UNIVERSITAT DE BARCELONA

Reinforcement Learning for Value Alignment

Manel Rodríguez Soto

UNIVERSITAT DE BARCELONA

Reinforcement Learning for Value Alignment

by

Manel Rodríguez Soto

A dissertation submitted in partial satisfaction
of the requirements for the award of the degree of
Doctor of Philosophy in the Program of Engineering and Applied Sciences.

Advisers:

Dr. Maite López Sánchez

Dr. Juan Antonio Rodríguez Aguilar

Barcelona, April 2023

Abstract

As autonomous agents become increasingly sophisticated and we allow them to perform more complex tasks, it is of utmost importance to guarantee that they will act in alignment with human values. This problem has received in the AI literature the name of the *value alignment* problem. Current approaches apply *reinforcement learning* to align agents with values due to its recent successes at solving complex sequential decision-making problems. However, they follow an *agent-centric* approach by expecting that the agent applies the reinforcement learning algorithm correctly to learn an ethical behaviour, without formal guarantees that the learnt ethical behaviour will be ethical. This thesis proposes a novel *environment-designer* approach for solving the value alignment problem with theoretical guarantees.

Our proposed environment-designer approach advances the state of the art with a process for designing ethical environments wherein it is in the agent's best interest to learn ethical behaviours. Our process specifies the ethical knowledge of a moral value in terms that can be used in a reinforcement learning context. Next, our process embeds this knowledge in the agent's learning environment to design an *ethical* learning environment. The resulting ethical environment incentivises the agent to learn an ethical behaviour while pursuing its own objective.

We further contribute to the state of the art by providing a novel algorithm that, following our ethical environment design process, is formally guaranteed to create ethical environments. In other words, this algorithm guarantees that it is in the agent's best interest to learn value-aligned behaviours.

We illustrate our algorithm by applying it in a case study environment wherein the agent is expected to learn to behave in alignment with the moral value of respect. In it, a conversational agent is in charge of doing surveys, and we expect it to ask the users questions respectfully while trying to get as much information as possible. In the designed ethical environment, results confirm our theoretical results: the agent learns an ethical behaviour while pursuing its individual objective.

Resum

A mesura que els agents autònoms es tornen cada cop més sofisticats i els permetem realitzar tasques més complexes, és de la màxima importància garantir que actuaran d'acord amb els valors humans. Aquest problema ha rebut a la literatura d'IA el nom del problema d'*alineació de valors*. Els enfocaments actuals apliquen *aprenentatge per reforç* per alinear els agents amb els valors a causa dels seus èxits recents a l'hora de resoldre problemes complexos de presa de decisions seqüencials. Tanmateix, segueixen un enfocament *centrat en l'agent* en esperar que l'agent apliqui correctament l'algorisme d'aprenentatge de reforç per aprendre un comportament ètic, sense garanties formals que el comportament ètic après serà ètic. Aquesta tesi proposa un nou enfocament de *dissenyador d'entorn* per resoldre el problema d'alineació de valors amb garanties teòriques.

El nostre enfocament de disseny d'entorns proposat avança l'estat de l'art amb un procés per dissenyar entorns ètics en què és del millor interès de l'agent aprendre comportaments ètics. El nostre procés especifica el coneixement ètic d'un valor moral en termes que es poden utilitzar en un context d'aprenentatge de reforç. A continuació, el nostre procés incorpora aquest coneixement a l'entorn d'aprenentatge de l'agent per dissenyar un entorn d'aprenentatge *ètic*. L'entorn ètic resultant incentiva l'agent a aprendre un comportament ètic mentre persegueix el seu propi objectiu.

A més, contribuïm a l'estat de l'art proporcionant un algorisme nou que, seguint el nostre procés de disseny d'entorns ètics, està garantit formalment per crear entorns ètics. En altres paraules, aquest algorisme garanteix que és del millor interès de l'agent aprendre comportaments alineats amb valors.

Il·lustrem el nostre algorisme aplicant-lo en un estudi de cas on s'espera que l'agent aprengui a comportar-se d'acord amb el valor moral del respecte. En ell, un agent de conversa s'encarrega de fer enquestes, i esperem que faci preguntes als usuaris amb respecte tot intentant obtenir la màxima informació possible. En l'entorn ètic dissenyat, els resultats confirmen els nostres resultats teòrics: l'agent aprèn un comportament ètic mentre persegueix el seu objectiu individual.

Acknowledgments

The past four years have been a wild ride of working on exciting projects, learning about the most interesting theories, and collaborating with brilliant minds. Looking back, this was not produced by luck: it was the result of the hard work of many people to whom I am very grateful. Without them this thesis would not have been possible.

First and foremost, I would like to thank my advisers, Maite López Sánchez and Juan Antonio Rodríguez Aguilar, for their invaluable guidance, support, and encouragement throughout my whole thesis. I was incredibly fortunate of meeting you and having you as my supervisors, you are the best teachers I could ever have. Thank you for believing in me ever since I did that undergraduate internship at IIIA back in 2017.

Thank you Maite for always giving me such insightful and exhaustive comments, your critical feedback, and for your constructive criticisms. Thank you for always being available for all our dialogues and discussions about Ethics and Values. Your constant commitment to excellence and hard work have been a source of inspiration and motivation for me.

Thank you Jar for teaching me how to think abstractly about research and how to make inquisitive questions. Thank you for always answering all my doubts. Your expertise and intellectual curiosity will never cease to fascinate me.

Thank you both for teaching me how to be researcher, for all your kindness, and for always treating me as a colleague and a friend.

My immense gratitude to Ann Nowé for opening the doors of her lab to me and bringing me the opportunity to learn from her expertise. I hope to work again with you soon.

I would like to express my heartfelt appreciation to my colleagues with whom I have had the opportunity to learn from and work with them. To Marc Serramià, for always guiding my path, for your camaraderie and for his insights on norms and moral values. To Roxana Rădulescu for sharing with me her invaluable knowledge on multi-agent reinforcement learning.

Thanks to all the people at the Artificial Intelligence Research Institute (IIIA-CSIC) for providing me an excellent working environment for developing my thesis. There I was able of meeting wonderful scientists from all the parts of the world. I want to give a special shoutout to all the PhD students in the lab for all the good memories and conversations.

Thanks to my two brothers-in-arms during my thesis. Thank you Tomeu for all the good moments, conversations and discussions. Reading all these books about reinforcement

learning has been a blast together with you. Thank you David for all your wisdom and for guiding me through the maze of FPU bureaucracy.

On a more personal note, I would like to thank my family and closest friends. They have been my pillar during the whole thesis. It is thanks to them that all of this was worth it.

Thanks to my friends for filling my life with joy. Jordi, Bernat, Miquel, Albert, Ramon, Alejandro, Souffian. I cannot imagine my life without you.

I want to thank my family for their unconditional and continuous support. To my mother Susi and my father Manolo for all your love and affection. I love you too.

Finally, I want to dedicate this thesis to my best friend and special one, Núria. Thank you for everything: for for all the time spent together, for always supporting me, all the laughs, and for shedding light when there was darkness. In short, thank you for existing.

Manel Rodríguez Soto
Barcelona, April 2023

This work has been funded by the Spanish government through the grant "Formación del Profesorado Universitario" (FPU) with reference FPU18/03387.

Table of Contents

Chapter 1

Introduction

As autonomous agents become more prevalent in our society, ensuring that they act in accordance with human values (i.e., *value-aligned*) has become a critical challenge [Russell et al., 2015, Soares and Fallenstein, 2014]. It is therefore of great concern to develop trustworthy AI [Chatila et al., 2021] aligned with ethical principles. Such AI needs to be capable of respecting human values in various emerging application domains, such as social assistive robotics [Boada et al., 2021], self-driving cars [Hansson, 2001], and conversational agents [Casas-Roma and Conesa, 2020]. The Machine Ethics [Rossi and Mattei, 2019, Yu et al., 2018] and AI Safety [Amodei et al., 2016, Leike et al., 2017] communities have recently shown a rising interest in using *Reinforcement Learning* (RL) to address the critical problem of value alignment. This interest is due to the recent successes of reinforcement learning in solving a plethora of complex problems such as winning at many arcade video games [Mnih et al., 2013], mastering competitive games such as go or StarCraft [Garisto, 2019, Silver et al., 2017], crewless aerial vehicles [Azar et al., 2021], and finding novel ways of fast matrix multiplication [Fawzi et al., 2022] among others. For this reason, reinforcement learning has gained traction in both AI communities as a promising candidate for solving the *value alignment* problem.

The standard approach these two communities share involves designing an environment with incentives encouraging ethical learning. In such an approach, an agent receives rewards through an exogenous function based on ethical knowledge (e.g., [Abel et al., 2016, Balakrishnan et al., 2019, Noothigattu et al., 2019, Riedl and Harrison, 2016, Vamplew et al., 2021, Wu and Lin, 2018]) employing a two-step process. First, the ethical reward function is *specified* from some ethical knowledge. Afterwards, rewards are aggregated to the agent's learning environment through an *ethical embedding* process. However, ensuring that agents learn to behave ethically in such an environment remains an open problem. The typical approach is agent-centric: to expect that the agent

applies the learning algorithm appropriately without formal guarantees that it will learn to behave ethically.

Against this background, this thesis follows a mechanism design approach [Vlassis, 2009] and aims to automate the design of *ethical environments*, taking an *environment-centric* approach. In an ethical environment, rewards are tailored so that it is in the agents' best interest to behave ethically-aligned with human values. Our proposed *ethical environment design process* is founded in *multi-objective* reinforcement learning (MORL) [Hayes et al., 2022, Roijers and Whiteson, 2017, Rădulescu et al., 2019]. This process computes how to reward the learning agent so that it is incentivised to learn an ethical behaviour.

This chapter is structured as follows. Section 1.1 motivates the research questions of this thesis. Section 1.2 continues by enumerating the research questions. After that, Section 1.3 summarises the contributions provided to answer the research questions. Section 1.4 outlines the structure of this thesis. Finally, Section 1.5 enumerates all the published work derived from this thesis.

1.1 Motivation

In this section, we further introduce the open problem of value alignment and the tools we currently have for solving it. The main ideas explored here motivate all the research questions in this thesis. Next, Section 1.1.1 introduces an explanation of the value alignment problem and discusses its importance. Then, Section 1.1.2 explains the basic concepts of reinforcement learning and its applicability to solve the value alignment problem. We continue in Section 1.1.3 by summarising the current research in reinforcement learning applied to value alignment and its open problems. Finally, Section 1.1.4 illustrates the potential of an environment-designer approach to solving the previously presented open problems.

1.1.1 The value alignment problem

Autonomous agents are becoming progressively more intelligent, and hence they increasingly imbue our daily lives. However, the advent of artificial intelligence comes with a big caveat: we need to ensure that agent's objectives and human objectives will not eventually become mismatched, which could have catastrophic consequences [Russell et al., 2015]. It becomes of utmost importance that we design methods for guaranteeing that agents always comply with human values, as expressed by Gabriel [2020]. This pressing challenge has received the name of the *value alignment* problem.

The *value alignment* problem is the problem of guaranteeing that agents act in alignment with human values [Russell et al., 2015, Soares and Fallenstein, 2014]. An agent exhibiting a value-aligned behaviour ought to always pursue objectives beneficial to humans [Arnold et al., 2017, Russell et al., 2015, Soares and Fallenstein, 2014, Sutrop, 2020]. Value alignment is still a novel research area, and thus there is still a lack of a standard definition of what a human *value* is when referring to the value alignment problem. In this thesis, we follow the philosophical views of Arnold *et al.* [Arnold et al., 2017, Gabriel, 2020, Sutrop, 2020] by defining *values* as: *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.* Moral values provide us with the ethical knowledge needed for our daily life. They tell us that, for example, behaving civilly and safely is good and that promoting inequality is wrong. With our stance, we are remarking that to behave value-aligned is to behave *ethically.* In the remainder of this thesis, we use both terms (*ethical* and *value-aligned*) interchangeably.

1.1.2 Reinforcement Learning

In most scenarios, an agent is deployed for a certain amount of time to fulfil its own objective (for example, reaching a goal position as fast as possible). This agent aims to learn a sequence of actions to fulfil its own objective. That means that the problem of guaranteeing that an agent acts ethically will require that when the agent is deployed, it always selects actions considering their ethical consequences. Therefore, in this thesis, we argue that the value alignment problem is, in essence, a *sequential decision-making* problem.

Once we frame the problem of guaranteeing ethical behaviour as sequential, it becomes natural to look at the reinforcement learning framework for value alignment (example applications include [Abel et al., 2016, Balakrishnan et al., 2019, Noothigattu et al., 2019, Wu and Lin, 2018]). Reinforcement learning is the most prominent framework for sequential decision-making (with or without uncertainty) nowadays. Indeed, the advent of reinforcement learning applications for guaranteeing value alignment is due to its potential to solve complex sequential decision problems. Since the surge of Deep Reinforcement Learning [François-Lavet et al., 2018], there has been an explosion of algorithms to solve sequential problems that range from beating world champions of chess and Go [Schrittwieser et al., 2019, Silver et al., 2017] to winning at realistic racing simulators like Gran Turismo [Wurman et al., 2022].

In reinforcement learning, an agent learns its behaviour via a trial-and-error scheme: while learning, the agent keeps acting upon its environment, and after each action, it

receives a reward as feedback that can be either positive, negative, or null, and also observes how its action changes the environment [Littman, 2015, Sutton and Barto, 1998]. By repeating this State-Action-Reward-State (SARS) loop, the agent eventually learns the sequence of actions that maximises its accumulation of rewards. Formally, rewards are specified in the *reward function* of the environment, and the behaviour capable of maximising the accumulation of rewards from its reward function is called the *optimal policy*. A reinforcement learning environment is known as a *Markov Decision Process* if there is a single learning agent or a *Markov Game* if there are multiple learning agents. Typically, the learning environment of the agent is assumed to be *single-objective*. The agent has a unique sequential problem to solve and receives a reward signal accordingly. When the agent has to deal with two or more objectives simultaneously, the environment is then formalised as a *Multi-Objective* Markov Decision Process [Hayes et al., 2022, Roijers and Whiteson, 2017, Rădulescu et al., 2019]. In a multi-objective environment, the agent has as many reward functions as objectives. The paradigmatic example in this thesis is an agent that has, apart from its original reward function regarding his individual objective, an extra ethical reward function. For the agent to learn to consider both objectives, the natural approach is to model its learning environment as multi-objective.

However, notice that there is no guarantee that a learning agent in a multi-objective environment learns to behave ethically when employing a MORL algorithm. The learnt behaviour will depend on whether it prioritises the ethical objective, but the environment designer has no guarantees that the latter will occur. Therefore, we argue for our environment-designer approach of incentivising ethical behaviour learning by designing *ethical* single-objective environments. Such a single-objective environment encapsulates the two previous without increasing the learning complexity of the agent.

Finally, in this thesis, we advocate for providing formal guarantees for our ethical environment design approach. We want to theoretically guaarntee that in the designed ethical environment, the agent is incentivised to behave value-aligned. Therefore, in this thesis, we limit ourselves to applying tabular reinforcement learning algorithms. Despite the impressive empirical results of deep reinforcement learning algorithms, they all share an important problem: no algorithm is formally guaranteed to obtain an optimal policy. Fortunately, *tabular* reinforcement learning (i.e., without deep learning) has been much more profoundly studied and counts with algorithms that are mathematically proven to yield optimal policies.

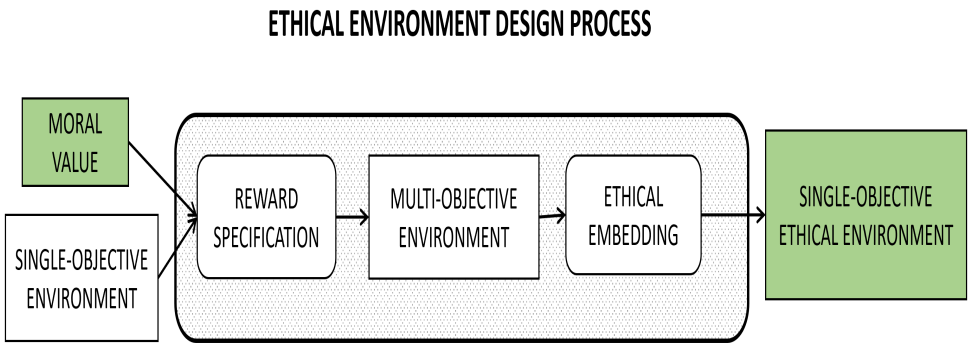


FIGURE 1.1: The ethical environment design process for aligning the behaviour of the agent to one moral value is performed in two steps: a reward specification and an ethical embedding. Rectangles stand for objects whereas rounded rectangles correspond to processes.

1.1.3 Ethics and Reinforcement Learning

As previously mentioned, one of the current mainstream approaches in the AI literature is to use Reinforcement Learning (RL) to align the agent's behaviour with human moral values. In summary, the reinforcement learning approach to ethically align the behaviour of an agent always follows the following two steps:

1. **Reward specification:** the ethical knowledge at hand is specified as an *ethical* reward function that positively rewards those actions that are ethical (i.e., that are valued as right from an ethical point of view) and penalises unethical actions. In other words, we transform the ethical knowledge of the moral value so it can be applied in reinforcement learning. It is the first process in Figure 1.1.

2. **Ethical embedding:** the ethical reward function is incorporated in the agent's learning environment so that the agent considers these ethical rewards. It is the second process in Figure 1.1.

When the two steps are correctly performed, the agent will learn to behave ethically, in alignment with a given moral value. However, as previously mentioned, the main problem with state-of-the-art approaches is that they have no theoretical guarantees. Hence, we only know if an agent will learn an ethical behaviour or not once we evaluate its learnt behaviour. There is nothing to do on this front for the approaches using deep reinforcement learning due to current theoretical limitations. However, we can already tackle the problem of providing guarantees for learning agents applying tabular reinforcement learning. Hence, the main focus of this thesis is on agents that learn by applying tabular reinforcement learning. For such kind of agent, *we provide formal guarantees* that it is in its best interest to behave ethically in the designed environment.

A remaining important question is how to formalise moral values to include them in our algorithms. This formalisation is indispensable to incorporate them into the behaviours of artificial agents. We already find in the Artificial Intelligence community some proposals to formally encode moral values (for instance, [Bench-Capon and Atkinson, 2009, Luo et al., 2017, Serramia et al., 2018]). This thesis follows the formalisation used in [Serramia et al., 2018] and adapts it to reinforcement learning environments.

We present now an example environment to illustrate the necessary elements that should be present in a reinforcement learning formalisation of a moral value. This example is called the *Public Civility Game* environment [Rodriguez-Soto et al., 2020]. In it, two autonomous agents move through a narrow path towards their respective goal positions. Despite being a small toy environment, it presents all the problems a learning agent faces when dealing with ethical decision-making. In more detail, the agent at the left faces this problem: someone has left a piece of garbage blocking its path. We expect the agent to solve this problem in accordance with *civility* (the moral value considered in this example) without disregarding its initial objective (reaching the goal). To proceed towards its destination, the agent can choose between these three actions: (i) immediately throw the garbage away, which may hurt the other agent, (ii) wait until the other agent is not nearby to push the garbage aside gently, or (iii) bring the garbage to the nearest bin. Each of the three actions is increasingly more aligned with the moral value of *civility*. For the agent to behave in alignment with such a moral value, we need to have all the necessary ethical knowledge stored in our formalisation of the moral value. In this case, throwing the garbage aside strongly demotes the value of civility, pushing it aside is neutral to it, and bringing it to the bin strongly promotes the value.

For the agent to learn to behave ethically (to bring the garbage to the bin in the case of the Public Civility Game), the agent must know for each possible action in the environment if it is ethical or not. The ethical reward function provides that ethical information in the form of rewards. The design of such an ethical reward function is the objective of the first step of our ethical environment design process: the reward specification process. In the case of the Public Civility Game, the specified ethical reward function would return a positive reward for bringing the garbage to the bin and a negative reward for throwing the garbage away.

The specification of the ethical reward function extends the agents' original learning environment (the input environment in Figure 1.1) into a multi-objective one (the environment resulting from the reward specification in Figure 1.1). As we have mentioned, in this thesis we propose to use an environment-designer approach in which the agent learns to behave ethically in an ethical single-objective environment. Therefore, in our approach, the ethical embedding process transforms the learning environment into an

(ethical) single-objective environment, the output in Figure 1.1. Recall that, in reinforcement learning, each objective is represented by its corresponding reward function. Hence, to transform a multi-objective environment into a single-objective one amounts to *scalarising* the vectoral reward function of the multi-objective into a scalar reward function. That is, we aggregate the two different reward functions by giving each of them a weight indicating their relative importance. This way, our ethical embedding process returns an ethical single-objective environment.

Back to the Public Civility Game, suppose that the agent's original reward function R_0 provides a greater reward the sooner it reaches its destination. For simplicity, assume that the reward for reaching the destination as fast as possible is $R_0(fast) = 10$ while the reward for reaching it later is $R_0(slow) = 5$. Assume that the ethical reward R_v for bringing the garbage to the bin is $R_v(bin) = 1$, and the ethical penalty for throwing it to the other agent is $R_v(throw) = -1$. Then, we would need to select a weight w such that:

$$R_0(slow) + w \cdot R_v(bin) > R_0(fast) + w \cdot R_v(throw).$$

In this particular case, any weight $w > 2.5$ is *guaranteed* to incentivise the agent to behave ethically. That is, in the learning environment designed with the single reward function $R_0 + w \cdot R_v$ with $w > 2.5$, the rational decision for the agent is to learn to behave ethically. Notice that we computed the value of the necessary weight analytically instead of fine-tuning it empirically. This computation is the gist of our ethical embedding process: to compute the weighting of the two objectives (individual and ethical) analytically so that it is formally guaranteed to incentivise ethical behaviour learning.

In conclusion, in this thesis, we argue for applying a two-step process for designing ethical environments that: (i) requires a formalisation of moral values which makes explicit the relationship between actions and values, so agents can apply it (this way, we will be able to encode all the ethical knowledge of a moral value as rewards of a reinforcement learning environment). (ii) After the ethical rewards are formalised in a multi-objective environment, we propose to *scalarise* the rewards for designing an ethical single-objective environment.

1.1.4 The environment designer approach

Reinforcement learning mainly focuses on the learning process of a single agent in a given environment. We refer to this as the *agent-centric* approach: to leave the environment as-is and instead focus on developing sophisticated learning algorithms for an agent. We

can also observe this approach in the current reinforcement learning proposal for value alignment. Once the ethical rewards are specified and incorporated into the agent's environment, the focus is on designing algorithms that help an agent to learn an ethical behaviour following the ethical rewards. Recall that the agent already has its own objective. Thus, by incorporating a second (ethical) objective in the environment, we expand it into a *multi-objective* environment.

Despite the predominance of the agent-centric approach in reinforcement learning, it has several essential inconveniences. First of all, by leaving the *multi-objective* environment unmodified, we would need that the agent uses more complex multi-objective learning algorithms to take ethical rewards into account. Secondly, the agent-centric approach not only makes the learning problem of the agent more complex: it also assumes that the agent is willing to take ethical rewards into account, which might not always be the case. Thirdly, it assumes that it can control the learning process of the agent, which might not be the case for third-party agents.

For those three reasons, in this thesis, we propose to follow an *environment-designer* (or *environment-centric*) approach: to design the environment so that it guarantees that it is in the agent's best interest to learn to behave ethically. Such an approach solves the three previously discussed problems:

- By focusing on the environment, our purpose is to ease the learning problem of an agent. The agent only needs to learn an optimal policy in our designed ethical environment. In other words, the agent can use any reinforcement learning algorithm, including the most basic ones: single-objective tabular algorithms.
- The environment-designer approach allows us to be robust against third-party agents in which we cannot control their preferences or learning algorithms.
- In this thesis, we assume that the agent may have its individual objective, which may not be completely compatible with behaving ethically. By giving ethical incentives we guarantee that it is in the agent's best interest to behave ethically.

Our environment designer approach draws inspiration from both the *mechanism design* literature [Vlassis, 2009] and the *social dilemmas* literature [Kollock, 1998]. Mechanism design assumes we have a multi-agent system in which each agent has its preferences. These preferences cannot be modified, but instead, we can design a *mechanism* in the form of incentives to lead the agents to reach a socially desired outcome.

Structural solutions represent a very similar idea in the social dilemmas literature [Kollock, 1998]. A social dilemma represents a situation in which if each agent pursues its

interests uniquely, then the reached outcome of the system is far from being the best one. As it is popularly known: *individual rationality leads to collective irrationality.* Structural solutions allow the agents to escape the negative outcome by changing the rules of the environment, typically by adding new sanctions or rewards.

With our environment designer approach, we translate these two approaches to a reinforcement learning environment. We aim to strategically include ethical rewards or penalties for incentivising ethical behaviour. Furthermore, since we need that value-aligned is guaranteed, we also aim to provide a method that theoretically guarantees that the chosen rewards are enough for incentivising the agent and a method to find them.

1.2 Research questions

As previously stated, in this thesis, we propose to apply an environment-designer approach to the problem of guaranteeing value-aligned behaviour. Such an approach must yield theoretical guarantees to ensure the agent learns to behave ethically. This approach leads us to several research questions, which we summarise as follows:

1. Is the ethical environment design approach possible in reinforcement learning?
2. How do we automate the ethical environment design process?
3. Can we formally guarantee that our ethical environment design process always succeeds?
4. Can we test our ethical environment design process with a case study?

The remainder of this section is devoted to enumerating and further detailing each research question.

As mentioned in the previous section, state-of-the-art value-alignment approaches with reinforcement learning apply a two-step process. First, the reward specification, in which the ethical knowledge of a moral value is *specified* as rewards of a reinforcement learning environment. In other words, an agent following exclusively ethical rewards should learn a behaviour aligned with the moral value. The problem is that the agent already follows other rewards related to its original objective. For that, in the second step, the ethical embedding, rewards are *embedded* in the agent's environment so that it will also take them into account. If done without formal guarantees, we have no way of knowing if the agent will pursue an ethical behaviour or not in the end.

In this thesis, we are taking an environment-designer stance, and thus we assume that we cannot modify the agent's learning algorithm. Our objective is thus to *design* an ethical learning environment that guarantees that it is in the agent's best interest to learn a value-aligned behaviour. Recall that the learning environment of the agent is multi-objective: there is the individual objective of the agent and the ethical objective for a given moral value. Following our environment-designer approach, we cannot assume that the agent will consider the ethical objective while learning. We aim to guarantee that even if the agent disregards the ethical objective, it learns to behave ethically. The natural first question is: Is it possible to design an ethical environment? More specifically:

Question Q1: Given an environment with some (already specified) ethical rewards, can we design an ethical single-objective environment wherein the agent learns to behave ethically?

If the answer to the previous question is affirmative, we can advance our research and make further research questions. Then, if we can create an ethical environment, the next question is obvious: how? and is there a way to computationally automate the design of such ethical environments? Therefore, our next research question is developing an ethical environment design algorithm. The output of such an algorithm would be an environment wherein it is in an agent's best interest to behave ethically with respect to a moral value while pursuing its individual objective. We make this research question explicit:

Question Q2: Given a moral value and an agent's reinforcement learning environment, can we develop an algorithm for transforming the environment into an ethical environment wherein it is in the agent's best interest to behave ethically?

Recall that in Section 1.1.3, we explained that the process for designing an ethical environment could be divided into two steps: (i) reward specification and (ii) ethical embedding. Similarly, we divide Research Question **Q2** into two questions. Each question asks how to develop the two steps for designing an ethical environment, as shown in Figure 1.1.

Given the two-step process for designing ethical environments, this thesis starts by focusing on the second step: the ethical embedding process. Assuming that we have the ethical knowledge already formalised as ethical rewards, we can tackle it with multi-objective reinforcement learning. Then, the question becomes how to aggregate the different reward functions (individual and ethical) so that the agent learns to behave ethically:

Question Q2.1: Can we develop an algorithm for the ethical embedding process so that it designs an ethical environment wherein the agent learns to behave ethically?

With such an ethical embedding algorithm, we expect to obtain an ethical environment where the agent behaves in alignment with a moral value. To perform the ethical embedding process, we require the ethical rewards to be specified. In general, ethical rewards might not be available, and we must perform the reward specification process beforehand.

Back to the reward specification process, we can apply current formalisations of moral values in the literature (such as [Serramia et al., 2018]) as a starting point. From them, the next front to address is incorporating a moral value into a reinforcement learning environment. In other words, how can an environment be extended to include the ethical knowledge of moral values? This extension is limited to only affecting the rewards of the environment. Formally:

Question Q2.2: Can we develop an algorithm for the reward specification process to transform a moral value into an ethical reward function?

Answering Research Questions Q2.1 and Q2.2 would automatically answer Research Question Q2. The developed ethical environment design algorithm receives the agent's environment and a moral value as input and returns the ethical environment, as illustrated in Figure 1.1.

However, developing an algorithm is not enough. To solve the value alignment problem, we need to have formal guarantees that it is in the agent's best interest to behave value-aligned in the designed environment. Recall that because we are following an environment-designer approach, we need to consider that our only impact in the agent's learning process is through the ethical reward function we provide.

Question Q3: Is our proposed ethical environment design process formally guaranteed to create an environment wherein it is in the agent's best interest to behave ethically?

If the answer to Research Question Q3 is positive, it means that the ethical environments designed by our algorithm guarantee ethical-behaviour learning. Nevertheless, in this thesis we consider that it is also important to further validate our algorithm with a case study. That is, we would like to see the ethical environment design process applied to an actual environment and observe the policy that the agent learns in the designed ethical environment. For that, we require a multi-objective reinforcement learning environment in which an agent faces ethical problems.

However, not even the main multi-objective reinforcement learning library, MO-Gymnasium Alegre et al. [2022], contains an example environment with ethical problems. Due to the

lack of value-alignment reinforcement learning environments in the literature, we need to provide it ourselves:

Question Q4: Can we test the ethical environment design process in a reinforcement learning environment to validate that an agent learns a behaviour aligned with a moral value?

Next, Section 1.3 shows, as contributions, that it is possible to answer positively to all research questions.

1.3 Contributions

The current thesis addresses all the previously-presented research questions through this publication compendium. In short, we contribute to the state of the art with an ethical environment design algorithm that returns an environment wherein an agent is guaranteed to behave value-aligned. The general structure of our algorithm is shown in Figure 1.1. Next, we present our contributions, each addressing its research question. We divided this section into one subsection per chapter of this thesis. Each chapter contains a single paper, following the chronological order of publication.

1.3.1 Designing an ethical environment

The paper in Chapter 2 provides our first contribution. We design an ethical environment (*ad hoc*, without any algorithm) wherein the agents of a multi-agent game learn to behave in alignment with a moral value:

- [Rodriguez-Soto et al., 2020]:

  Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, "A Structural Solution to Sequential Moral Dilemmas" Proceedings of the Nineteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020, Core A* conference) pp. 1152-1160. Auckland (New Zealand), May 9-13, 2020.

The designed ethical environment corresponds to the output environment in Figure 1.1. As a preliminary step to developing an algorithm, in [Rodriguez-Soto et al., 2020], we hand-craft the ethical environment instead of following the two-step process of Figure 1.1. Here, our way of embedding the ethical rewards in the original environment is by directly aggregating them to the agents' original reward function. That is, given the

original reward function of each agent R_0, and the new ethical reward function R_v, we design an environment in which each agent has a scalarised reward function $R = R_0 + R_v$.

We show that agents learn to behave ethically after embedding the ethical rewards in the learning multi-agent environment. The agents never exhibit ethical behaviour when learning in the original environment (the input environment in Figure 1.1, without ethical rewards). In summary, the main contribution of [Rodriguez-Soto et al., 2020] is to answer Research Question **Q1** with:

Contribution C1: Design of an ethical environment for the agents of a multi-agent game in which the agents learn to behave in alignment with a moral value.

The environment used in [Rodriguez-Soto et al., 2020] is the Public Civility Game environment. [Rodriguez-Soto et al., 2020]. In this environment, two agents need to learn in alignment with the moral value of *civility*. We expect agents always to bring any piece of garbage to the nearest bin. Results show how in the designed ethical environment, agents behave civilly and put all the garbage in a bin. Afterwards, the agents go as fast as possible to their respective goal position.

1.3.2 The ethical embedding process

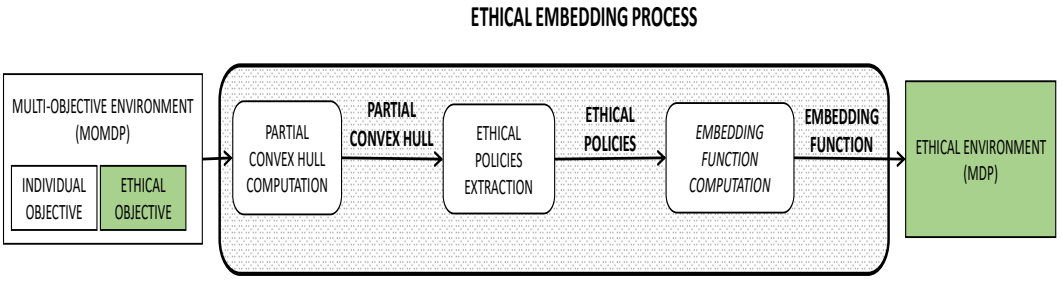


FIGURE 1.2: The ethical embedding design process for aligning the behaviour of the agent to an ethical reward function is performed in three steps: partial convex hull computation, ethical policies extraction, and embedding function computation. Rectangles stand for objects whereas rounded rectangles correspond to processes.

Due to the success in designing an ethical environment in Chapter 2, the paper in Chapter 3 starts building an algorithm for automating the design process of ethical environments. Chapter 3 provides an algorithm for the ethical embedding process, the second step of our ethical environment design process as shown in Figure 1.1. The presented ethical embedding algorithm takes as input a multi-objective environment with an ethical reward function already specified, as shown in Figure 1.2. From that input, it creates an environment wherein an agent using any reinforcement learning algorithm learns to behave in alignment with a single moral value. The contents of the chapter correspond to the following publication:

- [Rodriguez-Soto et al., 2021]:

  Manel Rodríguez-Soto, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Multi-Objective Reinforcement Learning for designing ethical environments", 30th International Joint Conference on Artificial Intelligence (IJCAI 2021 Core A* conference) pp. 545-551. Montreal (Canada), August 19-26, 2021.

Given the positive empirical results found in [Rodriguez-Soto et al., 2020], we proceeded in [Rodriguez-Soto et al., 2021] by proposing an algorithm to compute the desired ethical single-objective environment wherein an agent learns an ethical behaviour (i.e., aligned to a given moral value). This algorithm is built on top of a previous multi-objective reinforcement learning algorithm [Barrett and Narayanan, 2008]. It automates the aggregation of the two rewards of a Multi-Objective Markov Decision Process: the original individual reward function of the agent and the ethical reward function obtained in a reward specification process. In summary, we contribute to answering Research Question **Q2.1** with:

Contribution C2.1: Development of an embedding algorithm for designing an ethical environment that guarantees that it is in an agent's best interest to behave in alignment with a moral value.

The ethical embedding algorithm presented in [Rodriguez-Soto et al., 2021] designs an ethical environment in three steps, as illustrated in Figure 1.2. First, the algorithm receives as input a multi-objective environment with an original reward function R_0 and an ethical reward function specifying a moral value R_v. The objective of the algorithm is to design an ethical single-objective environment with a scalarised reward function. Hence, it has to find how much weight w to give to the ethical reward function concerning the individual one. We expect that, in the ethical environment, any policy that maximises the accumulation of scalarised rewards $R_0 + w \cdot R_v$ is also a policy that maximises the accumulation of ethical rewards R_v. Formally, we define an *ethical policy* (i..e, an ethical behaviour) as a policy that maximises the accumulation of ethical rewards.

To compute the desired ethical weight, the ethical embedding algorithm starts by computing the *partial convex hull* of the environment, which is a small subset of policies that will include at least one ethical policy. Thereafter, in the next step, the algorithm identifies an ethical policy to learn by finding the one that accumulates more ethical rewards. Thirdly, in the last step, the algorithm compares the value of the ethical policy with the rest of the policies of the convex hull. Comparing the ethical policy with this small subset is enough to find the desired ethical weighting, which we refer to as the *ethical embedding function*. Finally, the algorithm applies the embedding function to

the multi-objective environment to obtain the ethical environment. The output ethical environment is single-objective, with a reward function $R_0 + w \cdot R_e$.

1.3.3 The ethical environment design process and its formal guarantees

In the previous section, we discussed how Chapter 3 provides an algorithm for one of the two steps of the ethical environment design process, the ethical embedding. The paper in the following Chapter 4 completes the ethical environment design process by providing an algorithm for the remaining step: the reward specification. Thereafter, Chapter 4 presents a complete ethical environment design algorithm that implements the two steps in Figure 1.1. Furthermore, Chapter 4 provides the necessary theoretical results for guaranteeing that indeed in the obtained ethical environment, it is optimal for an agent to learn to behave in alignment with a moral value. The contents of this chapter are published in:

- [Rodriguez-Soto et al., 2022]:

  Manel Rodríguez-Soto, Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Instilling moral value alignment by means of multi-objective reinforcement learning", Ethics and Information Technology journal. Ed. Springer (ISSN 1388-1957, 2020 IF JCR: 4.449, Q1 Cat. Philosophy (19/317)). Vol 24:9.pp 1-17. 24 January 2022.

Developing an algorithm for automating the reward specification process is the first objective of [Rodriguez-Soto et al., 2022]. The paper starts by providing philosophical foundations for our chosen definition of moral value. After studying the Ethics literature, we consider moral values as principles for discerning right and wrong actions. Moreover, we argue that Ethics treats actions in two separate but related dimensions. First, there is a normative dimension that states whether an action is morally *obligatory*, *prohibited*, or *permitted*. Second, there is an evaluative dimension that states whether an action is morally *good*, *bad*, or *neutral*. Typically any action bad to do is morally prohibited as well. However, some actions that are good to do may be morally obligatory, while others are just morally permitted [Chisholm, 1963, Urmson, 1958]. Therefore, any formal definition of moral values needs actions to be considered from these two ethical dimensions. Thus, we propose to define a moral value as a tuple of two elements, following [Serramia et al., 2018]. The first element of the tuple is a set of norms classifying actions as permitted, prohibited, or obligatory. The second element of the tuple, the evaluation function, states numerically how good or bad the actions are.

With the proposed definition of a moral value in Chapter 3 the next step consists in including it in a reinforcement learning environment (Markov Decision Process). For that reason, [Rodriguez-Soto et al., 2022] provides a procedure for obtaining an exogenous reward function associated with a moral value that can be applied to any reinforcement learning environment. Such reward function has two components, the counterparts of those of the moral value: a *normative* reward component penalising prohibited actions, and an *evaluative* reward component rewarding morally good actions. In this way, we encapsulate all the ethical knowledge of the moral value as rewards. Formally, we answer Research Question Q3 with:

Contribution C2.2: Development of an ethical reward specification process for transforming the ethical knowledge of a moral value into an ethical reward function of a reinforcement learning environment.

The ethical reward specification process presented in [Rodriguez-Soto et al., 2022] designs an ethical reward function in two steps, which can be computed in any order. They transform each component of the moral value into its corresponding reward component. Regarding the normative component, each prohibition norm of the norm set is transformed into a negative reward of the normative reward function R_N to punish its corresponding action. Then, for the evaluative component, each action the evaluation function considers praiseworthy is transformed into a positive reward of the evaluative reward function R_E to incentivise the desired behaviour. Finally, since both components are considered equally important, they are aggregated to obtain the ethical reward function $R_v = R_N + R_E$.

In [Rodriguez-Soto et al., 2022], we argue that, given a moral value, to behave aligned with it means:

- First, for the normative component, an ethical behaviour needs to follow all the norms within the value: refrain from doing any morally prohibited action, and perform all morally obligatory actions. Therefore, in reinforcement learning terms, an ethical behaviour (ethical *policy*) needs to maximise the accumulation of normative ethical rewards.
- Second, for the evaluative component, an ethical behaviour needs to act in the most praiseworthy way possible: performing the most morally good actions. Therefore, in reinforcement learning terms, an ethical behaviour (ethical *policy*) needs to maximise the accumulation of evaluative ethical rewards.

Thus, we argue that an ethical behaviour (i.e., a behaviour aligned with a given moral value) is naturally defined as the one that maximises the accumulation of normative and

evaluative ethical rewards. For simplicity, we say that an ethical behaviour maximises the accumulation of ethical rewards. This definition leads us to the following conclusion: following our approach, behaving in alignment with a moral value is the same as maximising the accumulation of ethical rewards in a reinforcement learning environment.

Moreover, because [Rodriguez-Soto et al., 2022] provides an ethical reward specification process, combined with the ethical embedding algorithm from [Rodriguez-Soto et al., 2021] we obtain a complete two-step process for designing ethical environments. This algorithm answers Research Question Q2:

Contribution C2: Given a moral value and an agent's reinforcement learning environment, we provide a two-step algorithm for transforming the environment into an ethical environment that incentivises value-aligned behaviour.

Such an algorithm applies our two previous processes: the reward specification process and the ethical embedding process. Furthermore, in[Rodriguez-Soto et al., 2022], we provide formal guarantees for the proposed ethical environment design algorithm. Its main result is Theorem 2, which states that optimal policies are ethical in the designed ethical environment. That is, the designed ethical rewards incentivise the agent to perform value-aligned behaviours.

Theorem 2 of [Rodriguez-Soto et al., 2022] requires a small formal assumption to hold in a given environment. It is only possible to design an ethical environment if it is possible to follow an ethical policy in the original environment. This condition follows this simple logic: if we expect an agent to behave ethically, it should be possible for it to behave ethically. Formally, Theorem 2 leads to the following contribution:

Contribution C3: Formal proofs of our ethical environment design algorithm. Given the condition that behaving ethically is possible, our algorithm is guaranteed to yield an ethical environment wherein it is in an agent's best interest to behave in alignment with a moral value.

1.3.4 Evaluating the ethical environment design process

After developing an ethical environment design algorithm with theoretical guarantees in the previous chapter, there is only one remaining research question: to test the algorithm. The paper in Chapter 5 answers this question by empirically evaluating the ethical environment design algorithm in a case study. In that way, we illustrate how to use it and the resulting behaviour of the agent. The contents of this chapter are published in:

- [Roselló-Marín et al., 2022]:

  Eric Roselló-Marín, Maite Lopez-Sanchez, Inmaculada Rodríguez, Manel Rodríguez-Soto and Juan A. Rodríguez-Aguilar, "An Ethical Conversational Agent to Respectfully Conduct In-Game Surveys", Frontiers in Artificial Intelligence and Applications: Artificial Intelligence Research and Development, IOS Press Vol 356, pp. 335-344. October 2022.

In more detail, [Roselló-Marín et al., 2022] provides an empirical evaluation of our ethical environment design algorithm for the following problem: a conversational agent that needs to extract information from a human user while abiding by the moral value of respect. This environment represents a case study with potential applications in the real world.

In this case study, a user plays a video game while the conversational agent surveys the user about the video game to elicit as much feedback information as possible. Results showcase how a conversational agent that has learnt to behave in alignment with the value of respect manages to avoid disturbing a user's engagement. In summary, we answer Research Question **Q6.1** with:

Contribution C4: Validation in a case study of the ethical environment design algorithm. The case study is the learning environment of a conversational agent that needs to learn to behave in alignment with the moral value of respect. The evaluation of the ethical environment design algorithm applied to the conversational agent's environment shows that the agent abides by the moral value of respect in the designed ethical environment.

Chapter 5 provides the last contribution of this thesis, which answers the last research question. Hence, the papers presented in this thesis answer our four research questions.

1.4 Dissertation outline

Following the concepts introduced in this chapter, the rest of this thesis is structured as follows.

- **Chapter 2** presents our approach for designing ethical environments that motivates the main results in this thesis. In this chapter, we argue that an *environment-designer* approach incentivises agents to behave ethically by rewarding ethical behaviour. We empirically show in a multi-agent game how agents behave value-aligned in the designed ethical environment. The example problem used here is

called *the Public Civility Game*. In it, we expect that two agents learn to behave in alignment with the ethical objective of civility.

- **Chapter 3** sets the first results towards guaranteeing ethical behaviour learning in reinforcement learning environments. It introduces our first algorithm which designs environments in which it is in the agent's best interest to learn to behave aligned with a given moral value. The provided algorithm automates the ethical embedding process, the second step of our ethical environment design approach, as shown in Figure 1.1.

- **Chapter 4** proceeds by completing our ethical environment design algorithm. It provides a procedure for automating the ethical reward specification process which, when added to the previous algorithm, creates a complete two-step ethical environment design process. Furthermore, it provides theoretical guarantees of which environments the algorithm is guaranteed to succeed in incentivising ethical behaviour learning.

- **Chapter 5** provides an example application of the ethical environment design algorithm. Here, the objective is to guarantee that a conversational agent learns to behave following the moral value of respect. Our empirical results illustrate that the agent behaves in alignment with the moral value of respect in the designed ethical environment.

- **Chapter 6** is devoted to enumerating the main contributions from this thesis, discussing the conclusions extracted from them, and providing research directions for future work.

1.5 Publications derived from this thesis

This section provides a list of all the publications disseminated during this thesis. Publications are divided by category (journal papers, conference papers, and workshop papers), and they are presented in chronological order within each category. The last section also presents papers currently under review (as of June 2023).

This thesis is a compendium of all the published papers in a journal or proceedings. Thus, the content of workshop papers was excluded, and we only mention them here, even though they were also peer-reviewed.

1.5.1 Journal papers

- Manel Rodríguez-Soto, Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Instilling moral value alignment by means of multi-objective reinforcement learning", Ethics and Information Technology journal. Ed. Springer (ISSN 1388-1957, 2020 IF JCR: 3.633, **Q1 Cat. Ethics**). Vol 24:9.pp 1-17. 24 January 2022.

1.5.2 Conference proceedings

- Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, "A Structural Solution to Sequential Moral Dilemmas" Proceedings of the Nineteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020, **Core A* conference**) pp. 1152-1160. Auckland (New Zealand), May 9-13, 2020.
- Manel Rodríguez-Soto, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Multi-Objective Reinforcement Learning for designing ethical environments", 30th International Joint Conference on Artificial Intelligence (IJCAI 2021 **Core A* conference**) pp. 545-551. Montreal (Canada), August 19-26, 2021.
- Eric Roselló-Marín, Maite Lopez-Sanchez, Inmaculada Rodríguez, Manel Rodríguez-Soto and Juan A. Rodríguez-Aguilar, "An Ethical Conversational Agent to Respectfully Conduct In-Game Surveys", Frontiers in Artificial Intelligence and Applications: Artificial Intelligence Research and Development, IOS Press Vol 356, pp. 335-344. October 2022.

1.5.3 Workshop papers without proceedings

Workshop publications were not included in the current thesis due to not having proceedings (even though they were peer-reviewed), but we cite them here for completeness.

- Manel Rodríguez-Soto, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning", ALA 2021 Adaptive and Learning Agents Workshop at AAMAS, London (United Kingdom). pp 1-9. May 3rd-4th, 2021.
- Manel Rodríguez-Soto, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, "Building Multi-Agent Environments with Theoretical Guarantees on the Learning of Ethical Policies", ALA 2022 Adaptive and Learning Agents Workshop at AAMAS, Auckland (New Zealand). pp 1-9. May 9th-12th, 2022.

- Manel Rodríguez-Soto, Roxana Radulescu, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, Ann Nowe, "Multi-objective reinforcement learning for guaranteeing alignment with multiple values", ALA 2023 Adaptive and Learning Agents Workshop at AAMAS, London (United Kingdom). pp 1-9. May 29th-June 3rd, 2023.

1.5.4 Publications under review

The publication under review was not included in the current thesis, but we cite it here for completeness.

- **(Under review)** Manel Rodríguez-Soto, Maite Lopez-Sanchez, Juan A. Rodríguez-Aguilar, 'Multi-Objective Reinforcement Learning for Designing Ethical Multi-Agent Environments", Neural Computing and Applications. Ed. Springer.

Chapter 2

Designing an ethical environment

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

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

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:*

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

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 $P_n^i : \mathcal{S} \times \mathcal{A}^i \to \{0, p\}$ of every agent i as*

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

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 $R_0^{i=1,\dots,m}$, we define its normative extension as another Markov game such that the reward function R^i for each agent i is defined as $R^i = R_0^i + R_{\mathcal{N}}^i$, where $R_{\mathcal{N}}^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^-$ corresponds to the normative reward function and is defined as*

$$R_{\mathcal{N}}^i(s, a^i) \doteq -\sum_{n \in \mathcal{N}} P_n^i(s, a^i). \tag{5}$$

The normative reward function $R_{\mathcal{N}}^i$ 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 $R_0^{i=1,\dots,m}$, we define its ethical extension as another Markov game such that the reward function R^i of each agent i is defined as $R^i = R_0^i + R_{\mathcal{N}}^i + R_E^i$, where $R_{\mathcal{N}}^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^-$ is the normative reward function of norm set $\mathcal{N}$ applied to agent i and $R_E^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}^+$ is is a function of the form*

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

We will refer to R_E^i 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_E^i 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., $\Pi_v \neq \emptyset$).*

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

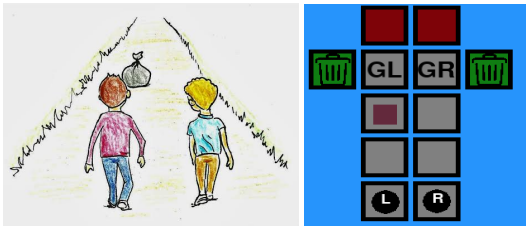


Figure 1: Left: garbage blocking the path of the agent at the left. Right: Our simulation representing the same state.

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 L \ Agent R	E	U
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 $V^i_\pi(s_0)$ (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 $R^i = R^i_0 + R^i_{\mathcal{N}} + R^i_E$ following Definition 8.

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

$$R^i_{\mathcal{N}}(s, a^i) = -P^i_{n_1}(s, a^i), \tag{17}$$

and following Eq. 6, the evaluative reward function R^i_E becomes:

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

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 $\mathtt{p}_1 = 10$ for not complying with norm n_1 (see equation 2) and a reward of 10 for behaving civilly $\mathtt{eval}_{\mathtt{civ}} = 10$ 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*.

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

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

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.

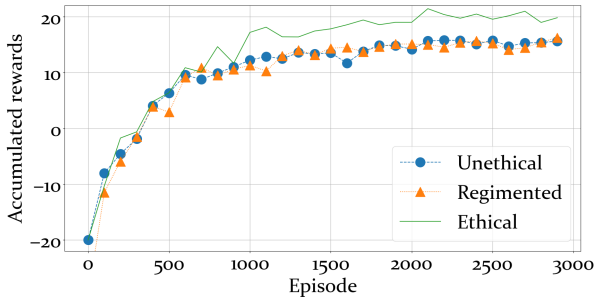


Figure 2: Evolution of the accumulated rewards per episode in the three scenarios: unethical, regimented, and ethical.

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].

[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.

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.

[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*.

Chapter 3

The ethical embedding process

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)

Multi-Objective Reinforcement Learning for Designing Ethical Environments

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

Abstract

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

1 Introduction

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

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

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

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

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

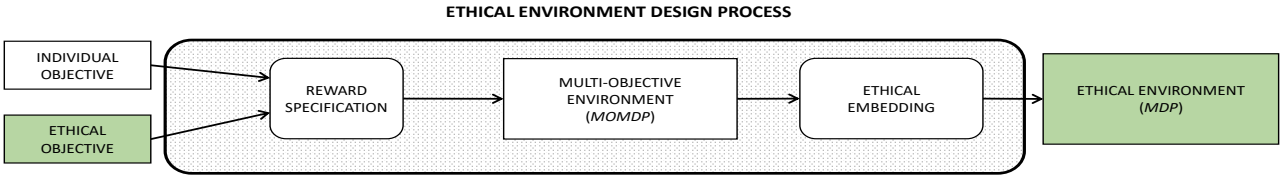


Figure 1: The process of designing an ethical environment is performed in two steps: a reward specification and an ethical embedding. Our algorithm computes the latter. Rectangles stand for objects whereas rounded rectangles correspond to processes.

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

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

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

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

2 Formalising the Ethical Embedding Problem

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

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

Definition 1. *A (finite)*[1] *n-objective Markov Decision Process (MOMDP) is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \vec{R}, T \rangle$ where $\mathcal{S}$ is a (finite) set of states, $\mathcal{A}(s)$ is the set of actions available at state s, $\vec{R} = (R_1, \ldots, R_n)$ is a* vectorial *reward function with each R_i as the associated scalar reward function to objective $i \in \{1, \ldots, n\}$, T is a transition function. Each MOMDP has its associated multi-dimensional state value function $\vec{V} = (V_1, \ldots, V_n)$ in which each V_i is the expectation of the obtained sum of i-objective rewards.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where Π_e is the set of ethical policies.

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

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

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

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

Problem 1 (Ethical embedding)**.** *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, (R_0, R_{\mathcal{N}}, R_E), T \rangle$ be an ethical MOMDP. Compute a weight vector $\vec{w}$ with positive weights such that all optimal policies in the MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, w_0 R_0 + w_N R_{\mathcal{N}} + w_E R_E, T \rangle$ are also ethical-optimal in $\mathcal{M}$ (as defined in Def. 4).*

Any weight vector $\vec{w}$ with positive weights that guarantees that all optimal policies (with respect to $\vec{w}$) are also ethical-optimal is a solution of Problem 1. The next section proves that such solutions always exist for any ethical MOMDP.

3 Solvability of the Ethical Embedding Problem

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

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

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

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

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

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

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

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

4 Solving the Ethical Embedding Problem

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

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

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

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

4.1 Computation of the Partial Convex Hull

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

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

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

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

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

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

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

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

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

4.2 Extraction of the Ethical-optimal Policies

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

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

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

4.3 Computation of the Embedding Function

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

Algorithm 1 Ethical Embedding

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

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

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

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

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

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

4.4 An Algorithm for Designing Ethical Environments

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

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

5 Example: The Public Civility Game

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

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

5.1 Reward Specification

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

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

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

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

5.2 Ethical Embedding

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

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

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

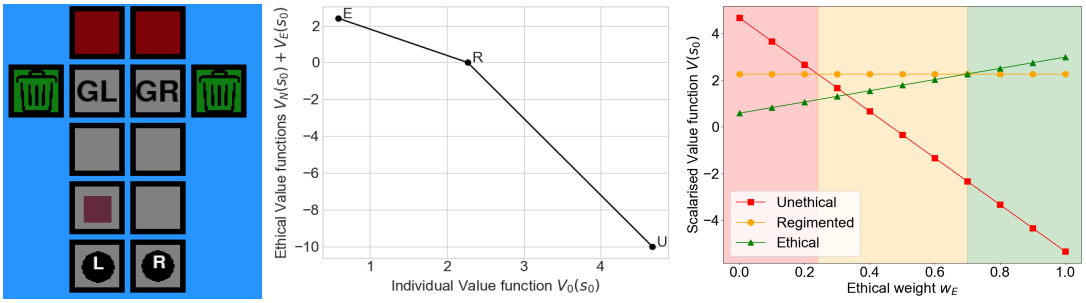


Figure 2: Left: Initial state of the public civility game. The agent on the left has to deal with the garbage obstacle, which has been located in front of it. Centre: Visualisation in Objective Space of the partial convex hull of $\mathcal{M}$ composed by 3 policies: E (Ethical), R (Regimented) and U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

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

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

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

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

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

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

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

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

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

6 Conclusions and Future Work

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

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

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 4

The ethical environment design process and its formal guarantees

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

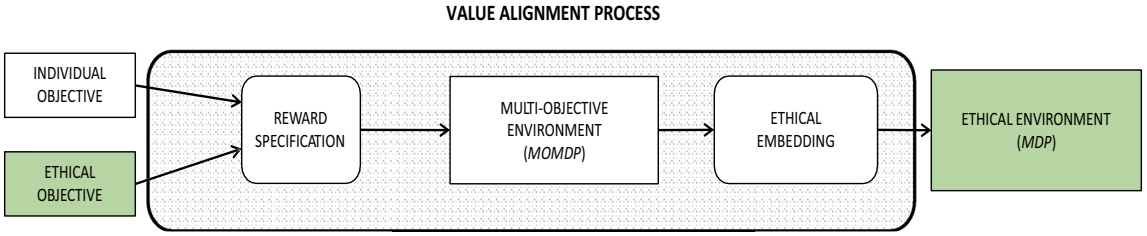


Fig. 1 The value alignment process is performed in two steps: a reward specification and an ethical embedding. Rectangles stand for objects whereas rounded rectangles correspond to processes

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 $R_1, \ldots, R_n$.

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.
- $E_v : \mathcal{A} \to [-1, 1]$ 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 $P_n : \mathcal{S} \times \mathcal{A} \to \{-1, 0\}$ 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 $R_\mathcal{N} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^-$, 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 $R_E : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$, 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. □

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

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

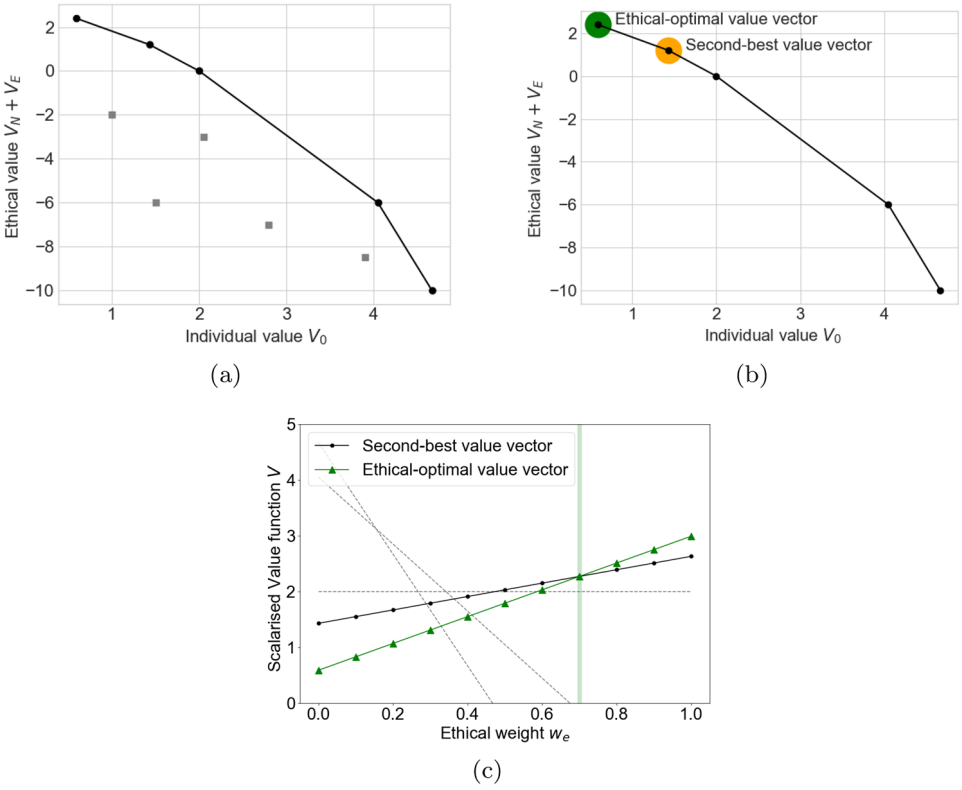


Fig. 3 **a** Example of convex hull $CH(\mathcal{M})$, represented in objective space. **b** Identification of the points of $CH(\mathcal{M})$ corresponding with the ethical-optimal value vector $\mathbf{V}^*$ (highlighted in green) and the second-best value vector $\mathbf{V'}^*$ (in yellow). **c** 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

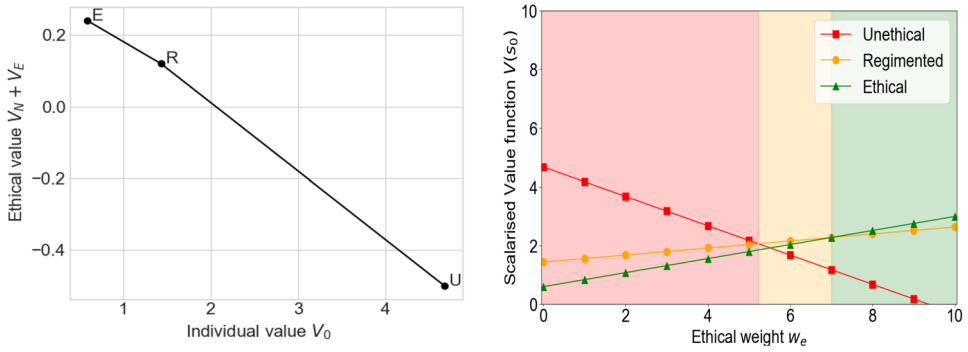


Fig. 4 Left: Visualisation in Objective Space of the convex hull of the public civility game composed by 3 policies: E (Ethical), R (Regimented) and U (Unethical). Right: Visualisation in Weight Space of the same convex hull. The painted areas indicate which policy is optimal for the varying values of 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 $\mathbf{V}^\pi = (V_0^\pi, V_N^\pi + V_E^\pi)$. 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 $\epsilon > 0$):

$$0.59 + (7 + \epsilon) \cdot (0 + 0.24) = 2.27 + 0.24\epsilon > 1.43 + 7 \cdot (0 + 0.12) = 2.27.$$

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.

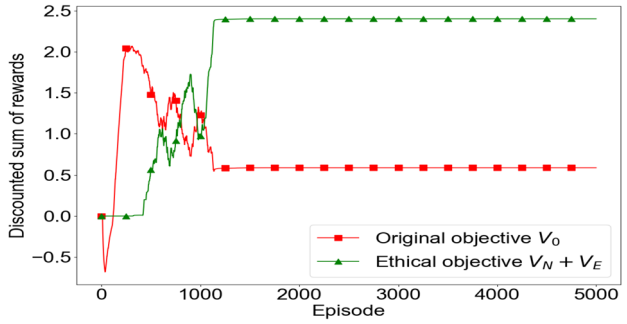


Fig. 5 Evolution of the accumulated rewards per episode that the agent obtains in the ethical environment

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 $R_0 + 7.1(R_N + R_E)$.

Analysis: learning in an ethical environment

After creating the ethical environment $\mathcal{M}'$ with reward function $R_0 + 0.71(R_N + R_E)$ 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 $\mathbf{V}$ stabilises, with less than 1500 episodes, at 0.59 (V_0 line) and 2.4 ($V_N + V_E$ 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 common-sense 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.

Chapter 5

Evaluating the ethical environment design process

Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220356

An Ethical Conversational Agent to Respectfully Conduct In-Game Surveys

Eric ROSELLÓ-MARÍN [a], Maite LOPEZ-SANCHEZ [a,1], Inmaculada RODRÍGUEZ [a], Manel RODRÍGUEZ-SOTO [b] and Juan A. RODRÍGUEZ-AGUILAR [b]
[a] *Department de Matemàtiques i Informàtica, Universitat de Barcelona (UB)*
[b] *Artificial Intelligence Research Institute (IIIA-CSIC)*

Abstract. The improvement of videogames highly relies on feedback, usually gathered through UX questionnaires performed after playing. However, users may not remember all the details. This paper proposes an ethical conversational agent, endowed with the moral value of respect, that interacts with the user to perform a survey during the game session. To do so, we use reinforcement learning and the ethical embedding algorithm to ensure that the agent learns to be respectful (i.e., avoid gameplay interruptions) while pursuing its individual objective of asking questions. The novelty is twofold: firstly, the application of ethical embedding outside toy problems; and secondly, the enrichment of a survey oriented conversational agent with this moral value of respect. Results showcase how our ethical conversational bot manages to avoid disturbing user's engagement while getting even a higher percentage of valid answers than a non-ethically enriched chatbot.

Keywords. Machine ethics, Reinforcement Learning, Conversational Agents, User Experience Questionnaires, Video Games

1. Introduction

Human Computer Interaction (HCI) and User eXperience design are fast evolving fields that pursue to improve the design of interactive systems [11]. In the context of UX empirical studies, questionnaires [13] have proven to be useful tools for assessing the user experience of using any computer application, and video games and virtual reality experiences are no exception. Thus, game designers resort to playtesting, which usually is conducted by first letting users play the game, and afterwards, once the playing session has concluded, asking questions about their playing experience [12].

However, users may not remember all details by the end of the experience and, if the number of questions is large, they may lead to user boredom or even user fatigue [25], which hinders the quality of the gathered feedback. Moreover, this disadvantage is aggravated when transitioning back to reality to perform a survey about a Virtual Reality (VR) experience, which can lead to systematic bias as the user is no longer immersed in the virtual world [1].

[1]Corresponding Author: maite_lopez@ub.edu. Funded by CI-SUSTAIN (Grant PID2019-104156GB-I00), Crowd4SDG (H2020-872944), COREDEM (H2020-785907), Barcelona City Council through the Fundació Solidaritat UB (code 21S01802-001). Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387).

Conversational agents –interactive systems (embodied or not) that engage in conversation with the user [8]–, offer a new way to collect information, allowing to substitute a traditional survey with an agent that prompts the questions to the user. Indeed, conversational agents have shown to be effective for this task, as they increase both user's commitment with the survey and the quality of the information elicited [10] .

Against this background, we propose to introduce a conversational agent that conducts the survey in-game, as part of the game experience, with the aim of avoiding the detrimental effects of post-game questionnaires, and to ease participation by allowing to stay closer to the context of an ongoing exposure [17]. Nevertheless, this has also the risk of disturbing the game flow [24] if the chatbot does not properly identify when to prompt the user, or even result in the abandonment of the interview due to the player's cognitive overload [10]. Therefore, we argue that the conversational agent should be respectful with the user's engagement, and thus, we propose to embed the chatbot with a moral value of *respect*, which should guide the agent to perform the questionnaire without disturbing the user experience.

As social interactions must be considered when designing artificial agents [5], it is becoming apparent that agents' behaviour should align to human values [2]. Unfortunately, although machine ethics [27,28] is an active research area, very little literature is found on alignment of ethical principles in conversational agents. Some discussions highlighted the need to furnish conversational agents with ethical awareness [7]. However, inducing an ethical behaviour requires some learning, since identifying at design time all situations where this may be required constitutes a complex task.

Our proposal ensures the conversational agent learns to behave ethically by applying ethical embedding, a reinforcement learning approach (see e.g., [18]). This methodology for instilling moral value alignment is founded in the framework of Multi-Objective Reinforcement Learning [20] and the philosophical consideration of values [3] as ethical principles that discern good from bad, and express what ought to be promoted. Examples of human values[2] include fairness, respect, freedom, security, or prosperity [9].

In particular, our proposal redesigns the conversational agent's learning environment so that it is ensured that the agent learns to pursue its individual objective of asking as many questions as possible while fulfilling the ethical objective of being respectful with the user's engagement. This advances the state of the art as it showcases the application of the ethical embedding method beyond toy problems and enriches current survey oriented conversational agents with this moral value of *respect*.

2. Problem Formulation and Scenario

Intuitively, our problem is that of designing an ethical conversational agent that performs in-game surveys. Briefly, we tackle this problem by transforming the learning environment of this agent so that it is guaranteed that the agent learns to be respectful with a user playing the game while eliciting as much player feedback as possible. The learning environment for the conversational agent is a (Multi-Objective) Markov Decision Process (see Subsection 3.1) specified based on the game being played, which in this case is a Pong game played by a simulated user. In this context, we understand respect as not

[2]Sociology and Psychology have also extensively studied human values, which are often defined as abstract ideals that guide people's behaviour [23].

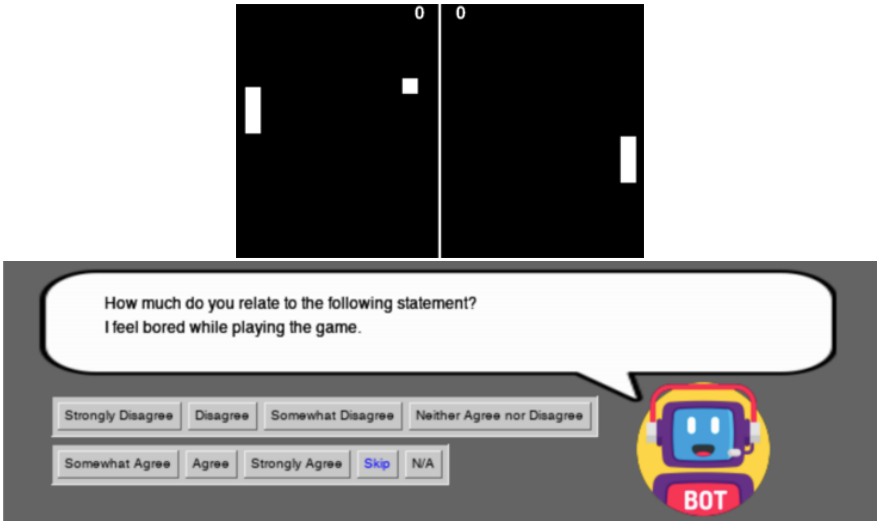


Figure 1. Screenshot of our Pong game illustrating an in-game period in which the chatbot is asking a question (resources from Flaticon, by Freepik). Skip and N/A response options are considered non-valid answers.

hindering the user engagement. In what follows, we introduce engagement and all other necessary elements that characterise our problem scenario.

2.1. Engagement

Within Human-Computer Interaction, engagement is a multi-stage process that becomes key to adapt the designs to the user [16]. The different stages of engagement can be distinguished by different levels of intensity of attributes [15] which, in video games mainly correspond to challenge, aesthetic, feedback, novelty and interactivity.

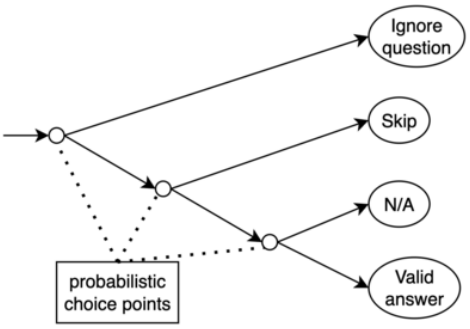
We can distinguish five different engagement stages. First, the *point of engagement*, is the stage where the user's attention is captured. Next, the *period of engagement* lasts while the attention and interest is maintained through feedback, novelty or challenge. Then, *disengagement* can be followed by the stage of *re-engagement*, which closes the cycle, or *nonengagement,* if the user engagement comes to an end.

In general, as game sessions consist on multiple engagement cycles of varying intensity, we require the survey conversational agent to behave respectful with the user by avoiding interrupting the user engagement, that is, just asking questions when the intensity of the engagement attributes is low.

2.2. Interaction with the User

For the sake of simplicity, we have chosen a single-player three-level Pong game. Levels in this game feature table-tennis games and are interleaved with several transition menus greeting the user or showing the score at the end of each level. Figure 1 depicts an in-game period, where the player uses keyboard arrow keys to move vertically the paddle and hit the bouncing ball. These in-game periods will be the ones typically having high user engagement, as they challenge the users and require from them higher interactivity than menus.

As Figure 1 shows, the conversational agent remains visible at the bottom of the screen throughout the whole game experience, and can prompt questions to the user at any time. Questions are taken from a short version of the Game User Experience



(a)

	Ignore question	Skip	N/A
Starting menu	0%	60%	100%
In-game	10%	30%	$40\% * 0.5^{level}$
Other menus	0%	5%	$30\% * 0.5^{level}$

(b)

Figure 2. Model of our simulated user, illustrating (a) the rule tree that dictates behaviour and (b) the threshold values of the probabilistic choice points for different in-game or in menu situations.

Satisfaction Scale (GUESS), the GUESS-18, which was designed to be used in iterative game design, testing, and research [12]. Our chatbot asks questions from a pool of 12 questions about enjoyment (see Figure 1), usability/playability, visual aesthetics, etc., discarding those about narrative, audio and social connectivity that do not apply to Pong.

The user can answer any of these questions by selecting the corresponding button in the user interface (see Fig 1). We distinguish two types of answers: *valid* and *non-valid*. Valid answers belong to the Likert scale used in GUESS-18 and are the ones the chatbot should gather to elicit useful data about the user's game experience. Non-valid answers correspond to "Skip" and "N/A": *Skip* denotes the user is not willing to answer a specific question, and thus it is discarded from the pool before being answered; and *N/A* (as in Not Available), indicates the user does not know the answer to the question yet, and should be asked at a later time, so the chatbot still has the chance to get a valid answer later.

Moreover, notice that the player also has the option of ignoring the survey question by simply continue playing. This leaves the chatbot waiting for an answer without being able to pose more questions and without requiring any particular action from the user.

2.3. Simulated user

As previously mentioned, we propose our survey conversational agent to learn to be respectful with the user by applying Reinforcement Learning (RL) [26] methods. However, RL constitutes a data-hungry approach, requiring numerous episodes to learn a policy, and human trials are expensive and time-consuming. Therefore, the repeatability and the acquisition of participants pose a serious challenge [6]. In this context, automatic user simulation tools [21] have been proposed as a handy alternative [14] for the first stages of agents' training, as they provide flexibility and repeatability [21]. Alternative simulators have been proposed based on probabilistic, heuristic, or stochastic models (or a combination of them) [6].

Following heuristic approaches [6] implemented by means of hierarchical patterns (such as HAMs) and rule sets, we have built a simulated user that reproduces human interactions by applying the rule tree in Figure 2a. Non-terminal nodes in the binary tree represent probabilistic *choice points* [22], and terminal nodes indicate the action to be taken. Whenever the chatbot asks a question, the simulated user traverses the tree to decide its reaction. Thus, the probabilities associated to choice point nodes, which

are shown in Figure 2b, allow the random selection of the outgoing edge (i.e., children) to follow. These probabilities vary if the user is playing or not (i.e., in-game or in a menu). We consider the user is collaborative and thus, it never ignores questions while being in a menu (i.e., the "Ignore question" branch in Figure 2 has 0% probability of being selected by the simulated user in Starting menu and Other menus) and just does it 10% in-game (which means it will select any other branch 90% of the times). Overall, we set the probabilities in Figure 2b so that the simulated user will be more likely to provide non-valid answers in-game (i.e., while playing) and in the starting menu than in subsequent menus. Moreover, the further the player gets in the game, the less chances of providing *N/A* answers. We include these probabilities in order to allow a degree of *lifelike* randomness in the behaviour [14].

3. Background

As previously introduced, we study how a conversational agent can learn to be respectful to the user while performing in-game surveys. The agent's environment is initially specified as a Multi-Objective Markov Decision Process, which in our approach we transform into a (single-objective) Markov Decision Process. This simplification of the environment is due to the fact that it is simpler for the agent to learn in a single-objective MDP, and thus, it is here where the agent learns its behaviour. Furthermore, we create such single-objective environment in a way that guarantees that the agent will learn a value-aligned behaviour (i.e., policy). This section is devoted to provide the necessary background to introduce our approach.

3.1. Markov Decision Process and Multi-Objective Markov Decision Process

In the context of Reinforcement Learning [26], the learning environment is characterised differently depending on the number of the agent's learning objectives:

Definition 1. *A (single-objective) Markov Decision Process (MDP) is defined as a tuple* $\langle S, A, R, T \rangle$ *where S is a set of environment states,* $A(s)$ *is the set of agent actions available at state s,* $R(s,a,s')$ *is a reward function specifying the reward the agent receives for performing action a at state s when the next state is* s'*, and* $T(s,a,s')$ *is the function specifying the probability of such transition.*

Definition 2. *An n-objective Markov Decision Process (MOMDP) is defined as a tuple* $\langle S, A, \vec{R}, T \rangle$ *where S, A and T are as in an MDP, and* $\vec{R} = (R_1, \ldots, R_n)$ *is a* vectorial *reward function composed of n scalar reward functions* R_i*, one per objective i.*

The agent's behaviour in an (MO)MDP is then described by a policy π, which indicates for each state-action pair $\langle s, a \rangle$, the probability of performing action a in state s. Moreover, a value vector $\overrightarrow{V}$ evaluates a policy π by computing the expected discounted sum of rewards obtained when following it:

$$\overrightarrow{V}^{\pi}(s) \doteq \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k \overrightarrow{r}_{t+k+1} | S_t = s, \pi] \text{ for every state } s \in S, \tag{1}$$

where $\gamma \in [0,1)$ is the discount factor and t is the time-step of each state s. An *optimal policy* in a single-objective MDP is, then, one that maximises the expected discounted reward accumulation for every state ($\pi_* \doteq arg\ max_{\pi} V^{\pi}$). π_* constitutes the behaviour the

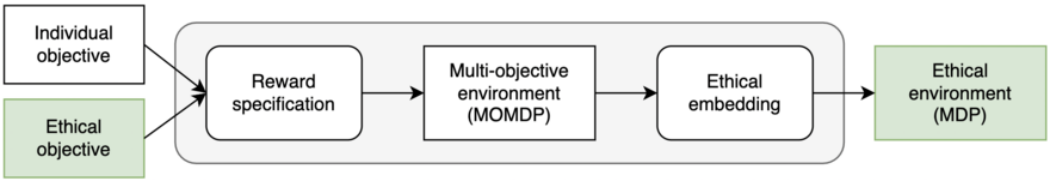


Figure 3. The ethical environment design process (as in [19]) for value alignment.

agent should learn, or, in other words, the solution to the MDP. Its computation is more complex for an MOMDP though, as it involves the optimisation of the value vector $\vec{V}^*$ instead of a single V^* value function.

3.2. Value Alignment

MOMDPs facilitate learning value-aligned behaviours, as they can be used to design the environment to incentivize ethical behaviour. Following the approach in [19], Figure 3 illustrates value alignment as a process consisting of two steps: *reward specification* and *ethical embedding*.

Firstly, the reward specification defines an MOMDP by considering both the individual objective (the agent's original objective translated into individual reward R_0) and the ethical objective (the moral value we introduce). This ethical objective encodes the moral value into rewards and is composed of two dimensions: the *normative* reward function R_N, which punishes the violation of normative moral requirements; and the *evaluative* reward function R_E, which rewards morally praiseworthy actions. In this context, we follow [19] and consider an *ethical policy* as one that abides to all norms while behaving as praiseworthy as possible, and an *ethical-optimal policy* as one that maximizes the individual objective as much as possible subject to being ethical. Formally, we refer to this value-enriched MOMDP as an *ethical* MOMDP, and define it as $\langle S, A, (R_0, R_N + R_E), T \rangle$.

Secondly, Figure 3 (right) depicts how the ethical embedding process transforms this ethical MOMDP into a single-objective MDP, where the agent is incentivized to learn an *ethical-optimal policy*. That is, the resulting MDP guarantees that the agent learns to fulfil the ethical objective while pursuing its individual objective (and, as it is single-objective, just requires the agent to apply a basic reinforcement learning method).

The ethical embedding process applies this transformation by computing a linear scalarisation function over the vectorial rewards $\vec{R}$ in the MOMDP that results in a scalar reward function R for an ethical MDP. This function has the form of:

$$f(\vec{V}^{\pi}) = \vec{w} \cdot \vec{V}^{\pi} = w_0 V_0^{\pi} + w_e (V_N^{\pi} + V_E^{\pi}) \tag{2}$$

Following [19], we fix the individual weight $w_0 = 1$ so that the ethical embedding process is reduced to looking for the ethical weight $w_e > 0$ that guarantees the learned behaviour in the resulting ethical MDP $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$ will prioritise the ethical objective over the individual one.

Algorithm 1 illustrates this computation. First, it applies Convex Hull Value Iteration [4], a modification of the original Bellman's Value Iteration algorithm [26] that allows learning the optimal policies for all linear preference assignments over multiple objectives. The resulting convex hull contains the subset of policies that are optimal for some value of the ethical weight w_e. Thus, second line of the algorithm exploits the convex hull to extract from it the value of the policy with the maximum amount of ethical value

$(V_N + V_E)$ (i.e., the value $\vec{V}^*$ of the ethical-optimal policy π^*), and the value of the policy with the second-best value ($\vec{V'}^*$). Next, third line finds the values of w_e for which the former policy becomes optimal by computing the minimal weight satisfying:

$$V_0^*(s) + w_e[V_N^*(s) + V_E^*(s)] > V'^*_0(s) + w_e[V'^*_N(s) + V'^*_E(s)]. \quad (3)$$

Algorithm 1 Ethical Embedding [19]

function EMBEDDING(Ethical MOMDP $\langle S, A, (R_0, R_N + R_E), T \rangle$)
Compute the convex hull for weight vectors $\vec{w} = (1, w_e)$ with $w_e > 0$
Find $\vec{V}^*$ the ethical-optimal value vector, and $\vec{V'}^*$ the second-best value vector in the convex hull
Find the minimal value for w_e that satisfies Eq. 3
return $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$

4. Environment design for an in-game survey agent to learn to be respectful

As previously mentioned, the ethical environment design process first defines an ethical MOMDP to then transform it into an ethical MDP by applying the embedding algorithm.

In our particular setting (see Figure 2), we define our ethical MOMDP $\langle S, A, \vec{R}, T \rangle$ so that states in S include information about current game status (level and if menu or in-game) and user's activity (if engaged[3] or if the answer to last question was valid/non-valid or quick/slow). Moreover, the agent can perform two actions $A = \{Ask, Wait\}$ and the reward vector $\vec{R} = (R_0, R_N + R_E)$ contains the individual and ethical reward functions:

- R_0 (individual reward): promotes collecting as many valid answers as possible.

$$R_0(s, a, s') \doteq \begin{cases} 1, & \text{if } a\text{=}Ask \text{ and } valid\_answer(s') \\ 0, & \text{otherwise} \end{cases}$$

- R_N (normative reward): punishes i) asking questions when the user is engaged or provides non-valid or slow answers; and ii) waiting (i.e., not asking questions) when the user is not engaged, as these moments of low engagement should not be wasted:

$$R_N(s, a, s') \doteq \begin{cases} -2, & \text{if } ((a\text{=}Ask \text{ and } (engaged(s) \text{ or not } valid\_answer(s') \text{ or } slow\_answer(s'))) \\ & \quad \text{or } (a\text{=}Wait \text{ and not } engaged(s))) \\ 0, & \text{otherwise} \end{cases}$$

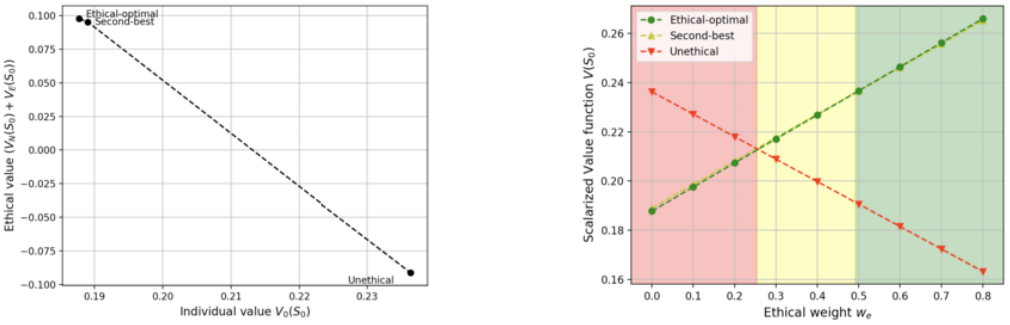
- R_E (evaluative reward): promotes asking questions that get a quick and valid response without interrupting engagement:

$$R_E(s, a, s') \doteq \begin{cases} 1, & \text{if } (a\text{=}Ask \text{ and } quick\_answer(s') \text{ and } valid\_answer(s') \text{ and not } engaged(s)) \\ 0, & \text{otherwise} \end{cases}$$

Thus $R_N + R_E$ encapsulates our notion of respect applied to the context of performing in-game questionnaires. Finally, state transition probabilities in $T(s, a, s')$ are approximated by observing the frequencies of such transitions in 500 game executions.

Next, we apply the ethical embedding algorithm. Figure 4a visualizes the convex hull, that is, those policies that are maximal for some value of w_e. Specifically, black dots signal the ethical-optimal policy ($\vec{V}^*$, the one that maximizes the ethical value function

[3]Notice that in our simple Pong game, engagement can be assumed if the user moves the paddle, but this varies for different games. Moreover, although moving the paddle can only be done in-game, and thus we assume low engagement in menus, it may also happen if the play is slow enough.



(a) The convex hull for our ethical MOMDP.

(b) Scalarized policies in the weight space.

Figure 4. The ethical embedding process: (a) visualizing the convex hull, and (b) finding the ethical weight.

$(V_N + V_E)$); the second-best ethical optimal policy ($\vec{V'}^*$); as well as the (unethical) policy that maximizes the individual value (V_0). Next, we solve Eq. 3 and obtain a value of $w_e > 0.49237$. In fact, this value can be empirically found by plotting, as in Figure 4b, the scalarised values for these tree policies, and by identifying the value of w_e for which the ethical-optimal policy has the highest scalarised value (and this is also the case for all w_e values in the green area). Then, we set the weight to $w_e = 0.5$ and return the ethical MDP $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$ as the environment that guarantees that the agent will learn to behave ethically. Finally, it is worth mentioning that Theorem 1 in [19] formally guarantees that the agent will still learn the same ethical optimal policy regardless of the scale[4] of the ethical rewards considered before scalarisation.

5. Results

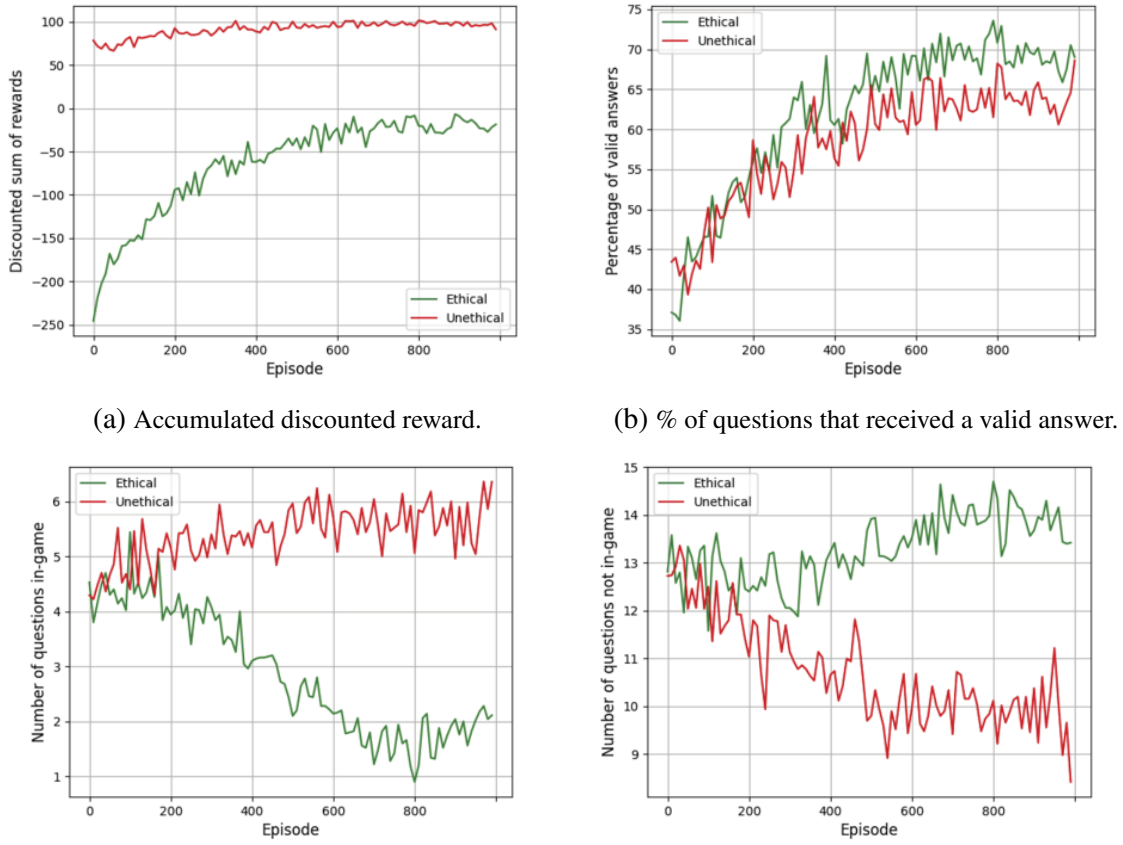
The resulting ethical MDP provides a simple environment for our conversational agent to learn to be respectful while asking survey questions. Here, we empirically prove so by applying Q-learning [26]. Specifically, we set a learning rate $\alpha = 0.7$, a discount factor $\gamma = 0.7$, and an ε-greedy policy for exploration along 1000 episodes, where each episode corresponds to a playthrough of our three-level Pong game[5].

To better assess the impact of the ethical embedding. Figure 5a illustrates the convergence, in terms of the accumulated reward, of the learning of two agents: in green, our ethical agent; in red, an unethical agent that just considers the individual reward R_0. Not surprisingly, our ethical agent takes longer to learn, and accumulates negative rewards as the R_N reward is quite demanding and punishes the agent for not taking advantage of all low engagement situations in slow play. However, this does not preclude our ethical agent to elicit necessary information. In fact, as depicted in Figure 5b, once it learns, it manages to get more valid answers than the unethical agent, which relies on the user to answer questions even if interrupted.

Beyond checking that the ethical agent manages to accomplish its individual objective, we need to assess it learns a respectful behaviour, asking questions when the user's engagement is low, which typically happens while the user is in menus. Thus, we focus on comparing the number of questions prompted in-game and in menus. Specifically, Figure 5c shows how the green ethical agent manages to drastically reduce the number of questions in-game (as opposed to the red unethical agent) and Figure 5d shows how the

[4] As long as the reward of praiseworthy actions are > 0 and the ones for blameworthy actions are < 0.

[5] Our code is publicly available at https://github.com/ericRosello/EthicalCA.



(a) Accumulated discounted reward.

(b) % of questions that received a valid answer.

(c) Number of questions in-game.

(d) Number of questions in menus.

Figure 5. Evolution of different metrics throughtout the learning process.

ethical agent focuses in asking most of the questions in menus (a behaviour that again contrasts with the one of the unethical agent). Thus, overall, we can claim that our conversational agent has successfully learnt to ask survey questions without disturbing the user play, that is, behaving in alignment with the moral value of respect.

6. Conclusions and Future Work

This paper proposes an ethical conversational agent in charge of gathering User eXperience data while the user is playing a game. The agent, applying the ethical embeddding method, learns to respectfully conduct the in-game questionnaire. This method transforms an ethical MOMDP into an ethical MDP that can be addressed by standard RL algorithms. Specifically, we defined the learning environment based on the Pong game, and used Q-learning with a simulated user to assess the ethical agent's learning. The results show that our ethical agent asks the user questions in more appropriate situations (low user engagement) than the unethical agent. Thus, it fulfils the ethical objective while still pursuing the individual one (i.e obtain as much UX data as possible). Indeed, the ethical agent obtained a higher proportion of valid answers than the unethical one, while reducing gameplay interruptions.

Future work should explore the generalization of our approach to alternative games and virtual reality experiences, as the activity of the user (and so engagement) is highly dependent on the (game) mechanics. The study of other moral values (e.g. fairness) is another interesting line of research.

References

[1] D. Alexandrovsky, S. Putze, M. Bonfert, S. Höffner, P. Michelmann, D. Wenig, R. Malaka, and J. D. Smeddinck. Examining design choices of questionnaires in vr user studies. In *CHI'20, 1–21*, 2020.

[2] M. Anderson, S. L. Anderson, and C. Armen. An approach to computing ethics. *IEEE Intelligent Systems*, 21(4):56–63, 2006.

[3] T. Arnold, D. Kasenberg, and M. Scheutz. Value alignment or misalignment - what will keep systems accountable? In *AAAI Workshops*, 2017.

[4] L. Barrett and S. Narayanan. Learning all optimal policies with multiple criteria. *Proc. of 25th ICML*, pages 41–47, 01 2008.

[5] A. Beck, B. Stevens, K. A Bard, and L. Cañamero. Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):1–29, 2012.

[6] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale. An evaluation methodology for interactive reinforcement learning with simulated users. *Biomimetics*, 6(1):13, 2021.

[7] J. Casas-Roma and J. Conesa. Towards the design of ethically-aware pedagogical conversational agents. In *Int. Conference on 3PGCIC*, pages 188–198. Springer, 2020.

[8] J. Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67–67, 2001.

[9] A. Cheng and K. R. Fleischmann. Developing a meta-inventory of human values. *Proc. of the ASIS&T*, 47(1):1–10, 2010.

[10] X. Han, M. Zhou, M. J. Turner, and T. Yeh. Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proc. of CHI'21*, pages 1–15, 2021.

[11] M. Hassenzahl. User experience and experience design. *The encyclopedia of HCI*, 2, 2013.

[12] J. R. Keebler, W. J. Shelstad, D. C. Smith, B. S. Chaparro, and M. H. Phan. Validation of the guess-18: a short version of the game user experience satisfaction scale. *J. of Usability Studies*, 16(1):49, 2020.

[13] E. L. Law. The measurability and predictability of user experience. In *ACM SIGCHI EICS, 1–10*, 2011.

[14] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23, 2000.

[15] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the ASIS&T*, 59(6):938–955, 2008.

[16] C. Peters, G. Castellano, and S. De Freitas. An exploration of user engagement in hci. In *Proc. of the Int. Workshop on Affective-Aware Virtual Agents and Social Robots*, pages 1–3, 2009.

[17] I. Rodríguez and A. Puig. Open the microphone, please! conversational ux evaluation in virtual reality. In *Workshop 'Evaluating user experiences in mixed reality' in CHI'21*, 2021.

[18] M. Rodriguez-Soto, M. Serramia, M. Lopez-Sanchez, and J. A. Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):1–17, 2022.

[19] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. Multi-objective reinforcement learning for designing ethical environments. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 1–7, 2021.

[20] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, October 2013.

[21] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126, 2006.

[22] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proc. of HLT*, volume 2, 2002.

[23] S. H. Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.

[24] V. J. Shute. Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2):503–524, 2011.

[25] A. Steinmaurer, M. Sackl, and C. Gütl. Engagement in in-game questionnaires-perspectives from users and experts. In *2021 7th Int. Conference of the iLRN*, pages 1–7. IEEE, 2021.

[26] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[27] W. Wallach and C. Allen. *Moral machines: teaching robots right from wrong*. Oxford Univ. press, 2008.

[28] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang. Building ethics into artificial intelligence. In *IJCAI*, page 5527–5533, 2018.

Chapter 6

Conclusions

This chapter provides a summary of the results and contributions obtained during this thesis, a general discussion of them, and the conclusions of this work.

This chapter is structured as follows. First, Section 6.1 provides a detailed list of the concrete answers to our research questions and discusses the results obtained. Then, Section 6.2 provides the conclusions and main takeaways from this thesis, and also directions for future work.

6.1 Results

During this work, we have addressed all the research questions posed in Section 1.2. In this section, we detail our answer to each research question, divided by chapter, following the order in which they were answered. Hence, each of the following subsections provides a summary of the results provided in its respective chapter.

6.1.1 Designing an ethical environment

The paper in Chapter 2 provided our first contribution. It was the design of an ethical environment wherein two agents of a multi-agent game learnt to behave in alignment with a moral value. With this ethical environment, we aimed to prove the viability of an environment-designer approach to align agents with moral values.

In [Rodriguez-Soto et al., 2020], we considered a multi-agent game in which agents needed to behave in alignment with a moral value. The game was a reinforcement learning environment in which each agent had its individual reward function R_0. Then, we aggregated the ethical reward function R_v directly into the environment so that the

agents received the reward function $R = R_0 + R_v$. The environment with reward function R was the *hand-crafted* ethical environment.

After letting the agents learn their policies in the ethical environment, both of them learnt to behave in alignment with the moral value. In summary, we positively answer Research Question Q1:

Question Q1: *Given an environment with some (already specified) ethical rewards, can we design an ethical single-objective environment wherein the agent learns to behave ethically?* Indeed, we show that given an example multi-agent environment, we can design an ethical environment in [Rodriguez-Soto et al., 2020] (contribution C1).

We illustrated our proposal with the Public Civility Game, a multi-agent game in which the agents need to behave in alignment with the moral value of civility. Recall that in the Public Civility Game, both agents may find pieces of garbage in their way, which we expect them to bring to the nearest bin.

We compared the policies learnt by the agents in the designed ethical environment and the original environment. We empirically showed that the multi-agent system improves its overall performance in terms of street cleanness in the ethical environment, because garbage is always brought to the bin. Furthermore, in the ethical environment, we also reduced the degree of violence of the environment by 100% (i.e., there was no violence) because no agent threw the piece of garbage to the position of the other.

In summary, the positive results [Rodriguez-Soto et al., 2020] opened the possibility of going from a hand-crafted to an automated ethical environment design process. Therefore, the following two Chapters were devoted to developing an algorithm for designing ethical environments.

6.1.2 The ethical embedding process

After the positive results of Chapter 2, we proceeded to develop an algorithm for automating the ethical environment design algorithm. Recall that the design of ethical environments consists of two parts. First, the reward specification process creates an ethical reward function R_v. Second, the ethical embedding process aggregates the ethical reward function R_v to the original function R_0 of an environment to create an ethical environment.

As mentioned in the Introduction section, in this thesis, we started with the second step, the ethical embedding process. We assumed that the ethical reward function was already provided to the environment designer so we could focus first on the problem of how to

aggregate the two reward functions ($R_0 + w \cdot R_v$) to create an ethical single-objective environment. The agent is expected to learn an ethical policy in the resulting ethical environment.

The paper in Chapter 3 presented an ethical embedding algorithm that designs an ethical environment in three steps. A general summary of the three steps is provided in Section 1.3.2. The main idea is that it applies a multi-objective reinforcement learning algorithm to compute the difference in rewards between behaving ethically and not. After computing this difference, it can find the weight necessary to *compensate* behaving ethically. Recall that a policy is ethical if it maximises the accumulation of ethical rewards. Hence, it is likely that an ethical policy does not accumulate as many individual rewards as an unethical one. With the computed weight, in the ethical environment, it is optimal for the agent to behave ethically. Hence, our answer to Research Question Q2.1 is *yes*:

Question Q2.1: *Can we develop an algorithm for the ethical embedding process so that it designs an ethical environment wherein the agent learns to behave ethically?* Yes, the Ethical Embedding Process presented in Algorithm 1 of [Rodriguez-Soto et al., 2021] (contribution C2.1).

We illustrated the algorithm with a single-agent version of the Public Civility Game. Again, the empirical evaluation confirmed that the learning agent learnt to behave civilly (i.e., always bringing the garbage to the nearest bin) while going to its destination.

6.1.3 The ethical environment design process and its formal guarantees

In [Rodriguez-Soto et al., 2021], we developed an algorithm for automating the ethical embedding process, the second step of the ethical environment design process. This result prompted us to complete the environment design algorithm by providing an algorithm for the remaining step: the reward specification process.

The paper in Chapter 4 provided an algorithm for the reward specification process. Given a moral value formalised following [Serramia et al., 2018] and a reinforcement learning environment, in [Rodriguez-Soto et al., 2022], we explain how to obtain an ethical reward function from them. Recall that a moral value is formalised as a tuple of two elements: (i) a set of norms that tell us which actions are prohibited for the value, and (ii) an evaluative function that quantitatively tells us how praiseworthy is each possible action for the value.

In [Rodriguez-Soto et al., 2022], we presented a reward specification process that independently transforms each element of the moral value into a component of a reward function. The set of norms is transformed into a normative reward function R_N, and the evaluative function is transformed into an evaluative reward function R_E. Finally, since both components are considered equally important, they are aggregated to obtain the ethical reward function $R_v = R_N + R_E$. Hence, our response to Research Question Q2.2 is *yes* again:

Question Q2.2: *Can we develop an algorithm for the reward specification process to transform a moral value into an ethical reward function?* Yes, the Reward Specification Process is formalised in Definition 7 of [Rodriguez-Soto et al., 2022] (contribution C2.1).

The resulting ethical reward function specifies all the elements of a moral value in terms of rewards of a reinforcement learning environment. Therefore, in [Rodriguez-Soto et al., 2022], we argue that to behave ethically means to maximise the accumulation of rewards from the ethical reward function. This proposed formalisation of ethical behaviour, albeit restrictive, is formally precise. Our definition means that we only consider as ethical the policy that maximises the accumulation of ethical rewards (i.e., optimal for the ethical reward function). One could ask why a policy that obtains 95% of the total possible ethical rewards should not be called ethical. Our reply to this is that if there are some *optional* ethical actions that the agent is not required to do (known as *supererogatory* actions in the Ethics literature, [Chisholm, 1963]), then it is a matter of not including them in the formalisation of the moral value (or when specifying the associated ethical reward). Nevertheless, it is interesting to consider a more relaxed definition of ethical behaviour in future work.

The proposed reward specification process in [Rodriguez-Soto et al., 2022] completed the ethical environment design process. After that, Chapter 4 presented the two-step ethical environment design algorithm, which answers Research Question Q2:

Question Q2: *Given a moral value and an agent's reinforcement learning environment, can we develop an algorithm for transforming the environment into an ethical environment wherein it is in the agent's best interest to behave ethically? ?* Yes, the Ethical Environment Design Process presented in Algorithm 1 of [Rodriguez-Soto et al., 2022] (contribution C2).

Our findings in [Rodriguez-Soto et al., 2022] lead to an algorithm for automating the design of ethical environments. In the built ethical environment, it will be in the agent's best interest to behave ethically. Again, we illustrated our algorithm by applying it to the single-agent version of the Public Civility Game.

When developing Algorithm 1 in [Rodriguez-Soto et al., 2022], our primary focus was on providing theoretical guarantees to it. We wanted to prove that optimal policies are ethical in the designed ethical environment. We proved this theoretical conjecture in Theorem 2 of [Rodriguez-Soto et al., 2022]. Thus, we proved that our ethical environment design process creates environments wherein the agent learns to behave ethically when its formal assumptions are satisfied. In particular, there are two assumptions:

- The first one is that the reinforcement learning environment is *finite*. That is, the number of states and actions is finite. This is a natural assumption usually expected in all tabular reinforcement learning applications.
- The second one is that behaving ethically should be possible in the reinforcement learning environment. That is, the agent can maximise the accumulation of praiseworthy and normative rewards (the two components of the ethical reward function). This assumption comes from the Ethics literature, in which if we demand a moral agent to perform some ethical actions, it must be possible for them to perform these actions. In the Ethics literature, this condition is known as *Ought implies Can* [Duignan, 2018].

In short, we can answer positively Research Question Q3:

Question Q3: *Is our proposed ethical environment design process formally guaranteed to create an environment wherein it is in the agent's best interest to behave ethically?* Yes, given two theoretical assumptions, as proven by Theorem 2 of [Rodriguez-Soto et al., 2022] (contribution C3).

Theorem 2 in [Rodriguez-Soto et al., 2022] is the main result of this thesis. This result, together with Algorithm 1 of [Rodriguez-Soto et al., 2022], advances the state of the art by providing the first algorithm with theoretical guarantees of ethical-behaviour learning, which is paramount to releasing autonomous agents in real-world environments. Furthermore, Theorem 2 validates all our previous empirical evaluations of ethical environments theoretically, confirming that it was guaranteed that it is in the agents' best interest to behave ethically.

In the designed ethical environment, optimal policies are ethical. It is worth remarking that we are assuming that the chosen learning algorithm of the agent must be at least formally guaranteed to converge to the optimal policy. Since we have chosen an environment-designer approach in this thesis, we assume that we have no control over the agent, and we only assume that it is a rational agent pursuing to maximise its reward accumulation.

To conclude this section, we wanted to discuss the applicability of the ethical environment design process to *deep* reinforcement learning environments. Currently, *no deep reinforcement learning algorithm is formally guaranteed to converge to optimality.* Such (lack of) theoretical results means that:

1. With deep reinforcement learning algorithms, we cannot guarantee that in the designed ethical environments it is optimal for an agent to behave ethically.
2. Even if we designed the ethical environment with tabular reinforcement learning by brute force if the agent uses a deep reinforcement learning algorithm for learning, there is no guarantee that it will converge to the ethical policy.

Nevertheless, considering that deep reinforcement learning is still a very recent research area, it is to be expected that future work in this direction would lead to deep learning algorithms with theoretical guarantees. In any case, the view in this thesis is that building a formal basis of algorithms for value alignment in *tabular* reinforcement learning environments is necessary before jumping to deep reinforcement learning algorithms.

6.1.4 Evaluating the ethical environment design process

The theoretical results in Chapter 4 guarantee that optimal policies are ethical in any ethical environment designed by our algorithm. However, as argued in the Introduction section, we deem it important to illustrate our ethical environment design algorithm in a case study.

Until Chapter 4 we only tested our ethical environment design process in the Public Civility Game. The paper in Chapter 5 provided a second environment with potential applications in the real world. As explained in Section 1.3.4, in [Roselló-Marín et al., 2022] we presented the learning environment of a conversational agent. The learning agent needs to extract user experience information from a human user that is playing a video game. We expect the agent to align with the moral value of *respect* in this environment. We expected the agent to ask the user while not disturbing the user's engagement with the video game.

After designing the ethical environment, we evaluated the policy learnt by the agent. As expected, the empirical results confirmed our theoretical results. The evaluation results showed that in the designed ethical environment, the agent abided by the moral value of respect: the conversational agent managed to avoid disturbing a user's engagement. In summary, we answered Research Question **Q4** with our case study:

Question Q4: *Can we test the ethical environment design process in a reinforcement learning environment to validate that an agent learns a behaviour aligned with a moral value?* Yes, with the conversational agent presented in [Roselló-Marín et al., 2022]. We validate our algorithm in Section 5 of the paper (contribution C4).

In conclusion, the results found in the papers provided in this compendium answered all the research questions posed in this thesis.

6.2 Conclusions and Future Work

This thesis aimed to develop ethical environments in which an agent is formally guaranteed to learn to behave ethically. Founded on the Ethics literature and the Multi-Objective Reinforcement Learning literature, we provided a two-step ethical environment design algorithm and validated it theoretically and empirically.

Guaranteeing value alignment is a challenging problem. While most of the AI community is following an agent-based approach to tackle this problem, here in this thesis, we have argued instead for using an environment-designer approach. That is, instead of modifying the agent's learning algorithm, we modify its environment so that it is in the agent's best interest to learn an ethical behaviour. With our ethical environment design process, we expected to ease the agent's learning process. Indeed, as our empirical evaluation confirms in the tested environments (The Public Civility Game and the survey-oriented conversational agent case study), the agents can learn an ethical policy using any tabular reinforcement learning algorithm. This is crucial because, as we have discussed throughout this thesis, we will not always be able of guaranteeing that the agent is willing to behave ethically. Our ethical environment design process provides a method to know exactly how to tailor the agent's reward function so it is optimal for the agent to behave value-aligned.

In this thesis, we have provided an algorithm for guaranteeing that a single agent learns to behave in alignment with a moral value. Hence, there are two natural future work directions: (i) to develop an algorithm for designing an ethical multi-agent environment for multiple agents, and (ii) to develop an algorithm for designing an ethical wherein it is in the agent's best interest to behave in alignment with multiple moral values. In both cases, we would like to follow the same methodology we applied in this thesis and provide theoretical guarantees for the eventual algorithms. Therefore, for the multi-agent case, we would require our algorithm to be generalised for learning environments with multiple agents (known in the reinforcement learning literature as *Markov games*). Similarly, for the multi-valued case, we would require our algorithm to be generalised for learning environments with many objectives (more than two). Both objectives are achievable given the state of the art in multi-agent and multi-objective reinforcement learning. In fact, our journal paper under review cited in the Introduction tackles the multi-agent ethical environment design problem.

Another possible line of work would be, on the theoretical side, to generalise our Theorems so that they hold in more general environments. In this thesis, our theorems remain valid for finite Markov Decision Processes. However, in large environments, it is common for the agent to learn with partial observations instead of considering the

environment's whole state at each step. Hence, it would be interesting to study how to generalise them for partially-observable environments. In the same direction, as previously mentioned, it would also be interesting to study how to expand them to deep reinforcement learning environments.

In summary, this thesis aimed to close an important gap in the AI value alignment research area. Despite the popularity of applying reinforcement learning for value alignment, before our approach, no methods in the literature formally guaranteed that it was in the agent's best interest (i.e., optimal) to behave ethically. With its theoretical guarantees, the presented ethical environment design process serves as a stepping stone towards solving the value alignment problem because there will not be general artificial intelligence if there is no value-aligned artificial intelligence.

References

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

Lucas N. Alegre, Florian Felten, El-Ghazali Talbi, Grégoire Danoy, Ann Nowé, Ana L. C. Bazzan, and Bruno C. da Silva. MO-Gym: A library of multi-objective reinforcement learning environments. In *Proceedings of the 34th Benelux Conference on Artificial Intelligence BNAIC/Benelearn 2022*, 2022.

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

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

Ahmad Taher Azar, Anis Koubaa, Nada Ali Mohamed, Habiba A. Ibrahim, Zahra Fathy Ibrahim, Muhammad Kazim, Adel Ammar, Bilel Benjdira, Alaa M. Khamis, Ibrahim A. Hameed, and Gabriella Casalino. Drone deep reinforcement learning: A review. *Electronics*, 10(9), 2021. ISSN 2079-9292. doi: 10.3390/electronics10090999. URL `https://www.mdpi.com/2079-9292/10/9/999`.

Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3–11, 07 2019. doi: 10.1609/aaai.v33i01.33013.

Leon Barrett and Srini Narayanan. Learning all optimal policies with multiple criteria. *Proceedings of the 25th International Conference on Machine Learning*, pages 41–47, 01 2008. doi: 10.1145/1390156.1390162.

Trevor J. M. Bench-Capon and Katie Atkinson. Abstract argumentation and values. In *Argumentation in Artificial Intelligence*, pages 45–64. Springer, 2009. URL `http://dx.doi.org/10.1007/978-0-387-98197-0_3`.

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

Joan Casas-Roma and Jordi Conesa. Towards the design of ethically-aware pedagogical conversational agents. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 188–198. Springer, 2020.

Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. Trustworthy ai. In *Reflections on Artificial Intelligence for Humanity*, pages 13–39. Springer, 2021.

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

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

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis, and Pushmeet Kohli. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47–53, 10 2022. doi: 10.1038/s41586-022-05172-4.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. 2018.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30: 411–437, 09 2020. doi: 10.1007/s11023-020-09539-2.

Dan Garisto. Google ai beats top human players at strategy game starcraft ii. *Nature*, 10 2019. doi: 10.1038/d41586-019-03298-6.

Sven Ove Hansson. *The Structure of Values and Norms*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, Cambridge, 2001. doi: 10.1017/CBO9780511498466.

Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 2022.

Peter Kollock. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24(1):183–214, 1998. doi: 10.1146/annurev.soc.24.1.183. URL `https://doi.org/10.1146/annurev.soc.24.1.183`.

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

Michael Littman. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521:445–51, 05 2015. doi: 10.1038/nature14540.

Jieting Luo, John-Jules Meyer, and Max Knobbout. Reasoning about opportunistic propensity in multi-agent systems. In *AAMAS 2017 Workshops, Best Papers.*, pages 1–16, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 12 2013.

Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Ramazon Kush, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*, PP:6377–6381, 09 2019. doi: 10.1147/JRD.2019.2940428.

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

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

Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez Aguilar. Multi-objective reinforcement learning for designing ethical environments. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 545–551. International Joint Conferences on Artificial Intelligence Organization, August 2021. Main Track.

Manel Rodriguez-Soto, Marc Serramia, Maite López-Sánchez, and Juan Rodríguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24, 03 2022. doi: 10.1007/s10676-022-09635-0.

Diederik Roijers and Shimon Whiteson. *Multi-Objective Decision Making*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool, California, USA, 2017. URL `http://www.morganclaypool.com/doi/abs/10.2200/S00765ED1V01Y201704AIM034`. doi:10.2200/S00765ED1V01Y201704AIM034.

Eric Roselló-Marín, Maite Lopez-Sanchez, Inmaculada Rodríguez, Manel Rodríguez-Soto, and Juan A Rodríguez-Aguilar. An ethical conversational agent to respectfully conduct in-game surveys. In *Artificial Intelligence Research and Development*, pages 335–344. IOS Press, Amsterdam, 2022.

Francesca Rossi and Nicholas Mattei. Building ethically bounded ai. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9785–9789, 07 2019. doi: 10.1609/aaai.v33i01.33019785.

Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36:105–114, 12 2015. doi: 10.1609/aimag.v36i4.2577.

Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34:1–52, 2019.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *CoRR*, abs/1911.08265, 2019. URL `http://arxiv.org/abs/1911.08265`.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–, October 2017. URL `http://dx.doi.org/10.1038/nature24270`.

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

Margit Sutrop. Challenges of aligning artificial intelligence with human values. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8:54–72, 12 2020. doi: 10.11590/abhps.2020.2.04.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 0262193981. URL `http://www.worldcat.org/oclc/37293240`.

James O. Urmson. Saints and heroes. In A. I. Melden, editor, *Essays in Moral Philosophy*. University of Washington Press, 1958.

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

Nikos A. Vlassis. A concise introduction to multiagent systems and distributed artificial intelligence. In *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*, 2009.

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

Peter Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael Thomure, and Hiroaki Kitano. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602:223–228, 02 2022. doi: 10.1038/s41586-021-04357-7.

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