Multi-objective reinforcement learning: an ethical perspective

Timon Deschamps, Rémy Chaput, Laetitia Matignon

Multi-objective reinforcement learning: an ethical perspective

T. Deschamps[1], R. Chaput[1], L. Matignon[1]

[1] Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

timon.deschamps@liris.cnrs.fr

Abstract

Reinforcement learning (RL) is becoming more prevalent in practical domains with human implications, raising ethical questions. Specifically, multi-objective RL has been argued to be an ideal framework for modeling real-world problems and developing human-aligned artificial intelligence. However, the ethical dimension remains underexplored in the field, and no survey covers this aspect. Hence, we propose a review of multi-objective RL from an ethical perspective, highlighting existing works, gaps in the literature, important considerations, and potential areas for future research.

Keywords

Reinforcement learning, multi-objective decision making, machine ethics.

Résumé

L'apprentissage par renforcement est de plus en plus employé pour des applications pratiques impactant l'humain, soulevant ainsi des questions éthiques. Spécifiquement, l'apprentissage par renforcement multi-objectif est considéré comme un cadre idéal pour la modélisation de problèmes concrets et le développement de systèmes d'intelligence artificielle alignés sur l'humain. Peu de travaux du domaine adoptent une perspective éthique, et les études existantes ne couvrent pas cet aspect. Ainsi, nous proposons une revue de l'apprentissage par renforcement multi-objectif d'un point de vue éthique, en détaillant les travaux existants, les lacunes de la littérature, les considérations importantes, et les potentielles pistes de recherche futures.

Mots-clés

Apprentissage par renforcement, prise de décision multi-objectifs, éthique computationnelle.

1 Introduction

The field of *reinforcement learning* (RL) has recently seen numerous breakthroughs, notably featuring artificial intelligence (AI) agents beating humans at a wide variety of games [50, 8]. RL has also been applied to multiple real-world problems, with a potentially large impact on societies, e.g., nuclear fusion control [16], healthcare [69]. This calls for the study of the ethical issues that may arise from such uses, and the development of techniques to ensure that the agents have a behavior deemed *ethically-aligned* with human principles; so as to guarantee this technology will be beneficial to humanity. This is a complex endeavor, and a few works have started paving the way [66, 52].

In this paper, we focus on *multi-objective reinforcement learning* (MORL), a sub-field of RL in which multiple potentially conflicting goals are considered rather than a single one. Following the RL trend, MORL is being increasingly used in real world applications such as public bicycle dispatching [14] or energy management [19]. It has been argued that aligning AI with human goals is a multi-objective problem [58], making the study of MORL interesting in this regard. A few multi-objective decision making surveys have been published [25, 48], focusing on the theory and applications of multi-objective decision making algorithms. The goal of this work is to highlight the need for ethically-aligned multi-objective methods and to conduct an analysis of MORL from a moral standpoint. To do so, we start by discussing and categorizing existing MORL methods, before introducing important ethical considerations, which we use to emphasize important gaps in the literature.

2 A motivating example

To illustrate the ethical concerns that can arise when AI agents are deployed in the real-world, we propose to study the case of self-driving vehicles. This sector has been increasingly interested in RL [28], which is viewed as a suitable paradigm: vehicles can be represented by agents taking actions such as steering and accelerating within an environment (road network).

RL agents typically optimize for a single objective, e.g., *speed*. However, when dealing with complex use-cases or when humans can be impacted, more flexibility is desirable to account for additional goals like *cost saving* and *comfort*.

MORL is ideal in such contexts, as it allows for representing and compromising between multiple objectives. This multi-objective aspect is essential when autonomous vehicles are deployed on real roads, as human error, technical malfunctions or unexpected situations will inevitably occur, leading the machine to have to handle complex ethical dilemmas which require weighting between conflicting moral values, e.g., ensuring safety for both passengers and surrounding pedestrians in an inevitable accident scenario.

This example motivates the study of MORL agents with an ethically-aligned behavior, and we will extend it throughout this paper to illustrate some of the notions discussed.

3 Background

3.1 Reinforcement learning

Reinforcement learning is a general framework to solve problems in which an agent alternatively takes *actions* and receives *observations* and *rewards* from an environment, and aims at maximizing the cumulative reward obtained. RL is usually modeled as a *Markov decision process* (MDP), defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle$, where:

- $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively;
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, i.e., the probability of transitioning to a state s_{t+1} given that the action a_t was taken at time step t in state s_t;
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, which outputs a scalar reward for a given (s_t, a_t, s_{t+1}) tuple;
- $\gamma \in [0, 1)$ is a discount factor modulating the importance of long term rewards.

The agent acts according to a *stochastic policy* $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which gives the probability of taking any action $a \in \mathcal{A}$ given the current state $s \in \mathcal{S}$. If in every state one of the actions is selected with probability 1, the policy becomes *deterministic*, denoted $\pi : \mathcal{S} \to \mathcal{A}$[1].

At any time step t, we can compute the sum of future rewards, or *return*, defined as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k. \quad (1)$$

The *value* of a state $V^\pi(s) = \mathbb{E}_\pi [G_t \mid S_t = s]$ is the expected return for an agent located in this state at time step t and following policy π. In turn, our goal is to find the optimal policy π^* which, when followed, maximizes the value for all states in $\mathcal{S}$.

To this day, RL remains a highly active discipline, with many emerging sub-fields such as multi-agent RL [23, 11], model-based RL [33] and multi-objective RL, the latter of which we discuss in the following section.

3.2 Multi-objective reinforcement learning

The field of multi-objective reinforcement learning (MORL) deals with *multi-objective Markov decision processes* (MOMDPs). MOMDPs differ from regular MDPs only in that the reward (and by extension the value) is vector-valued: $\mathbf{r} \in \mathbb{R}^m$ with m objectives[2]. This implies that finding a single optimal policy via a simple maximization process becomes impossible, as maximizing one of the component of the reward vector (called objective) could lead to a decrease in another one.

Utility functions, also referred to as scalarization functions, map the value vector $\mathbf{V}^\pi$ of a given policy π to a single scalar ($u : \mathbb{R}^m \to \mathbb{R}$). They provide a convenient way to formalize a decision maker's preferences and trade-offs over the objectives.

A common and simple class of utility functions are linear utilities, denoted as $u(\mathbf{V}^\pi) = \mathbf{w}^\top \mathbf{V}^\pi$, which combines a weight vector $\mathbf{w}$ in the $(m-1)$-simplex[3] and the value vector using a linear combination. Intuitively, each weight $w_o \in \mathbf{w}$ represents the importance of the associated objective $\mathbf{V}^\pi_o$.

If we have access to a linear utility function for the user, we can use it to simplify the problem back into the single-objective RL setting and solve it with classical methods. However, this is not an option when the utility function is not fully known in advance or is non-linear, which represents a large portion of real-life scenarios (see the motivating scenarios presented in [25]).

In these settings, we focus instead on a set of optimal policies: the *Pareto front* (PF). A policy $\pi \in \Pi$ belongs to the Pareto front PF(Π) if it is not Pareto-dominated by any other policy. The Pareto-dominance of a policy π over a policy π' is defined as:

$$\pi \succ_P \pi' := (\forall o : \mathbf{V}^\pi_o \geq \mathbf{V}^{\pi'}_o) \wedge (\exists o : \mathbf{V}^\pi_o > \mathbf{V}^{\pi'}_o). \quad (2)$$

In plain words, π's associated value vector is greater or equal to the one associated with π' for all objectives o, and strictly greater for at least one.

As the PF can have multiple policies with the same induced value function, we often refer to a *Pareto coverage set* (PCS), which simply retains a single policy for each non Pareto-dominated value function. Computing a PCS guarantees that we have access to all policies that are optimal under some monotonically increasing utility function. This allows to adapt to changes in the user's preferences while making minimal assumptions about u. In practice, however, PF and PCS can be prohibitively large to compute. Recent works [48, 25, 41] have argued for a utility-based approach, in which we use information we have about the utility function to guide our search in the space of policies. For example, when u is known to be linear, we can restrict our focus to subsets of the PF referred as *convex coverage sets* (CCS), which contain all maximal policies under this assumption.

To illustrate these concepts, let's take our example from section 2. Keeping only 2 objectives (speed and comfort) for ease of representation, we can visualize the PF and a CCS in figure 1. Each point represents a policy and its associated value vector, compromising between the two objectives. We can see that increasing speed usually leads to a decrease in comfort, but it is not always the case (for instance, faster speeds on very uneven roads could smooth out the cruise). Notice that points belonging to the represented CCS are also part of the Pareto front (in fact CCS(Π) $\subseteq$ PF(Π)). Here, point b is not Pareto-dominated by any other point. Furthermore, there is no $\mathbf{w}$ for which a linear scalarization would lead to b being maximal. Thus, we can conclude that b belongs to the PF but not to a CCS.

When using a scalarization function, two optimization criteria naturally arise: *scalarized expected returns* (SER) and

[1] Some works also use $\mu(s) = a$ specifically for deterministic policies.

[2] Note that we use the standard notation of boldface for vector variables.

[3] The m-simplex, denoted Δ^m, is the set of all nonnegative vectors of $m + 1$ dimensions whose components sum to 1.

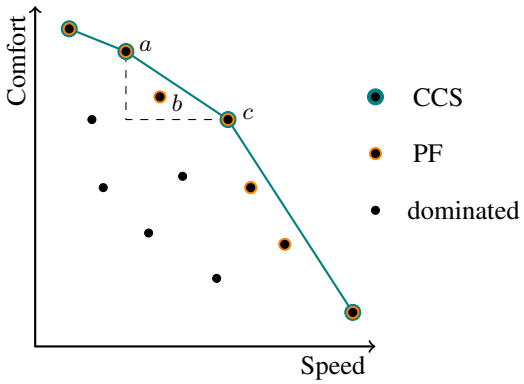

Figure 1: Visualization of the Pareto front and a convex coverage set for a 2-objective self-driving car example.

expected scalarized returns (ESR). To optimize for SER, we scalarize an expectation over multiple runs of the vector-valued returns of a policy, whereas optimizing for ESR requires having a scalarized return for each run, and then computing an expectation over them. These two criteria have different properties and should be used in different scenarios. SER, the most studied one, is particularly suited when we aim to optimize over many policy executions, whereas using the ESR criterion is better to ensure that each execution is maximal over our utility function.

See [48, 25] for a detailed overview of the theory and methods of multi-objective reinforcement learning.

3.3 Machine ethics

As autonomous machines are increasingly integrated into domains with significant human implications, their impact, whether it be positive or negative, requires investigation. *Machine ethics* is concerned with ensuring that AI agents demonstrate ethically-aligned behaviors, i.e., behaviors whose outcomes are acceptable according to some human-chosen ethical framework [6]. In turn, we aim for them to be *explicit ethical agents* [35], i.e., agents who are not simply constrained to avoid unethical behaviors but who integrate algorithmic capabilities [18] allowing them to perform ethics-related computations and to consider ethical considerations in their decision-making process. To evaluate the ethical alignment of these behaviors, we leverage insights from *normative ethics*. As it is concerned with the morality of actions, this field provides a suitable framework for such an analysis.

Normative ethics encompasses three main schools of thought: *consequentialism*, *virtue ethics* and *deontology*. According to consequentialism, only the outcomes of actions are necessary to judge whether these actions are ethical or not. Consequentialist ethics are most known for utilitarianism, which argues that in every situation, the ethical action is the one that maximizes happiness and well-being for all. Virtue ethics shift the focus from the action to its motivation. In this view, an agent is ethical if it acts according to set values (e.g., confidence, honour, freedom). Deontology takes a rule-based approach, in which actions can either be right or wrong according to a list of principles. Kantian ethics is a prime example of deontological ethical theories. We refer the interested reader to [54] for an extensive review of western moral philosophy.

As discussed in section 3.1, a defining feature of reinforcement learning agents is their ability to take actions in an environment, making normative ethics a natural framework for studying the ethical alignment of their behavior.

In fact, reinforcement learning has been characterized as an ideal framework to develop ethical agents [1], and recent work has surveyed RL-based moral learning agents [52]. Furthermore, we argue that the formulation of the reinforcement learning objective as the maximization of a future reward signal naturally aligns with a number of branches of consequentialism. Although some methods allow for the application of deontological ethics into RL [24, 5], none to our knowledge directly takes a moral perspective and is adapted to the multi-objective setting. Finally, it has been argued that MORL, on top of being ideal to model a number of real-world problems [25], is a particularly fitting framework to develop human-aligned artificial intelligence [58]. Moreover, we suggest that it is also suited for modeling virtue ethics, as each component of the vector-valued reward can encode a virtue to be followed. For a comprehensive overview of machine ethics implementations, refer to the survey of Tolmeijer et al. [55].

4 Classical MORL methods

The most commonly used taxonomy for multi-objective sequential decision making [48, 25] classifies methods depending on the type of policy and utility function they consider, resulting in a number of criteria:

- *single vs. multiple policies*: As mentioned in sec. 3.2, algorithms can either output a single solution (if the utility is fixed and known in advance) or a set of optimal policies. Multi-policy methods are **more costly**, but allow for **greater flexibility**: since fewer assumptions are made on the utility function, the user can adapt in the face of new data or changing contexts.

- *deterministic vs. stochastic policies*: While it was shown that **stochastic policies can outperform deterministic ones** in some environments [63, 56], their use can become ethically questionable or impossible in domains requiring strong guarantees (e.g., medical treatments).

- *linear vs. monotonically increasing u*: Using linear utility functions **simplifies the learning process**, allowing the MORL problem to be reduced to a single-objective one (for single-policy algorithms) or to restrict the policy search to a CCS (for multi-policy algorithms). Using monotonically increasing utility functions enables the expression of a much **richer relationship between the objectives**, at the cost of a **more complex learning process**, as the entire PF has to be considered.

	single policy (known u)		multiple policies (unknown u)	
	deterministic	stochastic	deterministic	stochastic
linear scalarization	one policy in Π_{DS}: DQN [32], REINFORCE [51]		CCS of policies in Π_{DS}: Envelope [68], PG-MORL [67], PD-MORL [9], CN [2]	
monotonically increasing scalarization	one policy in Π_{D}: EUPG [46], MOCAC [43], Q-steering [60]	mixture of policies in Π_{DS}: π-mix [56], S-rand [63]	PCS of policies in Π_{D}: PQL [34], PCN [42]	mixture of policies in Π_{DS}: CAPQL [29], π-mix [56], S-rand [63]

Table 1: Non-exhaustive classification of MORL algorithms, following the common utility-based taxonomy from [48, 25]. Here, Π_{D} and Π_{DS} denote the policy space restricted to deterministic and deterministic stationary policies, respectively.

For each combination of criteria, this taxonomy allows us to define a *solution set*, i.e., the type of policies that will constitute the solution to our given problem. In table 1, we categorize a non-exhaustive list of popular MORL methods according to said taxonomy. In this section, we present each class of solution set alongside its corresponding methods.

4.1 Linear scalarization

When the utility function is linear, Roijers et al. [48] show that **deterministic stationary**[4] **policies are optimal**. Furthermore, adding non-stationarity and stochasticity greatly increases the size of the policy space. Thus, MORL methods developped for linear utility functions tend to limit their search to deterministic stationary policies. In scenarios where u is known, only a single optimal policy is required. Conversely, when the utility is unknown or may change, we seek to retrieve a convex coverage set.

Note that by definition, the SER and ESR optimization criteria are equivalent under linear utility, and as such no distinction is made between them in this section.

4.1.1 One deterministic stationary policy

When a linear utility function is used, any single policy MORL problem can be cast into single-objective RL by scalarizing the reward vector. This setting can be solved with most of the existing RL methods (e.g., value-based methods, policy gradients).

For example, take the autonomous driving example discussed in section 2. Let's assume our user is budget-conscious, not in a hurry, and has recurrent back pain. They might then decide on a preference (weight) vector of $[0.1, 0.5, 0.4]$, meaning that they assign an importance factor of 0.1 to speed, 0.5 to cost saving, and 0.4 to comfort. When driving towards a speed bump, the car can either brake or accelerate. The brake option yields a reward of $[-0.4, 0.4, 2.1]$ which gets scalarized to $0.1 \cdot -0.4 + 0.5 \cdot 0.4 + 0.4 \cdot 2.1 = 1$. Accelerating gives $[5, -0.2, -1]$, resulting in a scalarized reward of $u([5, -0.2, -1]) = 0$. This indicates that braking is to be favored in this context. When the agent receives a reward vector from the environment, single-objective RL methods like REINFORCE [51] or DQN [32] can scalarize it as such before using the resulting value as their reward input.

[4] A policy π is stationary if the distribution of actions is constant in all states, i.e., it is not conditional on time step-dependent information.

4.1.2 CCS of deterministic stationary policies

As mentioned in section 3.2, using a linear utility function implies that all optimal policies lie on a convex coverage set. This means that a multi-policy algorithm able to recover a CCS has access to an optimal policy for any possible weight vector $\mathbf{w}$.

Most algorithms use some form of neural network conditioned on a weight vector in their architecture and train it with random values, allowing the model to produce robust outputs over any input $\mathbf{w}$. Conditioned Networks (CN) [2] popularized this approach by showing the potential of conditioned deep Q-networks to generalize across the weight space. Following work kept the same general structure, while focusing on efficient exploration and alignment of weight vectors. The authors of Envelope [68] propose to use multiple schemes such as homotopy optimization and Hindsight Experience Replay [7] and show that it allows them to consistently outperform CN. PG-MORL [67] was one of the first methods to tackle environments with large continuous action spaces. It features an evolutionary stage that allows it to efficiently search the space of policies and weights to best improve the CCS. PD-MORL [9] was able to beat Envelope and PG-MORL (on discrete and continuous action tasks respectively) by adding a preference guidance term to a double deep Q-network loss [62]. Note that some of these works use the terms Pareto coverage sets and convex coverage sets interchangeably, but their nature in fact strictly limit them to the retrieval of CCS.

4.2 Monotically increasing scalarization

When the utility function is non-linear, **deterministic stationary policies are not guaranteed to be optimal**. To retrieve policies from the Pareto front that do not lie on convex coverage sets, we need to introduce either non-stationarity or stochasticity.

Note that in this context of non-linear scalarization functions, the ESR and SER optimization criteria are distinct. Although not explicitly mentioned here, each method presented in this section optimizes for one of them.

4.2.1 Deterministic non-stationary policies

When the solution policies must be deterministic and the utility function is non-linear, White shows that non-stationary policies can dominate stationary ones [65]. Consequently, it is necessary to consider non-stationary policies to retrieve a PCS in this context.

Imagine an autonomous delivery company working for two large clients A and B. Its goal is to distribute as many items as possible, while avoiding to neglect either A or B as not to lose an important partnership. An autonomous truck receives a reward of $[1, 0]$ when customer A gets a successful delivery, and $[0, 1]$ for customer B. The utility function to use could then be $u(\mathbf{V}^\pi) = \min(V_A^\pi, V_B^\pi)$, effectively maximizing the total number of deliveries while ensuring no client is left out. Here, a deterministic non-stationary policy would be able to yield a satisfying utility while a stationary one would not. Indeed, instead of always acting the same in each state — which would be equivalent to always picking the same client and thus yielding a utility of 0 — the non-stationary policy could condition on the time-dependent past rewards. This allows the agent to make informed decisions about actions to take depending on whether A or B was most chosen until now.

The first and third cells in the second row of table 1 respectively represent the single and multi-policy (PCS) solution sets for deterministic non-stationary policies. Constructing such policies is often done by conditioning them on the current timestep t (EUPG [46], PCN[42][5]), or by splitting $\mathbf{G}$ (see eq. 1) into past (also known as accrued) and future returns (PQL [34], EUPG [46], MO-CAC [43]). For example, the EUPG algorithm employs a modified policy gradient loss including both accrued rewards and a t-conditioned policy. Q-steering [60] takes another approach, forming non-stationary combinations of deterministic stationary base policies. Q-steering is based on Q-learning, and as such is limited to discrete state and action spaces.

4.2.2 Deterministic stationary mixture policies

As previously mentioned, there are contexts in which having a predictable, deterministic policy is essential. Conversely, other applications can tolerate some degree of stochasticity. For example, when designing a fleet of autonomous cars, we might want to add randomness to the path-finding algorithm, such that not all agents converge to the same road, thus avoiding congested traffic and globally sub-optimal behaviors.

When allowed, stochastic policies should be considered as part of the solution, as they can dominate deterministic policies under non-linear utility function [48]. It was shown that in some cases, we can construct a Pareto front from a mixture (i.e., a stochastic combination) of deterministic stationary policies [56, 63]. This is ideal, as it means that recovering a CCS is sufficient to construct the entire PF, greatly reducing the amount of computation needed to find optimal policies.

For example, Vamplew et al. [56] introduce a new algorithm, which we refer to as π-mix, that randomly selects a deterministic policy at the start of each episode and for its entire duration. Although this method works as expected under SER, using one deterministic policy per episode is not suitable for learning under ESR. Following our autonomous delivery example from section 4.2.1, π-mix could learn to alternate between two policies, each favoring only client A or B. In expectation over multiple episodes, this would indeed result in a fair delivery between them. However, on a per-episode basis, one customer would not be supplied, and thus could end the contract.

The ESR case is more complex, as the choice of policy needs to happen at each state (instead of each episode), being effectively equivalent to a stochastic policy. Wakuta [63] introduces a such method in a simplified setting, which we designate as S-rand, where the probability of picking one of k policy is the same at each state.

However, Lu et al. [29] show that finding the correct weights of a stochastic policy to retrieve a specific value vector is in practice infeasible. They propose CAPQL which uses reward augmentation to recover otherwise unreachable value functions from the Pareto front, although the resulting policies are not stochastic.

4.3 Challenges and way forward

As seen throughout this section, the field of multi-objective reinforcement learning, despite its growing popularity, remains sparse and fragmented. The recent work of Hayes et al. [25] identifies a few understudied areas of MORL that require further exploration: *complex multi-objective benchmarks*, dedicated *many-objectives methods*, specificities of *multi-agent settings* and the *dynamical identification and evolution of objectives*.

In particular, the study of many-objectives methods seems like an important future research area for MORL. Indeed, most MORL algorithms suffer from the *curse of dimensionality*, i.e., the exponential growth of the search space in the number of objectives makes retrieving satisfying policies highly complex. Note that the lack of MORL benchmarks has been partly addressed since the survey. Notably, the widely-used RL library *Gymnasium* was extended to the multi-objective case with *MO-Gymnasium* [21].

5 MORL and ethics

While it is important to take into account the normative ethics considerations mentioned in section 3.3, deploying MORL agents in society introduces additional concerns. Drawing from the machine ethics literature and considering potential issues caused by the use of naive MORL algorithms in real life scenarios, we identify four desirable features associated with ethical MORL agents.

They should have the ability to: (a) **prioritize user experience**, (b) **adapt to an evolving society**, (c) **adhere to a set of norms**, and (d) **account for other agents**. Interestingly, the evolution of objectives and the multi-agent aspect are part of the list of open challenges for MORL research mentioned in section 4.3. Note that these properties are pointers for researchers wanting to consider the impact of their algorithms, and not an exhaustive list of required attributes to develop agents with ethically-aligned behaviors. These features can even be contradictory in some cases, e.g., when a user's preferences are incompatible with the set of norms the agent ought to follow.

[5] Pareto Conditioned Networks can be seen as a sort of deterministic non-stationary policy method, as the agent follows a policy trained using supervised learning that conditions on the "desired horizon".

In this section, we define each of the aforementioned properties, review their place in the MORL literature, highlight potential future work, and conclude by discussing ways of benchmarking ethics in a MORL settings. A summarizing classification of existing methods according to our four principles is presented in table 2.

5.1 The user-centric approach

User-centric methods bring an **explicit consideration of the user** alongside the traditional performance goals. These approaches aim to empower users with agency, helping them to make informed decisions while minimizing their cognitive load. Algorithms mentioned in section 4 are capable of producing one policy (or a set of policies) that efficiently solves the input problem. However, most of them do not tackle how to find what utility function to use or which policy to pick from the Pareto front. Consequently, the end-user is tasked with making these decisions which can be non-trivial, for when the Pareto Front is not easily visualizable ($m > 3$). Etzioni and Etzioni [20] advocate for the *ethics bot*, an AI program that "extracts specific ethical preferences from a user and subsequently applies these preferences to the operations of the user's machine". This resonates with the example discussed in sec. 4.1.1 in which we want the agent to learn the passenger's preferences (e.g., prioritize speed if they are in a hurry or low costs if they want to save up) and adapt its driving profile accordingly. Zintgraf et al. [71] noticed this gap in the literature and made a first step to address it by proposing and evaluating several preference elicitation strategies. Following this work, a number of papers have focused on making the human decision maker a bigger part of the MORL process.

With GUTS [47], Roijers et al. introduce an interactive approach for multi-armed bandits, where the agent learns simultaneously about the environment and the user's preferences. Contrary to previous methods, GUTS is able to learn non-linear utility functions, while querying the user a provably limited number of times.

MORAL [40] proposes a two-step method for aligning an agent's behavior with the preferences of a user. First, a set of reward functions is learned from expert demonstrations using adversarial inverse reinforcement learning [22]. The user is then faced with multiple queries, allowing the agent to find a preference vector between expert reward functions, while simultaneously optimizing a policy on this combination. Empirically, the authors show that an adversarial user would not be able to teach the agent behaviors actively avoided by the expert demonstrations, although no formal proof is given. DWPI [30] learns the user's preference vector from demonstrations of their behavior in the environment (in a way reminiscent of inverse RL [70]). Chaput et al. [13] argue for a more contextual and intelligible approach, and propose QSOM-MORL, which learns to identify and solve ethical dilemmas using contextual human preferences.

Although not discussed in this work, it is important to consider potential biases in the construction of the utility function when developing single-policy user-centric algorithms. For example, some work (notably in the economics literature) show that there can be a gap between observed and ground truth preferences [10]. As MORL algorithms get better, this discrepancy may become a bottleneck in user satisfaction, further emphasizing the need to take these factors into account.

5.2 Evolving values and preferences

The methods for learning a user's preferences or utility function introduced in the previous section assume that this target is fixed and not subject to change. However, the owner of a self-driving vehicle, who usually favors comfort and savings over speed, may radically change their preferences in the case of an emergency. Similarly, the vehicle could be part of an autonomous taxi fleet, having to adapt to each customer profile. Therefore, it can be desirable for autonomous agents to have the ability to **detect and adapt to user preference changes**.

A few MORL methods have been developed to tackle this problem. CN [2] and DMCRL [37] take similar approaches, using prior information from learned policies to adapt to changing preferences. Q-steering [60] includes an interactive mode, allowing the user to update the target during or after the learning phase.

As society evolves, the three values proposed in our example of section 2 could fail to address emerging considerations such as environmental impact. Pavaloiu and Koose [39] emphasize that morality is subjective, varies across cultures, and continuously evolves. Thus, we may want our agent to **adapt to newly introduced objectives** while retaining previously learned knowledge. One naive way to approach this aspect could be to use a linear scalarization function, and take advantages of methods which support non-stationary reward functions (e.g., continual RL [27], Q(D)SOM [12]). Hayes et al. [25] identify the challenge of dynamic identification and addition of objectives as one of the main areas for future work in MORL, and to our knowledge the formulation of a variable sized vector-valued reward function has not been studied yet.

5.3 Lawful agents

Approaches for the ethical alignment of agents behavior can be categorized into 3 classes [4]:

- *Bottom-up* approaches do not enforce any obligatory or prohibited actions. Instead, the ethical behavior is learned through experience, and emerges from the definition of the agent and environment.
- *Top-down* approaches are rule-based, and incorporate a priori knowledge (such as deontological duties).
- Some works [52, 17] argue for *hybrid* methods which combine the top-down and bottom-up approaches.

When discussing their ethics bots, Etzioni and Etzioni [20] mention that they only address moral preferences, and disregard normative aspects (e.g., a legal framework). Thus, a MORL-based implementation of an ethics bot would only learn in a bottom-up fashion. Although some works

MORL methods	user-centered	adaptable	normative	multi-agent
CN [2], DMCRL [37], Q-steering [60]	✓	✓		
MAEE [44]			✓	✓
GUTS [47], MORAL [40], DWPI [30], QSOM-MORL [13]	✓			
EE [45], TLO [59]			✓	
MO-MIX [26], PRBS/D [31], moral rewards [53]				✓

Table 2: Qualification of MORL methods with regards to ethical properties.

[64, 53] argues that top-down approaches are challenging and pose some risks, having a set of guarantees (via top-down or hybrid agents) can be crucial in some applications. Typically, we want to ensure that self-driving vehicles deployed on real roads act according to the locally enforced traffic regulations, so that their behavior is safe and predictable for human drivers. In fact, Pagallo [38] argues that values alone are not enough for the coordination of AI agents and that rules are needed. Thus, it is desirable for our agents to **be able to follow a set of norms**.

In MORL, Rodriguez-Soto et al. [45] take the perspective of the environment designer, allowing them to derive theoretical guarantees for the alignment of agents w.r.t. chosen ethical values. To do so, they start from a MOMDP whose reward functions are built upon a value system. Their proposed Multi-Valued Ethical Embedding (EE) algorithm then proceeds to compute a solution weight vector, resulting in a linearly scalarized MDP with the desired properties.

Using potential-based rewards, TLO [59] focuses on impact-minimizing agents, i.e., agents performing a primary task while aiming at disrupting the environment as little as possible. This approach is bottom-up by design, yet the authors demonstrate strong empirical results showing the ethical alignment of trained agents. These results are for now limited to discrete states and actions, although the algorithms proposed are theoretically extensible to the continuous cases.

For single-objective RL, a few works propose top-down or hybrid approaches. Shielding [5] uses temporal logic to enforce a set of properties on the resulting policy. AJAR [3] uses argumentation-based judges to compute the rewards based on a set of moral values. Extending such methods to the multi-objective case presents promising possibilities for future research.

5.4 Ethics as a multi-agent problem

Murukannaiah et al. [36] argue that the study of ethics intrinsically needs to be done in a multi-agent context, highlighting that research in AI ethics is to this day largely constituted of single-agent works and ignores the societal context. As trained MORL algorithms are deployed a real-life situations, they are likely to encounter other actors, both artificial and human. Therefore, we argue that our agents should **be able to account for and interact with other actors**. The field of multi-objective multi-agent reinforcement learning (MOMARL) accounts by design for the interactions that can emerge in these cases. Being at the intersection of two sub-fields, MOMARL remains relatively understudied. Rădulescu et al. [41] have surveyed the field of multi-objective multi-agent decision making and concluded that many gaps still exist in the literature, particularly for RL-based methods. Although some MOMARL approaches have been proposed [26, 31], and there has been work on ethics in the multi-agent setting [15], very few MOMARL papers specifically take an ethical perspective. Rodriguez-Soto et al. [44] propose a method (MAEE) to construct environments in which agents are guaranteed to have an ethically-aligned behavior, while pursuing their individual goals. However, the multi-objective reward function they use is very simple, with only two component: an individual objective and an ethical objective (itself split between a normative and evaluative part). QSOM and QDSOM [12] are multi-agent algorithms based on self-organizing maps. Although not multi-objective, they were tested with various reward functions combining ethical stakes, analogously to ESR-optimized MORL. Tennant et al. [53] analyze the behavior of intrinsically-motivated RL agents rewarded according to moral theories when faced with moral dilemmas.

5.5 Benchmarking ethics

While some papers tackle the evaluation of MORL algorithms and the available benchmarks [57], few environments have become standard, and most of them are too simple for modern methods [25].

When trying to ensure the ethical alignment of an AI agent's behavior, the metric of success may be more complex than a simple sum of reward signals. Few MORL environments with an ethics-first approach have been proposed. The ethical gathering game by Rodriguez-Soto et al. [44] extends the regular gathering game, with the addition of beneficence as a moral value. Scheirlinck et al. [49] introduce the ethical smart grid, a complex environment with continuous actions and observations. They propose to use a number of (sometimes conflicting) moral values from the literature to evaluate the behavior of agents.

Additionally, there is a number of environments which are not created with ethics in mind but allow for the inclusion of one or more of the constraints previously mentioned. As such, any MORL environment (e.g., DST [61]) can be viewed through a user-centric lens by changing the setting or adding queries to a user to learn their preferences. Similarly, we can modify multi-agent multi-objective environ-

ments (e.g., MOBDP [31]) to shift the focus towards the alignment of agents with some specified ethical values.

6 Conclusion

As artificial intelligence agents are being increasingly deployed in society, there is a growing need to study ways of ensuring the ethical alignment of their behaviors. In this paper, we have focused on multi-objective reinforcement learning, a framework that has been deemed ideal for modeling the complexities of both ethics and real-world problems. First, we proposed a classification of existing multi-objective RL methods according to the prevalent taxonomy. Then, we explored the considerations required when one wishes to work in MORL while adopting an ethics-centered perspective. The literature at the intersection of MORL and ethics is still very limited, and a lot of work remains to be done, notably on methods explicitly implementing one or more of the four desirable properties for ethical agents highlighted in section 5: adherence to user preferences, adaptability to societal changes, compliance with norms and regulations, and considerations of other agents. We hope that this work can serve researchers at the intersection of MORL and ethics to visualize the state of current research and the still lacking areas deserving of further investigations.

Acknowledgements

This work was funded by ANR project ACCELER-AI (ANR-22-CE23-0028-01).

References

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

[2] Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic Weights in Multi-Objective Deep Reinforcement Learning. In *ICML*, 2019.

[3] Benoît Alcaraz, Olivier Boissier, Rémy Chaput, and Christopher Leturc. Ajar: An argumentation-based judging agents framework for ethical reinforcement learning. In *AAMAS*, 2023.

[4] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 2005.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI*, 2018.

[6] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 2007.

[7] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight Experience Replay. In *NeurIPS*, 2018.

[8] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *ICML*, 2020.

[9] Toygun Basaklar, Suat Gumussoy, and Umit Y. Ogras. PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm. In *ICLR*, 2023.

[10] John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. How are preferences revealed? *Journal of public economics*, 2008.

[11] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on SMC*, 2008.

[12] Rémy Chaput, Olivier Boissier, and Mathieu Guillermin. Adaptive reinforcement learning of multi-agent ethically-aligned behaviours: the QSOM and QDSOM algorithms. *arXiv e-prints*, 2023.

[13] Rémy Chaput, Laetitia Matignon, and Mathieu Guillermin. Learning to identify and settle dilemmas through contextual user preferences. In *ICTAI*, 2023.

[14] Jianguo Chen, Kenli Li, Keqin Li, Philip S Yu, and Zeng Zeng. Dynamic bicycle dispatching of dockless public bicycle-sharing systems using multi-objective reinforcement learning. *ACM TCPS*, 2021.

[15] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. In *AAMAS*, 2016.

[16] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 2022.

[17] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. 2019.

[18] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S Kließ, Maite Lopez-Sanchez, et al. Ethics by design: Necessity or curse? In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[19] Muhammad Diyan, Bhagya Nathali Silva, and Kijun Han. A multi-objective approach for optimal energy management in smart home using the reinforcement learning. *Sensors*, 2020.

[20] Amitai Etzioni and Oren Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 2017.

[21] Florian Felten, Lucas Nunes Alegre, Ann Nowe, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno Castro da Silva. A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[22] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *ICLR*, 2018.

[23] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 2022.

[24] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[25] 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. *AAMAS*, 2022.

[26] Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making With Deep Reinforcement Learning. *IEEE PAMI*, 2023.

[27] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *JAIR*, 2022.

[28] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[29] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-Objective Reinforcement Learning: Convexity, Stationarity and Pareto Optimality. In *ICLR*, 2022.

[30] Junlin Lu, Patrick Mannion, and Karl Mason. Inferring Preferences from Demonstrations in Multi-objective Reinforcement Learning: A Dynamic Weight-based Approach. In *ALA (AAMAS)*, 2023.

[31] Patrick Mannion, Sam Devlin, Jim Duggan, and Enda Howley. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning. *The Knowledge Engineering Review*, 2018.

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

[33] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 2023.

[34] Kristof Van Moffaert and Ann Nowé. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *JMLR*, 2014.

[35] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 2006.

[36] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. New Foundations of Ethical Multiagent Systems. In *AAMAS*, 2020.

[37] Sriraam Natarajan and Prasad Tadepalli. Dynamic preferences in multi-criteria reinforcement learning. In *ICML*, 2005.

[38] Ugo Pagallo et al. Even angels need the rules: Ai, roboethics, and the law. In *ECAI*, 2016.

[39] Alice Pavaloiu and Utku Kose. Ethical artificial intelligence-an open question. *JOMUDE*, 2017.

[40] Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. In *AAMAS*, 2022.

[41] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. Multi-objective multi-agent decision making: A utility-based analysis and survey. In *AAMAS*, 2020.

[42] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto Conditioned Networks. In *AAMAS*, 2022.

[43] Mathieu Reymond, Conor F. Hayes, Denis Steckelmacher, Diederik M. Roijers, and Ann Nowé. Actor-critic multi-objective reinforcement learning for non-linear utility functions. In *AAMAS*, 2023.

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

[45] Manel Rodriguez-Soto, Roxana Rădulescu, Juan A Rodriguez-Aguilar, and Maite Lopez-Sanchez. Multi-objective reinforcement learning for guaranteeing alignment with multiple values. In *ALA (AAMAS)*, 2023.

[46] Diederik Roijers, Denis Steckelmacher, and Ann Nowe. Multi-objective Reinforcement Learning for the Expected Utility of the Return. In *ALA (AAMAS)*, 2018.

[47] Diederik M. Roijers, Luisa M. Zintgraf, Pieter Libin, and Ann Nowé. Interactive multi-objective reinforcement learning in multi-armed bandits for any utility function. In *ALA (AAMAS)*, 2020.

[48] Diederik Marijn Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A Survey of Multi-Objective Sequential Decision-Making. *JAIR*, 2013.

[49] Clément Scheirlinck, Rémy Chaput, and Salima Hassas. Ethical Smart Grid: A Gym environment for learning ethical behaviours. *JOSS*, 2023.

[50] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

[51] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 1999.

[52] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Learning machine morality through experience and interaction. 2023.

[53] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *IJCAI*, 2023.

[54] Mark Timmons. *Moral theory: An introduction*. Rowman & Littlefield publishers, 2012.

[55] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 2021.

[56] Peter Vamplew, Richard Dazeley, Ewan Barker, and Andrei Kelarev. Constructing Stochastic Mixture Policies for Episodic Multiobjective Reinforcement Learning Tasks. In *Advances in Artificial Intelligence*. 2009.

[57] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 2011.

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

[59] 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*, 2021.

[60] Peter Vamplew, Rustam Issabekov, Richard Dazeley, Cameron Foale, Adam Berry, Tim Moore, and Douglas Creighton. Steering approaches to pareto-optimal multiobjective reinforcement learning. *Neurocomputing*, 2017.

[61] Peter Vamplew, John Yearwood, Richard Dazeley, and Adam Berry. On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Advances in Artificial Intelligence*, 2008.

[62] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, 2016.

[63] Kazuyoshi Wakuta. A note on the structure of value spaces in vector-valued Markov decision processes. *Mathematical Methods of Operations Research*, 1999.

[64] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[65] D. J White. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 1982.

[66] Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. The societal implications of deep reinforcement learning. *JAIR*, 2021.

[67] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control. *ICML*, 2020.

[68] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation. In *NeurIPS*, 2019.

[69] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM CSUR*, 2021.

[70] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[71] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. Ordered preference elicitation strategies for supporting multi-objective decision making. *AAMAS*, 2018.