moral value $v = \langle \mathcal{N}_v, E_v \rangle$ and an MDP, we define the ethical reward function of v as a reward function $R_v : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defined as:

$$R_v(s,a) = R_{\mathcal{N}}(s,a) + R_E(s,a), \tag{4}$$

where $R_{\mathcal{N}}$ is the reward function of $\mathcal{N}_v$, and R_E is the reward function of E_v.

Finally, recall, from Fig. 1, that the output of the Reward Specification process we are describing here corresponds to a Multi-Objective MDP. This MOMDP extends the individual objective –represented trough the R_0 reward function– with an ethical objective by adding the value-aligned reward function R_v. Formally:

Definition 7 (*Ethical extension of a Markov decision process*) Given a moral value v and an MDP with a reward function R_0, we define its *ethical extension* as a Multi-Objective MDP with a vectorial reward function $\mathbf{R} = (R_0, R_v)$, where R_v is the ethical reward function of v.

For simplicity, when there is no confusion, we refer to the ethical extension of an MDP simply as an *ethical MOMDP*.

Our definition of an Ethical extension of an MDP is a refined translation for Multi-Objective MDPs of an Ethical extension of a (single-objective) Markov game, as defined in Rodriguez-Soto et al. (2020). This modular framing of the objectives allows us to utilise multi-objective algorithms to later obtain the desired ethical environment, as we will see in the following section.

Example 3 Continuing with previous Example 1 about the moral value of *civility* $= \langle \{n\}, E_v \rangle$, we can formalise the public civility game as an ethical MOMDP. In this MOMDP, states represent the positions of the agents and the garbage, and the individual objective for the learning agent is to reach its destination as fast as possible. Thus, the individual reward function R_0 returns a positive reward of 20 to the agent whenever located at its goal. Otherwise, it returns a negative reward of −1. Furthermore, we consider the ethical reward function $R_v = R_{\mathcal{N}} + R_E$, and we proceed to first define the normative component $R_{\mathcal{N}}$ based on norm $n = Prh(hit)$:

$$R_{\mathcal{N}}(s,a) = \begin{cases} -1 & \text{if } a = hit \text{ and } hit \in \mathcal{A}(s), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This normative component $R_{\mathcal{N}}$ of the ethical reward function punishes the agent for not complying with the moral requirement of being respectful with other agents. Thus, the agent on the left will be punished with a negative reward of −1 if it throws the garbage to the agent on the right.

Secondly, we define R_E from E_v as:

$$R_E(s,a) = \begin{cases} E_v(bin) & \text{if } a = bin, \text{and } a \in \mathcal{A}(s), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Thus, our evaluative component R_E of the ethical reward function rewards the agent positively (with a reward of 1) when performing the praiseworthy action of pushing the garbage inside the wastebasket.

Formal discussion on the soundness of the proposed solution

This subsection is devoted to prove that the ethical reward function previously introduced actually solves Problem 1. In other words, we aim at showing that R_v guarantees that an agent trying to maximise it will learn a value-aligned behaviour according to Definition 2.

In order to do so, let us first recall, from 'Dealing with the value alignment problem', that agent behaviours are formalised as policies in the context of MDPs. Thus, we refer to the ethical behaviour from Definition 2 as an ethical policy. Consequently, we consider a policy to be ethical if it complies with all the norms of a moral value, and if it is also praiseworthy in *the long term*. In Reinforcement Learning, this notion of the long term is formalised with the *state-value function* V^{π}, that for any policy π it returns how many rewards will the agent obtain in total. In an MOMDP, there is a state-value function V_i for each objective i.

Thus, we can formalise an ethical policy as a policy that: (1) never accumulates normative punishments; and (2) maximises the accumulation of evaluative rewards. Formally:

Definition 8 (*Ethical policy*) Let $\mathcal{M}$ be an ethical MOMDP. We say that a policy π_* is an ethical policy in $\mathcal{M}$ if and only if it is optimal for both its normative $V_{\mathcal{N}}$ and evaluative V_E components:

$$V_{\mathcal{N}}^{\pi_*} = 0,$$
$$V_E^{\pi_*} = \max_{\pi} V_E^{\pi}.$$

Our definition of ethical policy in an ethical MDP is an adaptation of the definition of ethically-aligned policy in an ethical Markov game from Rodriguez-Soto et al. (2020). Notice however that unlike in Rodriguez-Soto et al. (2020), our definition is a translation of the definition of ethical behaviour (Def. 2) to MDPs.

For all the following theoretical results, we assume the following condition for any ethical MOMDP: if we want the agent to behave ethically, it must be actually possible for it to behave ethically[6]. Formally:

Condition 1 (Ethical policy existence) Given an ethical MOMDP, there is at least one ethical policy (as formalised by Def. 8).

With Condition 1 we are capable of finally proving that our translation of moral values to reward functions solves Problem 1:

Theorem 1 (*Specification soundness*) *Given a moral value v and an ethical MOMDP $\mathcal{M}$ with an ethical reward function R_v in which Condition 1 is satisfied, all optimal policies of $\mathcal{M}$ with respect to R_v are ethical policies with respect to v.*

Proof This theorem relies on the fact that any policy that is optimal with respect to an ethical reward function R_v given a moral value $v = \langle \mathcal{N}_v, E_v \rangle$ will maximise the accumulation of $V_N + V_E$. Then, Condition 1 also implies that $V_N + V_E$ will be maximised if and only if both V_N and V_E are maximised. Therefore, such optimal policy will be an ethical policy. □

The ethical embedding problem

Reward specification is followed, within the overall value-alignment process, by the ethical embedding process. As depicted in Fig. 1, this ethical embedding process takes as input the MOMDP –which contains reward functions R_0 and R_v– and produces an ethical (single-objective) MDP by linearly combining these reward functions. Next Formalising the ethical embedding problem' specifies our formalisation of the so-called ethical embedding problem. Subsequently, 'Solving the ethical embedding problem' details our proposal to solve this problem.

Formalising the ethical embedding problem

As previously mentioned, our main goal is to guarantee that an agent will learn to behave ethically, that is, to behave in alignment with a moral value whilst pursuing its individual objective. With that aim, we combine the reward functions that represent these two objectives in the ethical MOMDP by means of a so-called *embedding function* to obtain an ethical (single-objective) MDP where the agent will learn its policy.

Although the previous section introduced ethical policies, in fact, we are interested in the so-called *ethical-optimal* policies. These policies pursue the individual objective subject to the ethical objective being fulfilled. Specifically, we say that a policy is *ethical-optimal* if and only if it is ethical (following Def. 8), and it maximises the individual objective V_0 (i.e., the accumulation of rewards R_0) among ethical policies. Formally:

Definition 9 (*Ethical-optimal policy*) Given an ethical MOMDP $\mathcal{M}$, a policy π_* is *ethical-optimal* in $\mathcal{M}$ if and only if it is maximal among the set Π_e of ethical policies:

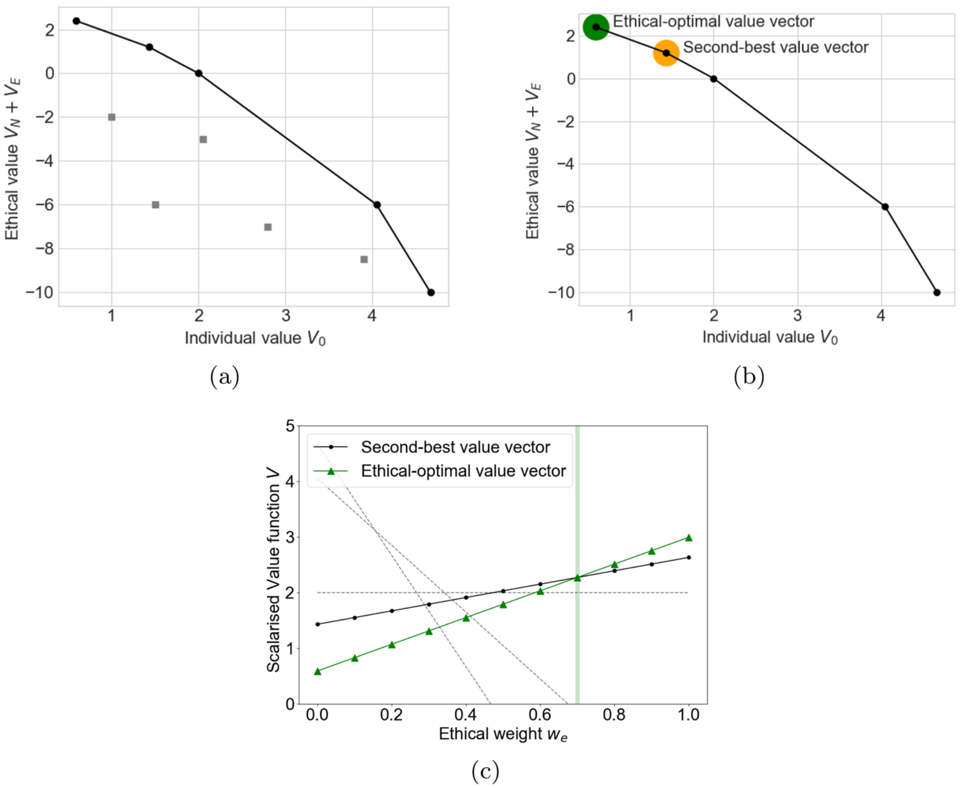
$$V_0^{\pi_*} = \max_{\pi \in \Pi_e} V_0^{\pi}.$$

Due to the mathematical properties of MOMDPs, while there can be several ethical-optimal policies in an ethical MOMDP, all of them will share the same *value vector* (the vector of all the state-value functions of the agent). We refer to such value vector as the *ethical-optimal* value vector $\mathbf{V}^*$.

Example 4 In the context of the public civility game, an ethical-optimal policy is a policy that brings the garbage to the bin (the ethical behaviour, as explained in Example 2) while getting to its goal as fast as possible (its individual objective).

In the literature on MOMDPs, any function that combines all the objectives of the agent into a single one receives the name of a *scalarisation function* (Roijers & Whiteson, 2017). We refer to this scalarisarion function as the

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



(a)

(b)

(c)

Fig. 3 **a** Example of convex hull $CH(\mathcal{M})$, represented in objective space. **b** Identification of the points of $CH(\mathcal{M})$ corresponding with the ethical-optimal value vector $\mathbf{V}^*$ (highlighted in green) and the second-best value vector $\mathbf{V'}^*$ (in yellow). **c** Representation in weight space of $CH(\mathcal{M})$. The minimal weight value w_e for which $\mathbf{V}^*$ is optimal is identified with a green vertical line. (Color figure online)

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

$$f(\mathbf{V}^\pi) = \mathbf{w} \cdot \mathbf{V}^\pi = w_0 V_0^\pi + w_e(V_\mathcal{N}^\pi + V_E^\pi) \tag{7}$$

where $\mathbf{w} = (w_0, w_e)$ is a weight vector with weights $w_0, w_e > 0$ to guarantee that the agent is taking into account all rewards (i.e., both objectives). We will be referring thus to w_0 as the *individual* weight and w_e as the *ethical weight*. Without loss of generality, hereafter we fix the individual weight to $w_0 = 1$.

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

Problem 2 (*Ethical embedding*) Let $\mathcal{M}$ be an ethical MOMDP with reward functions $(R_0, R_N + R_E)$. The ethical embedding problem amounts to computing the weight vector $\mathbf{w}$ with positive weights such that all optimal policies in the MDP $\mathcal{M}'$ with reward function $R_0 + w_e(R_\mathcal{N} + R_E)$ are also ethical-optimal in $\mathcal{M}$ (following Def. 9).

A weight vector $\mathbf{w}$ with positive weights guaranteeing that all optimal policies are also ethical-optimal is a solution of Problem 2. Moreover, we aim at finding solutions of the form $\mathbf{w} = (1, w_e)$ that design a so-called *ethical* environment as similar as possible to the original one, in which the agent only cared for its individual objective. Therefore, we aim at knowing the *minimal* ethical weight w_e for which $(1, w_e)$ is a solution of Problem 2 (i.e., for which $\mathbf{V}^*$ is the only optimal policy).

Solving the ethical embedding problem

This section explains how to compute a solution weight vector $\mathbf{w}$ for the ethical embedding problem (Problem 2). Such weight vector $\mathbf{w}$ combines individual and ethical rewards into a single reward to create an ethical environment in which the agent learns an ethical behaviour, that is, an ethical-optimal policy.

Figure 3 illustrates our proposed steps for solving this embedding problem. The first step focuses on obtaining the *convex hull* $CH(\mathcal{M})$ (Roijers & Whiteson, 2017) of the ethical MOMDP. The convex hull is one of the main concepts of

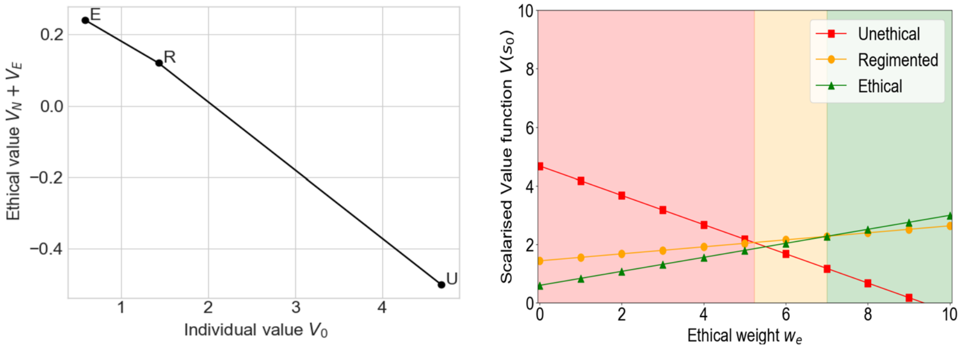


Fig. 4 Left: Visualisation in Objective Space of the convex hull of the public civility game composed by 3 policies: E (Ethical), R (Regimented) and U (Unethical). Right: Visualisation in Weight Space of the same convex hull. The painted areas indicate which policy is optimal for the varying values of the ethical weight w_e: red for the Unethical policy, yellow for the Regimented one, and green for the Ethical one. (Color figure online)

MOMDPs: it contains all the policies (and their associated value vectors) that are optimal for at least one linear scalarisation function $\mathbf{w}$ with positive weights (i.e., $w_i > 0$ for all $w_i \in \mathbf{w}$, as it is actually the case in our embedding function). Figure 3a shows an example of $CH(\mathcal{M})$ where black-rounded points constitute the convex hull while grey points are values of policies never maximal for any weight.

The second step requires the computation of the ethical-optimal value vector $\mathbf{V}^*$. Figure 3b highlights in green $\mathbf{V}^*$, which accumulates the greatest ethical value (Y axis). This ethical-optimal value vector $\mathbf{V}^*$ will serve as a reference value vector to find the minimal weight vector $\mathbf{w} = (1, w_e)$ that solves Problem 2. For such weight vector, $\mathbf{w} \cdot \mathbf{V}^*$ is maximal (and the only maximal one) among all value vectors of $CH(\mathcal{M})$.

Computing the minimal ethical weight does not require to consider all value vectors on the convex hull. In fact, it suffices to consider the so-called *second-best* value vector (highlighted in yellow in Fig. 3b) to compute it. The second-best value vector accumulates the greatest amount of ethical value after the ethical-optimal one.

Figure 3c plots how the scalarised values of the points in the convex hull $CH(\mathcal{M})$ (Fig. 3a) change as the ethical weight increases. This figure illustrates how immediately after the line representing the ethical-optimal value vector $\mathbf{V}^*$ intersects the second-best value vector, $\mathbf{V}^*$ becomes maximal. Computing such intersection point constitutes the last step to find the solution, as it provides a *tight lower bound* for the value of the ethical weight w_e (see the green vertical line for $w_e = 0.7$ in Fig. 3c).

To summarise, we compute the ethical embedding function $\mathbf{w} = (1, w_e)$ with the minimal ethical weight w_e in three steps:

1. *Computation of the convex hull* (Fig. 3a).
2. *Extraction of the two value vectors with the greatest ethical values* (Fig. 3b).
3. *Computation of the ethical embedding function* $(1, w_e)$ *with minimal* w_e (Fig. 3c).

The remaining of this section is devoted to provide some more details about these three steps.

1. Computation of the convex hull. The convex hull can be readily computed by means of the well-known Convex Hull Value Iteration algorithm (Barrett & Narayanan, 2008). Here, we illustrate the convex hull obtained for our running example:

Example 5 Considering $\mathcal{M}$, the ethical MOMDP of the public civility game, we compute its convex hull $CH(\mathcal{M})$ [7]. Figure 4 depicts the result. It is composed of 3 different policies named after the behaviour they encapsulate: (1) an **U**nethical (uncivil) policy that would make the agent move towards the goal and throw away the garbage without caring about any ethical implication; (2) a **R**egimented policy that would allow the agent to comply with the norm that prohibits throwing the garbage to the other agent; and finally, (3) an **E**thical policy that would make the agent behave civically

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

Policy π	Value $\mathbf{V}^\pi$	w_e ranges
Unethical	(4.67, -0.5 + 0)	[0.0, 5.2]
Regimented	(1.43, 0 + 0.12)	[5.2, 7]
Ethical	(0.59, 0 + 0.24)	[7, ∞)

[7] Recall that the convex hull is formed by those policies that are optimal for some weight vector with positive weights.

as desired. Table 1 provides the specific vectorial value $\mathbf{V}^{\pi} = (V_0^{\pi}, V_{\mathcal{N}}^{\pi} + V_E^{\pi})$ of each policy π.

Recall that we find these three policies (**U**nethical, **R**egimented and **E**thical) in the convex hull because they are the only three policies that are optimal for some weight vector with positive weights.

2. Extraction of the two value vectors with the greatest ethical values (as illustrated in Fig. 3b). Firstly, in order to find the value vector in the convex hull $CH(\mathcal{M})$ that corresponds to an ethical-optimal policy, we look for the one that maximises the ethical reward function $(R_{\mathcal{N}} + R_E)$ of the ethical MOMDP. Formally, to obtain the ethical-optimal value vector within $CH(\mathcal{M})$, we compute:

$$\mathbf{V}^* = \underset{(V_0, V_{\mathcal{N}}+V_E) \in CH}{\arg\max} [V_{\mathcal{N}} + V_E]. \tag{8}$$

Secondly, we compute $\mathbf{V'}^*$, the so-called *second-best* value vector, which accumulates the greatest amount of ethical rewards in $CH(\mathcal{M})$ if we disregard $\mathbf{V}^*$ (i.e., when considering $CH \setminus \{V^*\}$). Formally:

$$\mathbf{V'}^* \doteq \underset{(V_0, V_{\mathcal{N}}+V_E) \in CH \setminus \{V^*\}}{\arg\max} [V_{\mathcal{N}} + V_E]. \tag{9}$$

In fact, we only need to compare $\mathbf{V}^*$ and $\mathbf{V'}^*$, and hence disregard the rest of value vectors in the convex hull, in order to find the minimal ethical weight w_e for which $\mathbf{V}^*$ is the only maximal value vector. Thus, these two value vectors $\mathbf{V}^*$ and $\mathbf{V'}^*$ are all we need to compute the embedding function $\mathbf{w} = (1, w_e)$ with minimal ethical weight w_e. Notice that $\mathbf{V}^*$ and $\mathbf{V'}^*$ can be found simultaneously while sorting the value vectors of $CH(\mathcal{M})$. Furthermore, $\mathbf{V}_{\mathcal{N}} + \mathbf{V}_E$ are already available for these two value vectors because they are both part of the previously computed convex hull $CH(\mathcal{M})$.

Example 6 In the case of the public civility game, the **E**thical policy turns out to be the one that has associated the ethical-optimal value vector. The third row in Table 1 indicates so, since it is the policy with greatest ethical value within the convex hull. Specifically, if we denote the ethical policy as π_e, we have $\mathbf{V}^{\pi_e} = (V_0^{\pi_e}, V_{\mathcal{N}}^{\pi_e} + V_E^{\pi_e}) = (0.59, 0 + 0.24)$ and $\mathbf{V}^* = \mathbf{V}^{\pi_e}$ because π_e is the only policy that maximises both the normative and the evaluative components ($V_{\mathcal{N}}$ and V_E respectively).

Similarly, the second most ethical value vector in $CH(\mathcal{M})$ corresponds to the value of the **R**egimented policy π_R, which (as the second row in Table 1 shows) has value $\mathbf{V}^{\pi_R} = (V_0^{\pi_R}, V_{\mathcal{N}}^{\pi_R} + V_E^{\pi_R}) = (1.43, 0 + 0.12)$. Therefore, $\mathbf{V'}^* = \mathbf{V}^{\pi_R}$.

3. Computation of the ethical embedding function $(1, w_e)$ *with minimal* w_e. We use the two previously extracted value vectors $\mathbf{V}^*$ and $\mathbf{V'}^*$ to find the minimal solution weight vector $\mathbf{w} = (1, w_e)$ that guarantees that optimal policies are ethical-optimal. In other words, such weight vector $\mathbf{w}$ will create an ethical environment (a single-objective MDP) in which the agent will learn an ethical-optimal policy. Specifically, we need to find the minimal value for $w_e \in \mathbf{w}$ such that:

$$V_0^* + w_e[V_{\mathcal{N}}^* + V_E^*] > V_0' + w_e[V_{\mathcal{N}}' + V_E'], \tag{10}$$

for every state $s \in \mathcal{S}$, where $\mathbf{V}^* = (V_0^*, V_N^* + V_E^*)$ and $\mathbf{V'}^* = (V_0', V_N' + V_E')$. This process is illustrated in Fig. 3c. Notice that in Eq. 10 the only unknown variable is w_e.

Example 7 Back again to the public civility game, we can compute the weight w_e in $\mathbf{w} = (1, w_e)$ for which π_e is the only optimal policy of CH by solving Eq. 10. This amounts to solve:

$$V_0^{\pi_e} + w_e[V_{\mathcal{N}}^{\pi_e} + V_E^{\pi_e}] > V_0^{\pi_R} + w_e[V_{\mathcal{N}}^{\pi_R} + V_E^{\pi_R}]. \tag{11}$$

By solving it, we find that if $w_e > 7$, then the **E**thical policy becomes the only optimal one. We can check it (set $\epsilon > 0$):

$$\begin{aligned} &0.59 + (7 + \epsilon) \cdot (0 + 0.24) \\ &\quad = 2.27 + 0.24\epsilon > 1.43 + 7 \cdot (0 + 0.12) = 2.27. \end{aligned}$$

Figure 4 (right) illustrates the scalarised value of the three policies for varying values of w_e in [0,10] (for $w_e > 10$ the Ethical policy remains optimal). The painted areas in the plot help to identify the optimal policies for specific intervals of w_e. Focusing on the green area, we observe that the **E**thical policy becomes the only optimal one for $w_e > 7$.

An algorithm for designing ethical environments

At this point, we now count on all the tools for solving the value alignment problem (formulated as Problems 1 and 2), and hence build an ethical environment where the learning of ethical policies is guaranteed.

The ethical environment design algorithm

Algorithm 1 implements the reward specification and ethical embedding processes outlined in Fig. 1. The algorithm receives as input an MDP $\mathcal{M}_0$ with an individual reward function R_0, and a moral value v. It starts in line 2 by computing the associated ethical MOMDP that contains both the individual and the ethical objectives of the agent. This step corresponds to the whole reward specification process detailed in 'The reward specification problem'.

Then, the rest of lines (from 3 to 6) deal with the ethical embedding process detailed in 'The ethical embedding problem'. In line 3, the algorithm computes the convex hull $CH(\mathcal{M})$ of the ethical MOMDP $\mathcal{M}$. Next, line 4 obtains the ethical-optimal value vector $\mathbf{V}^*$ and the second-best value vector $\mathbf{V'}^*$ out of those in $CH(\mathcal{M})$. Thereafter, line 5 applies $\mathbf{V}^*$ and $\mathbf{V'}^*$ in Equation 10 to compute the minimal ethical weight w_e. The algorithm then builds a single-objective *ethical* MDP $\mathcal{M}'$ with reward function $R_0 + w_e(R_\mathcal{N} + R_E)$ where all optimal policies in $\mathcal{M}'$ are ethical. Thus, since $\mathcal{M}'$ solves the ethical embedding problem (Problem 2),—and hence, the whole value alignment problem—the algorithm returns $\mathcal{M}'$ in line 6.

Algorithm 1 Ethical Environment Design

```
1: function ( MDP $\mathcal{M}_0$ with reward function $R_0$, moral value $v = \langle \mathcal{N}_v, E_v \rangle$ )
2:     Compute an ethical MOMDP $\mathcal{M}$ with reward functions $(R_0, R_v)$, where $R_v = R_N + R_E$ is the ethical reward function associated with $v$.
3:     Compute $CH(\mathcal{M})$, the convex hull of $\mathcal{M}$
4:     Find $\mathbf{V}^*$ the ethical-optimal value vector, and $\mathbf{V'}^*$ the second-best value vector, within $CH(\mathcal{M})$ by solving Eq.'s 8 and 9.
5:     Find the minimal value for $w_e$ that satisfies Eq. 10.
6:     Return the *ethical* MDP $\mathcal{M}'$ with reward function $R_0 + w_e(R_\mathcal{N} + R_E)$.
7: end function
```

We finish this subsection by proving that Algorithm 1 is complete, that is, for any finite MDP $\mathcal{M}$ and any moral value v, it returns another MDP $\mathcal{M}'$ in which it is guaranteed that optimal policies are value-aligned with v. Formally:

Theorem 2 (*Algorithm completeness*) *Let a moral value v (as formalised in Def.* 1*), and a finite MDP $\mathcal{M}$ in which condition* 1 *is satisfied, be the inputs of Algorithm* 1*. Then, Algorithm* 1 *returns an MDP $\mathcal{M}'$ in which all optimal policies are ethical-optimal with respect to v.*

Proof If there exists an ethical weight w_e for which all optimal policies are ethical-optimal, lines 4, 5 and 6 of our algorithm can be computed guaranteeing that in the resulting MDP $\mathcal{M}'$ all optimal policies are ethical-optimal.

To prove that there always exists a solution ethical weight for any input MDP with reward function R_0 is equivalent to proving that $\mathbf{V}^*$ always belongs to the convex hull. Consider the ethical-optimal value vector $\mathbf{V}^* = (V_0^*, V_N^* + V_E^*)$, and any value vector $\mathbf{V} = (V_0, V_\mathcal{N} + V_E)$ of an unethical (i.e., not ethical) policy of $\mathcal{M}$ such that $V_0 > V_0^*$. We will prove that there is an w_e for which the value of $\mathbf{V}^*$ is greater than $\mathbf{V}$, hence proving also that $\mathbf{V}^*$ indeed belongs to the convex hull.

Consider the lines that the two aforementioned value vectors form in the weight space: $(1-w) \cdot V_0 + w \cdot (V_\mathcal{N} + V_E)$ for the unethical policy, and $(1-w) \cdot V_0^* + w \cdot (V_N^* + V_E^*)$ for the ethical-optimal value vector. Consider the line of their subtraction as a function f depending of w:

$$f(w) = (1-w) \cdot (V_0^* - V_0) + w \cdot (V_N^* + V_E^* - V_\mathcal{N} - V_E).$$

It is clear that $f(0) < 0$ and $f(1) > 0$. Thus, by Bolzano's Theorem, there exists another point $0 < w_e < 1$ such that w_e is a root of f, that is, $f(w_e) = 0$. Since $f(w)$ is linear, then $f(w)$ will be positive for any $w \in (w_e, 1)$. Therefore, if we select the unethical policy such that $f(w)$ has the greatest root w_e^*, for any $w \in (w_e^*, 1)$, the value vector of the ethical-optimal policy will be greater than that of any other policy. In conclusion, $\mathbf{V}^*$ belongs to the convex hull. □

In practice, Theorem 2 ensures that Algorithm 1 will always yield an environment where the optimal policy is ethical-optimal. If an agent situated in such ethical environment is endowed with a learning algorithm capable of finding the optimal policy, then the agent will learn an ethical behaviour.

It is important to highlight that an autonomous agent in our ethical environment is free to either behave ethically or not. Actually, when learning, an agent not following the norms is penalised. Our design of the environment makes that the optimal policy to learn, the one that gives more reward to the agent, fulfils all the norms of a given moral value and behaves as much praiseworthily as possible. This is what we refer to when we say that Algorithm 1 guarantees the learning of an ethical-optimal policy.

The next subsection illustrates, in our example, that a simple algorithm like Q-learning can do the job.

Example 8 For the public civility game, the last step in our algorithm returns an MDP $\mathcal{M}'$ whose reward comes from scalarising the MOMDP by $\mathbf{w} = (1, w_e)$, being w_e strictly greater than 7. Thus, adding any $\epsilon > 0$ will suffice.

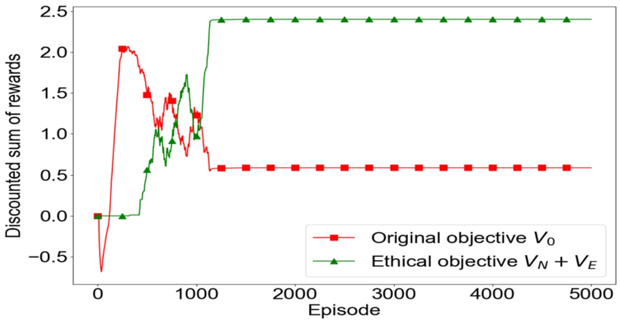


Fig. 5 Evolution of the accumulated rewards per episode that the agent obtains in the ethical environment

If, for instance, we set $\epsilon = 0.1$ then, the weight vector $(1, 7 + 0.1) = (1, 7.1)$ solves the Public Civility Game. More specifically, an MDP created from an embedding function with such ethical weight w_e incentivises the agent to learn the **E**thical (civic) policy. Such MDP will have the reward function $R_0 + 7.1(R_{\mathcal{N}} + R_E)$.

Analysis: learning in an ethical environment

After creating the ethical environment $\mathcal{M}'$ with reward function $R_0 + 0.71(R_{\mathcal{N}} + R_E)$ for the public civility game, we can illustrate our theoretical results by letting the agent learn an optimal policy in $\mathcal{M}'$. With that aim, we endow the learning agent with Q-learning (Watkins & Dayan 1992) as its learning algorithm. In Q-learning, we need to specify two hyperparameters: the learning rate $\alpha \in (0, 1]$ and the discount factor $\gamma \in (0, 1]$. In our case, we set them to $\alpha = 0.8$ and $\gamma = 0.7$. A large discount factor γ makes sense for environments that are episodic such as ours, while the impact of the value of the learning rate α is not significant in deterministic environments such as ours. Furthermore, we set the learning policy to be ϵ-greedy (Sutton & Barto, 1998), the simplest option. Applying Q-learning with the ϵ -greedy learning policy, the agent is guaranteed to learn an optimal policy if it trains during enough iterations (Sutton & Bar,m 1998).

After letting the agent learn for 5000 iterations, it ends up learning the Ethical policy: to bring the garbage to the wastebasket while moving towards its goal. The result was expected because: (1) Theorem 2 guarantees that all optimal policies are ethical-optimal; and (2) the use of Q-learning by the agent ensures the learning of the optimal policy (that is also ethical-optimal).

Figure 5 shows how the agent's value vector **V** stabilises, with less than 1500 episodes, at 0.59 (V_0 line) and 2.4 ($V_{\mathcal{N}} + V_E$ line), which is precisely the value of the **E**thical policy.

Related work

The AI literature on value alignment is typically divided between top-down, bottom-up, and hybrid approaches, as surveyed in Allen et al. (2005), Tolmeijer et al. (2021). In brief, top-down approaches focus on formalising ethical knowledge to encode it directly into the agent's behaviour, whereas bottom-up approaches resort on the agent learning the ethical knowledge by itself. Hybrid approaches combine bottom-up and top-down approaches.

Some top-down proposals of formalising moral values include the work of Sierra et al. (2019), in which values are formalised as preferences, and also the work of Mercuur et al. (2019), in which values and norms are formalised as two distinct concepts, where values serve as a static component in agent behaviour, whereas norms serve as a dynamic component. There has also been studies about the formal relationship between norms and values by Hansson and Hendricks (2018), and even some attempts at formalising supererogatory actions (for instance, in McNamara (1996), Hansson (2013)). Other top-down approaches more related with AI Safety focus on defining a set of safety constraints that an agent must comply with, hence formalising its problem as a Constrained MDP (Chow et al., 2018; García & Fernández 2015; Miryoosefi et al., 2020). Notice, however, that the framework of Constrained MDPs cannot express an ordering between objectives such as the one performed in this work. In summary, while all of the mentioned formal work is a clear contribution to the area, it is also widely accepted that pure top-down approaches cannot deal with the whole value alignment problem, as explained by Arnold *et al.* in Arnold et al. (2017).

Regarding bottom-up approaches, they almost exclusively focus on reinforcement learning for teaching moral values, following the proposed approaches of Russell, Soares and Fallenstein, among others (Russell et al., 2015; Soares & Fallenstein 2014). In particular, *inverse* reinforcement learning (IRL) (Abbeel & Ng 2004) has been proposed as a viable approach for solving the value alignment problem. Inverse reinforcement learning deals with the opposite problem of reinforcement learning: to learn a reward function from a policy. Hence, applying IRL, the agent would be able to infer the values of humans by observing their behaviour. Examples of the use of IRL for the value alignment problem include (Abel et al., 2016; Hadfield-Menell et al., 2016; Noothigattu et al., 2019; Riedl & Harrison, 2016; Wu & Lin, 2017).

One of the first criticisms that IRL received about tackling the value alignment problem was expressed by Arnold et al. (2017). The authors claim that IRL cannot infer that there are certain norms that the agent needs to follow. Arnold et al. propose instead to combine the strength of RL and logical representations of norms as a hybrid approach. Following the proposal of Arnold et al., an agent would learn to maximise a reward function while satisfying some norms at the same time. While we consider this approach related to ours, we differ in that we are capable of also integrating norms directly into the agent's ethical reward function via carefully dividing it into two components.

Another major criticism of the majority of bottom-up approaches consider the problem of reward specification as equivalent to the whole value alignment problem. This has only recently started to be considered as a two-step process (reward specification and ethical embedding) that must take into account that the agent will have its own objectives (for instance, in Wu and Lin (2017), Noothigattu et al. (2019), Balakrishnan et al. (2019)).

While the value alignment literature typically considers a single learning agent, results for multi-agent systems are still scarce (notice how all the aforementioned works were approaches for a single agent). Some related areas for multi-agent systems are mechanism design and co-utility. They both address the development of agent-interaction protocols or mechanisms in which no agent is worse off by participating (Domingo-Ferrer et al., 2017; Nisan & Ronen, 2001). In more detail, the problem in mechanism design is to design a mechanism for a multi-agent system that yields a socially desirable outcome. Similarly, co-utility aims at promoting a mutually beneficial collaboration between agents. Both methods differ from value alignment in that they only consider the individual utility function of each agent, disregarding any external ethical objective nor considering whether or not the maximisation of the agents' utility functions is compatible with a value-aligned behaviour..

Finally, recent studies in cognitive science also remark the influence of the environment on human moral behaviour (Gigerenzer, 2010). According to Gigerenzer, moral behaviour in real environments is not based on maximising an ethical utility function, but instead on following some heuristics. This is also the point of our work: that instead of demanding the agent to maximise the ethical reward function, we design the environment in such a way that it is naturally inclined to behave ethically even with the simplest reinforcement learning algorithms.

Conclusions and future work

Designing algorithms for guaranteeing agents' value alignment is a challenging problem. We make headway in tackling this problem by providing a novel algorithmic approach for tackling the whole value alignment problem. Our approach builds upon formal philosophy and multi-objective reinforcement learning. In particular, our approach ensures that the agent wholly fulfils its ethical objective while pursuing its individual objective.

Overall, we design a method for guaranteeing value-alignment by considering a two-step process. It firstly specifies ethical behaviour as ethical rewards, and then embeds such rewards into the learning environment of the agent.

We formalise the first step as the ethical reward specification problem, and we provide a solution to it via specifying our formalisation of moral values with MORL, a valuable framework to handle multiple objectives. We do so by first formalising *moral values* based on moral philosophy. Our reward specification of a moral value guarantees that any agent following it will be value-aligned. We formalise the second and last step as the ethical embedding problem, and provide a method –within the MORL framework– to solve it.

Our findings lead to an algorithm for automating the whole value-alignment process. Our algorithm builds an ethical environment in which it will be in the best interest of the agent to behave ethically while still pursuing its individual objective. We illustrate our approach by means of an example that embeds the moral value of civility.

As to future work, we would like to go beyond a single moral value, as considered in this paper, and extend our approach to be capable of coping with multiple moral values in a value system. We expect to create such extension by, for instance, considering a (pre-defined) ranking over moral values that allows us to accommodate opposing moral norms in our approach. As a reference, we have identified the work in Serramia et al. (2018, 2020) as promising regarding how to tackle clashing norms that support different moral values.

We would also like to further investigate the potential applicability of our approach in more complex environments (such as P2P networks, multi-agent environments, agent-human collaboration environments and so on) and study how to include an ethical reward function of a given moral value to those environments.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Research supported by projects AI4EU (H2020-825619), Crowd4SDG (H2020-872944), TAILOR (H2020-952215), COREDEM (H2020-785907), NANOMOOC (COMR-DI18-1-0010-02), and 21S01802-001 from Barcelona City Council through the Fundació Solidaritat de la UB. Financial support was also received from grant PID2019-104156GB-I00 funded by MCIN/AEI/10.13039/501100011033. Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387).

Code availability The code generated and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, p. 1. ACM, New York, NY, USA. https://doi.org/10.1145/1015330.1015430.

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Work.: AI, Ethics, and Society* (vol. 92).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7,* 149–155. https://doi.org/10.1007/s10676-006-0004-4.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. CoRR arXiv:1606.06565.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment - what will keep systems accountable?. In AAAI Workshops

Audi, R. (1999). *The Cambridge Dictionary of Philosophy*. Cambridge University Press.

Balakrishnan, A., Bouneffouf, D., Mattei, N., & Rossi, F. (2019). Incorporating behavioral constraints in online ai systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 33,* 3–11. https://doi.org/10.1609/aaai.v33i01.33013.

Barcaro, R., Mazzoleni, M., & Virgili, P. (2018). Ethics of care and robot caregivers. *Prolegomena, 17,* 71–80. https://doi.org/10.26362/20180204.

Barrett, L., & Narayanan, S. (2008). Learning all optimal policies with multiple criteria. In Proceedings of the 25th International Conference on Machine Learning pp. 41–47. https://doi.org/10.1145/1390156.1390162

Camps, V. (2013). *Brief history of ethics*. BA.

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

Chow, Y., Nachum, O., Duenez-Guzman, E., & Ghavamzadeh, M. (2018). A lyapunov-based approach to safe reinforcement learning. NIPS'18.

Conee, E. (1982). Against moral dilemmas. *The Philosophical Review,91*(1), 87–97. http://www.jstor.org/stable/2184670.

Cooper, D. (1993). *Value pluralism and ethical choice*. St. Martin Press Inc.

Domingo-Ferrer, J., Martínez, S., Sínchez, D., & Soria-Comas, J. (2017). Co-utility: Self-enforcing protocols for the mutual benefit of participants. *Engineering Applications of Artificial Intelligence, 59,* 148–158. https://doi.org/10.1016/j.engappai.2016.12.023.

Duignan, B. (2018). Ought implies can. Retrieved January 15, 2015, from https://www.britannica.com/topic/ought-implies-can

Etzioni, A., & Etzioni, O. (2016). Designing ai systems that obey our laws and values. *Communications of the ACM, 59*(9), 29–31. https://doi.org/10.1145/2955091.

Fieser, J., & Dowden, B. (2000). *Ethics*. https://www.iep.utm.edu/ethics/ (The Internet Encyclopedia of Philosophy).

Frankena, W. K. (1973). *Ethics* (2nd ed.). Prentice-Hall.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30,* 411–437. https://doi.org/10.1007/s11023-020-09539-2.

García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research, 16*(1), 1437–1480.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science,2*(3), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2010.01094.x.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in Neural Information Processing Systems 29, pp. 3909–3917. Berkeley.

Hansson, S. O. (2013). Representing supererogation. *Journal of Logic and Computation, 25*(2), 443–451. https://doi.org/10.1093/logcom/exs065.

Hansson, S. O., & Hendricks, V. (2018). *Introduction to Formal Philosophy*. Springer.

Heyd, D. (2016). Supererogation. In Zalta E. N. (ed.) The Stanford encyclopedia of philosophy, spring 2016 edn. https://plato.stanford.edu/entries/supererogation/

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *The Journal of Artificial Intelligence Research, 4*(1), 237–285.

Leike, J., Martic, M., Krakovna, V., Ortega, P., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). Ai safety gridworlds. arXiv:1711.09883.

Lin, P. (2015). *Why ethics matters for autonomous cars* (pp. 69–85). Springer. https://doi.org/10.1007/978-3-662-45854-9_4.

Littman, M. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature, 521,* 445–51. https://doi.org/10.1038/nature14540.

McNamara, P. (1996). Doing well enough: Toward a logic for commonsense morality. *Studia Logica, 57*(1), 167–192. https://doi.org/10.1007/BF00370674.

Mercuur, R., Dignum, V., Jonker, C., et al. (2019). The value of values and norms in social simulation. *Journal Artificial Societies and Social Simulation, 22*(1), 1–9.

Miryoosefi, S., Brantley, K., Iii, H., Dudík, M., & Schapire, R. (2020). Reinforcement learning with convex constraints. In Advances in Neural Information Processing Systems.

Nisan, N., & Ronen, A. (2001). Algorithmic mechanism design. *Games and Economic Behavior,35*(1), 166–196. https://doi.org/10.1006/game.1999.0790. https://www.sciencedirect.com/science/article/pii/S089982569990790X.

Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Kush, R., et al. (2019). Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development, PP,* 6377–6381. https://doi.org/10.1147/JRD.2019.2940428.

Riedl, M. O., & Harrison, B. (2016). Using stories to teach human values to artificial agents. In AAAI Workshop: AI, Ethics, and Society.

Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodríguez-Aguilar, J. A. (2020). A structural solution to sequential moral dilemmas. In

Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Aystems (AAMAS 2020).

Roijers, D., & Whiteson, S. (2017). Multi-Objective Decision Making. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool, California, USA. https://doi.org/10.2200/S00765ED1V01Y201704AIM034. http://www.morganclaypool.com/doi/abs/10.2200/S00765ED1V01Y201704AIM034.

Rossi, F., & Mattei, N. (2019). Building ethically bounded ai. *Proceedings of the AAAI Conference on Artificial Intelligence, 33,* 9785–9789. https://doi.org/10.1609/aaai.v33i01.33019785.

Russell, S. (2019). *Human compatible. AI and the problem of control*. Penguin Books.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine, 36,* 105–114. https://doi.org/10.1609/aimag.v36i4.2577.

Serramia, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2020). A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, p. 1233–1241. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., & Ansotegui, C. (2018). Moral values in norm decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18)*, pp. 1294–1302. International Foundation for Autonomous Agents and Multiagent Systems.

Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., & Perello-Moragues, A. (2019). Value alignment: A formal approach. In *Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS 2019*.

Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report 8*.

Sutrop, M. (2020). Challenges of aligning artificial intelligence with human values. *Acta Baltica Historiae et Philosophiae Scientiarum, 8,* 54–72. https://doi.org/10.11590/abhps.2020.2.04.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning—an introduction. Adaptive computation and machine learning*. MIT Press. http://www.worldcat.org/oclc/37293240.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in machine ethics: A survey. *ACM Computing Surveys*. https://doi.org/10.1145/3419633.

Urmson, J. O. (1958). Saints and heroes. In A. I. Melden (Ed.), *Essays in moral philosophy*. University of Washington Press.

van de Poel, I., & Royakkers, L. (2011). *Ethics, technology, and engineering: An introduction*. Wiley-Blackwell.

Watkins, C. J. C. H., & Dayan, P. (1992). Technical note q-learning. *Machine Learning, 8,* 279–292. https://doi.org/10.1007/BF00992698.

Wu, Y. H., & Lin, S. D. (2017). A low-cost ethics shaping approach for designing reinforcement learning agents. arXiv.

Wynsberghe, A. (2016). Service robots, care ethics, and design. *Ethics and Information Technology, 18*(4), 311–321. https://doi.org/10.1007/s10676-016-9409-x.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In: *IJCAI*, pp. 5527–5533.

Zimmerman, M. J. (1987). Remote obligation. *American Philosophical Quarterly, 24*(2), 199–205.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.