On the Limitations of Markovian Rewards to Express Multi-Objective, Risk-Sensitive, and Modal Tasks

Abstract

In this paper, we study the expressivity of scalar, Markovian reward functions in Reinforcement Learning (RL), and identify several limitations to what they can express. Specifically, we look at three classes of RL tasks (multi-objective RL, risk-sensitive RL, and modal RL), and show that most of the instances in each of these three classes cannot be expressed using scalar, Markovian rewards. Among these three classes, we provide necessary and sufficient conditions for when a problem can be reduced to ordinary, scalar reward reinforcement learning. Modal problems have so far not been given any systematic treatment in the RL literature; we thus call attention to them as a new class of problems. Finally, we also show that many of these problems *can* be solved by means of bespoke RL approaches: this rules out the possibility that those problems that cannot be expressed using Markovian reward functions also are impossible to learn effectively.

1 INTRODUCTION

To solve a task with reinforcement learning (RL), we must first encode that tasks as a reward function [Sutton and Barto, 2018]. Typically, these rewards are *scalar* and *Markovian*. In this paper, we study some limitations in the expressivity of scalar and Markovian reward functions. Specifically, we will examine three broad classes of tasks, all of which are both intuitive to understand, and useful in practical situations. We will then show that *almost all* tasks in each of these three classes cannot be expressed using Markovian reward functions. This suggests that scalar, Markovian reward functions are semantically limited, and raises a warning for their indiscriminate use. Moreover, we also show that many of these problems *can* be solved effectively with RL algorithms. This rules out the possibility that those problems that cannot be expressed using scalar, Markovian reward functions also are impossible to learn effectively in RL.

The first class of problems we look at, in Section 3, are single-policy, multi-objective RL tasks (MORL). In such problems, the agent receives multiple reward signals, and the aim is to learn a single policy that achieves an optimal trade-off amongst those rewards, according to some specified criterion [Roijers et al., 2013, Liu et al., 2015]. For example, a single-policy MORL algorithm might attempt to maximise the rewards lexicographically [Skalse et al., 2022b]. We will provide necessary and sufficient conditions describing when a MORL problem can be reduced to scalar-reward RL, by providing a single reward function that induces the same preferences as the MORL problem. We find that this can *only* be done for MORL problems that correspond to a linear weighting of the rewards, which means that it cannot be done for the vast majority of all interesting MORL problems.

The next class of problems we study, in Section 4, is risks-sensitive RL. In expected utility theory, risk-aversion is often modelled using utility functions that are concave in some of their variables. We will show that these tasks cannot be expressed as Markovian reward functions, by demonstrating that no non-affine monotonic transformations of the trajectory return function are possible. This demonstrates another limitation in the expressive power of Markovian rewards.

In Section 5, we introduce a new class of tasks, which we call *modal* tasks. These are tasks where the agent is evaluated not only based on what (distribution of) trajectories that it generates, but also based on what it *could have done* along those trajectories. As an example, consider the instruction "you should always be *able* to return to the start state". We provide a formalisation of such tasks, argue that there are many situations in which these tasks could be useful, and finally prove that these tasks also typically cannot be formalised using ordinary reward functions.

In Section 6, we discuss how to solve tasks from each of these classes using specialised RL solutions: we provide

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

references to existing literature, and also sketch both an approach for learning a wide class of MORL problems, and an approach for learning a wide class of modal problems. Finally, in Section 7, we discuss the implications of our results, together with several pieces of related work.

2 PRELIMINARIES

The standard RL setting is formalised using *Markov Decision Processes* (MDPs) Sutton and Barto [2018], which are tuples $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$ where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\tau : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function, μ_0 is an initial state distribution over $\mathcal{S}$, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, and $\gamma \in (0, 1)$ is a discount factor. A *trajectory* ξ is in general an element of $(\mathcal{S} \times \mathcal{A})^\omega$, i.e. a sequence $s_0, a_0, s_1 \ldots$. We use G to denote the *trajectory return function*, where $G(\xi) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$. A *policy* is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and Π is the set of all policies. Given a policy π, its *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ is the function where $V^\pi(s)$ is the expected future discounted reward when following π from s, and its *Q-function* Q^π is $\mathbb{E}_{s' \sim \tau(s,a)}[R(s, a) + \gamma \cdot V^\pi(s')]$. The *policy evaluation function* $J : \Pi \to \mathbb{R}$ is $J(\pi) = \mathbb{E}_{s_0 \sim \mu_0}[V^\pi(s_o)]$. If a policy maximises J, then we say that this policy is *optimal*. We denote optimal policies by $\pi^\star$, and their value function and Q-function by $V^\star$ and $Q^\star$. Moreover, given an MDP $\mathcal{M}$, we say that $\mathcal{M}$'s policy order is the ordering $\prec$ on Π where $\pi_1 \prec \pi_2 \iff J(\pi_1) < J(\pi_2)$ for any π_1, π_2.

In this paper, we will say that a reward function R is *trivial* if $J(\pi_1) = J(\pi_2)$ for all π_1, π_2. Moreover, we say that R_1 and R_2 are *equivalent* if $J_1(\pi_1) < J_1(\pi_2) \iff J_2(\pi_1) < J_2(\pi_2)$ for all π_1, π_2, and that they are *opposites* if $J_1(\pi_1) < J_1(\pi_2) \iff J_2(\pi_1) > J_2(\pi_2)$ for all π_1, π_2.

MORL problems are formalised using *Multi-Objective MDPs* (MOMDPs), which are tuples $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \mathbf{R}, \gamma \rangle$, with the only difference from MDPs being $\mathbf{R}$, which is now a function $\mathbf{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^k$ that, for each pair (s, a), returns k different rewards (for some finite k). We denote the i'th component of $\mathbf{R}$ as the scalar reward function R_i, and use V_i^π, Q_i^π, J_i, and G_i, etc, to refer to its value-, Q-, evaluation-, and return function, etc. There are two types of MORL problems; single-policy MORL, where the goal is to compute one policy that achieves an optimal trade-off of the rewards, and multi-policy MORL, where the aim is to compute several policies (typically with the aim of approximating the Pareto front of the rewards). In this paper, we are concerned with single-policy MORL. Since there may not be a single policy that maximises each component of $\mathbf{R}$, a single-policy MORL problem needs some additional rule for combining and trading off each reward.

In economics and psychology, risk-aversion is often modelled using utility functions $U(c)$ that are concave in some relevant variable c. The most common risk-averse utility functions are the *exponential*, the *isoelastic*, and the *quadratic* utility functions. The exponential utility function is given by $U(c) = -e^{\alpha c}$, where $\alpha > 0$ is a parameter controlling the degree of risk aversion. The isoelastic utility function is given by $U(c) = c^{1-\alpha}$, for $\alpha > 0, \alpha \neq 1$, or by $U(c) = \ln(c)$ (corresponding to the case when $\alpha = 1$). The quadratic utility function is given by $U(c) = c - \alpha c^2$, where $\alpha > 0$. Since this function is decreasing for sufficiently large c, its domain is typically restricted to $(-\infty, 1/2\alpha]$.

A Remark on "Tasks": In this paper, we are investigating the question of when a given task can be expressed using a scalar, Markovian reward function. To do this, we must first formalise what it should mean for a reward function to express a task. One option is to say that a task corresponds to a desired policy π, and that a reward function R expresses the task if π is optimal under R (possibly with the additional requirement that π is the *only* policy that is optimal under R). With this definition, we find that *any* task can be expressed as a Markovian reward function, at least as long as π is stationary and deterministic (see Appendix B). With this definition, the problem is therefore rather trivial.

An alternative, stronger formalisation is to say that a task corresponds to an ordering $\prec$ on Π, which encodes a preference ordering over all policies, and that a reward function R expresses the task if its corresponding evaluation function J orders Π according to $\prec$. It is primarily this latter definition that we will use in this paper. The main reason for this is that it is often impossible to find the optimal policy in complex environments. For example, in a robotics problem, it is typically not feasible to find a policy that is globally optimal. This means that it is not enough for R to admit the correct optimal policy; it must also induce the right preferences between the all the (sub-optimal) policies that the policy synthesis algorithm might in fact generate. The only way to robustly ensure that this is the case is if R induces the right policy ordering. For this reason, we think it is more informative to think of a problem setting (i.e. a "task") as corresponding to an ordering on Π.

3 MULTI-OBJECTIVE PROBLEMS

In this section, we examine the MORL setting. We first need a general definition of what a single-policy MORL problem is. Recall that a MOMDP $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \mathbf{R}, \gamma \rangle$ by itself has no one canonical objective to maximise. We therefore introduce the notion of a *MORL objective*:

Definition 1. *A **MORL objective** over k rewards is a function $\mathcal{O}$ that takes k policy evaluation functions $J_1 \ldots J_k$ and returns a (total) ordering $\prec_\mathcal{O}$ over the set of all policies Π.*

Given a MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \mathbf{R}, \gamma \rangle$, a MORL objective $\mathcal{O}$ gives us an ordering over Π that tells us when a policy is preferred over another. We use $\prec_\mathcal{O}^\mathcal{M}$ to denote the

ordering that is obtained when we apply $\mathcal{O}$ to $\mathcal{M}$'s policy evaluation functions. For the purposes of this paper, we will not need to impose any further requirements on $\prec_{\mathcal{O}}$. For example, we will not insist that $\prec_{\mathcal{O}}$ must have a greatest element in Π, or that $\pi_1 \prec_O \pi_2$ whenever π_2 is a Pareto improvement over π_1, etc, even though a reasonable MORL objective presumably would have these properties. We next provide a few examples of MORL objectives, where we denote by π_1, π_2 any given pair of distinct policies.

Definition 2. *Given $J_1 \dots J_k$, the **LexMax** objective $\prec_{\texttt{Lex}}$ is given by $\pi_1 \prec_{\texttt{Lex}} \pi_2$ iff there is an $i \in \{1 \dots k\}$ such that $J_i(\pi_1) < J_i(\pi_2)$ and $J_j(\pi_1) = J_j(\pi_2)$ for $j < i$.*

Definition 3. *Given $J_1 \dots J_k$, the **MaxMin** objective $\prec_{\texttt{Min}}$ is given by $\pi_1 \prec_{\texttt{Min}} \pi_2 \iff \min_i J_i(\pi_1) < \min_i J_i(\pi_2)$.*

Definition 4. *Given $J_1 \dots J_k$ and some $c_1 \dots c_m \in \mathbb{R}$, the **MaxSat** objective $\prec_{\texttt{Sat}}$ is given by $\pi_1 \prec_{\texttt{Sat}} \pi_2$ if and only if the number of rewards that satisfy $J_i(\pi_1) \geq c_i$ is larger than the number of rewards that satisfy $J_i(\pi_2) \geq c_i$.*

Definition 5. *Given J_1, J_2 and some $c \in \mathbb{R}$, the **ConSat** objective $\prec_{\texttt{Con}}$ is given by $\pi_1 \prec_{\texttt{Con}} \pi_2$ if and only if either $J_1(\pi_1) < c$ and $J_1(\pi_1) < J_1(\pi_2)$, or $J_1(\pi_1), J_1(\pi_2) \geq c$ and $J_2(\pi_1) < J_2(\pi_2)$.*

In other words, the LexMax objective has *lexicographic* preferences over $R_1 \dots R_m$, so that policies are first ordered by their expected discounted R_1-reward, and then policies that obtain the same expected discounted R_1-reward are ordered by their expected discounted R_2-reward, and so on. The MaxMin objective orders policies by their *worst* performance according to any of $R_1 \dots R_m$ (which could be used to obtain worst-case guarantees). The MaxSat objective only cares whether a policy reaches a certain *threshold* for each reward, and ranks policies based on how many thresholds they reach. The ConSat objective aims to maximise J_2, but under the constraint that J_1 reaches a certain threshold. Note that these objectives are not necessarily the most important MORL objectives. Rather, they are simply a short list of illustrative examples, meant to demonstrate the flexibility of the MORL framework, and give an intuition for what types of problems it can be used to express. A few more examples can be found in Appendix C.

We next define what it means to *reduce* a MORL problem to a scalar RL problem:

Definition 6. *A MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \boldsymbol{R}, \gamma \rangle$ with objective $\mathcal{O}$ is **equivalent** to the MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$ if and only if $\mathcal{M}'$'s policy order is $\prec_{\mathcal{O}}^{\mathcal{M}}$. We then say that $\mathcal{M}$ with $\mathcal{O}$ is **scalarized** by R. If $\mathcal{M}$ with $\mathcal{O}$ is scalarized by some R then we say that $\mathcal{M}$ with $\mathcal{O}$ is **scalarizable**, otherwise we say that it is **unscalarizable**.*

Note that $\mathcal{M}'$ must have the same states, actions, transition function, initial state distribution, and discount factor, as $\mathcal{M}$. This definition therefore says that $\mathcal{M}$ with $\mathcal{O}$ is equivalent to $\mathcal{M}'$ if $\mathcal{M}'$ is given by replacing $\mathbf{R} = \langle R_1 \dots R_k \rangle$ with a single reward function R, and R induces the same preferences between all policies as $\mathcal{O}(J_1 \dots J_k)$. Note also that we require R to express the same *policy order* as $\mathcal{O}(J_1 \dots J_k)$; it is not enough for R and $\mathcal{O}(J_1 \dots J_k)$ to have the same *optimal policies* (see Section 2).

Given this definition, we can now provide the necessary and sufficient conditions for when a MORL problem can be reduced to a scalar-reward RL problem. All proofs are provided in the supplementary material.

Theorem 1. *If a MOMDP $\mathcal{M}$ with objective $\mathcal{O}$ is scalarizable, then there exist $w_1 \dots w_k \in \mathbb{R}$ such that $\mathcal{M}$ with $\mathcal{O}$ is scalarized by the reward $R(s,a) = \sum_{i=1}^{k} w_i \cdot R_i(s,a)$.*

Theorem 1 tells us that a MORL objective can be expressed using a scalar, Markovian reward function if and only if that objective corresponds to a linear weighting of the individual rewards. In other words, scalar, Markovian rewards are unable to express all non-linear MORL problems. As we will see, this imposes a strong limitation on what MORL tasks can be encoded using scalar, Markovian rewards.

Theorem 1 also entails the following corollary, which is useful to elucidate when a MORL objective cannot be expressed using scalar reward functions. Given an ordering $\prec$ over Π, depending on some evaluation functions $J_1 \dots J_k$, we say that a function $U : \Pi \to \mathbb{R}$ *represents* $\prec$ if $U(\pi_1) < U(\pi_2) \iff \pi_1 \prec \pi_2$. We say that U is a *linear representation* if $U(\pi) = f(\sum_{i=1}^{k} w_i \cdot J_i(\pi))$ for some $w_1 \dots w_k \in \mathbb{R}$ and some strictly monotonic f.

Corollary 1. *If $\mathcal{O}(J_1 \dots J_k)$ has a non-linear representation U, and $\mathcal{M}$ is a MOMDP whose J-functions are $J_1 \dots J_k$, then $\mathcal{M}$ with $\mathcal{O}$ is unscalarizable.*

Therefore, we can prove that $\mathcal{M}$ with $\mathcal{O}$ is unscalarizable by finding a non-linear representation of $\prec_{\mathcal{O}}^{\mathcal{M}}$. Accordingly, we now show that none of the MORL objectives given in Definitions 2-5 can be expressed using scalar, Markovian reward functions, except in a few degenerate cases.

Corollary 2. *$\mathcal{M}$ with **LexMax** is unscalarizable, as long as $\mathcal{M}$ has at least two reward functions that are neither trivial, equivalent, or opposite.*

Note that if all reward functions are either trivial, equivalent, or opposite, then the only reward function that matters for **LexMax** is the highest-priority non-trivial reward function. In that case, $\mathcal{M}$ with **LexMax** is equivalent to the MDP which contains only this reward function.

Corollary 3. *$\mathcal{M}$ with **MaxMin** is unscalarizable, unless $\mathcal{M}$ has a reward function R_i such that $J_i(\pi) \leq J_j(\pi)$ for all $j \in \{1 \dots k\}$ and all π.*

Note that if $\mathcal{M}$ has a reward function R_i such that $J_i(\pi) \leq J_j(\pi)$ for all j and π, then this is the only reward function that matters for the **MaxMin** objective. In that case, $\mathcal{M}$ with **MaxMin** is equivalent to the MDP which contains only R_i.

Corollary 4. *$\mathcal{M}$ with **MaxSat** is unscalarizable, as long as $\mathcal{M}$ has at least one reward R_i where $J_i(\pi_1) < c_i$ and $J_i(\pi_2) \geq c_i$ for some $\pi_1, \pi_2 \in \Pi$.*

Note that if $\mathcal{M}$ has no reward R_i where $J_i(\pi_1) < c_i$ and $J_i(\pi_2) \geq c_i$ for some $\pi_1, \pi_2 \in \Pi$, then either all policies satisfy all constraints, or no policy satisfies any constraint. In either case, $\mathcal{M}$ with **MaxSat** would be equivalent to an MDP with a trivial reward function.

Corollary 5. *$\mathcal{M}$ with **ConSat** is unscalarizable, unless either R_1 and R_2 are equivalent, or $\max_\pi J_1(\pi) \leq c$, or $\min_\pi J_1(\pi) \geq c$.*

Note that if $\max_\pi J_1(\pi) \leq c$ then no policy satisfies the constraint, in which case $\mathcal{M}$ with **ConSat** is equivalent to the MDP with R_1. If $\min_\pi J_1(\pi) \geq c$ then all policies satisfy the constraint, in which case $\mathcal{M}$ with **ConSat** is equivalent to the MDP with R_2. If R_1 and R_2 are equivalent, then $\mathcal{M}$ with **ConSat** is scalarized by R_1 and R_2.

Corollaries 2-5 thus show that none of the MORL objectives given in Definition 2-5 can be expressed using a scalar, Markovian reward function, except in a few degenerate cases where those MORL objectives are trivialised. This demonstrates that most MORL problems cannot be scalarized in a satisfactory way.

To get an intuition for this result, note that the expected cumulative return of a Markovian reward function always is maximised by some stationary policy, whereas some of these MORL objectives may require the optimal policy to be non-stationary. For example, consider the **MaxMin** objective, and suppose the agent can choose between an action giving one R_1-reward, and an action giving one R_2-reward. Then the optimal choice may depend on how much R_1 and R_2-reward the agent has got in the past. This means that the optimal policy may be non-stationary, and thus not correspond to any Markovian reward.

4 RISK-SENSITIVE PROBLEMS

The next area we will look at is that of *risk-sensitive* RL. An ordinary RL agent tries to maximise the *expectation* of its reward function. However, there are many cases where it is natural to require the agent to be *risk-averse*. For example, we might prefer a policy that reliably achieves 5 reward, over one that achieves 11 reward with probability 0.5, and otherwise gets 0 reward, even though the latter policy achieves a higher expected reward. In this section, we will examine when scalar, Markovian reward functions can be used to encourage such behaviour.

In expected utility theory, risk-aversion is often modelled using concave utility functions. In particular, suppose we have a set of *outcomes* C, each of which is associated with some utility via a function $U_1 : C \to \mathbb{R}$. We can then construct a second utility function $U_2 : C \to \mathbb{R}$ by letting $U_2(c) = f(U_1(c))$ for some concave function f. Then, an agent which maximises expected utility according to U_2, will be risk-averse with respect to utility as defined by U_1. For example, suppose each outcome is associated with some monetary payoff, and that U_1 measures how much money is obtained in each outcome. If we were to maximise expected utility according to U_1, then we would prefer a 50% chance of obtaining $\$2,000,000$, to a certain chance of obtaining $\$900,000$. However, in the real world, most people would prefer the latter option. One reason for this is that, while getting $\$2,000,000$ is better than getting $\$900,000$, it is less than twice as good. We can model these preferences by using a second utility function U_2 that is concave in U_1. Intuitively, U_2 should measure how much *benefit* we get from the money. Then the expected U_2-utility might be higher for the safe option than the risky option, even though the expected U_1-utility is higher for the risky option.

In reinforcement learning, the *outcomes* are the trajectories that might occur in the environment, and the *utility* of a trajectory ξ is induced by the reward function R via the return function, G. If the transition function τ is nondeterministic, then the agent cannot reliably enact a particular outcome (i.e., move along a particular trajectory), but can instead only choose between some distributions over outcomes. By default, the agent may then be compelled to pursue a policy that achieves a high reward with small probability, as long as the expectation remains high. A natural question is then whether we could avoid this by constructing a second reward function that is concave in the original reward function, similar to what is done in expected utility theory. That is, given a reward function R_1 and a concave function f, can we construct a second reward function R_2 such that $G_1 = f(G_1)$? Our next theorem demonstrates that this is impossible. As before, the proof is in Appendix A.

Theorem 2. *Given $\mathcal{S}$, $\mathcal{A}$, and γ, let R_1 and R_2 be two reward functions. If $\gamma \geq 0.5$, and for all $\xi_1, \xi_2 \in (\mathcal{S} \times \mathcal{A})^\omega$,*

$$G_1(\xi_1) \leq G_1(\xi_2) \iff G_2(\xi_1) \leq G_2(\xi_2),$$

then $\exists a \in \mathbb{R}$, $b \in \mathbb{R} > 0$ such that for all $\xi \in (\mathcal{S} \times \mathcal{A})^\omega$,

$$G_1(\xi) = b \cdot G_2(\xi) + a.$$

Theorem 2 effectively tells us that *only affine transformations of G are possible*. From this result, it straightforwardly follows that none of the standard risk-averse utility functions (exponential utility, isoelastic utility, and quadratic utility) can be expressed using Markovian reward functions:

Corollary 6. *For any non-trivial reward R_1 and any constant $\alpha \neq 0$, if $\gamma \geq 0.5$ then there is no reward R_2 such that $G_2(\xi) = -e^{\alpha G_1(\xi)}$ for all $\xi \in (\mathcal{S} \times \mathcal{A})^\omega$.*

Corollary 7. *For any non-trivial reward R_1 and any constant $\alpha > 0$, $\alpha \neq 1$, if $\gamma \geq 0.5$ then there is no reward R_2 such that $G_2(\xi) = G_1(\xi)^{1-\alpha}$ for all $\xi \in (\mathcal{S} \times \mathcal{A})^\omega$.*

Corollary 8. *For any non-trivial reward R_1, if $\gamma \geq 0.5$ then there is no reward R_2 such that $G_2(\xi) = \ln(G_1(\xi))$ for all $\xi \in (\mathcal{S} \times \mathcal{A})^\omega$.*

Corollary 9. *For any non-trivial reward R_1 and any $\alpha > 0$ where $\max_\xi G_1(\xi) \leq \frac{1}{2\alpha}$, if $\gamma \geq 0.5$ then there is no reward R_2 such that $G_2(\xi) = G_1(\xi) - \alpha G_1(\xi)^2$ for all $\xi \in (\mathcal{S} \times \mathcal{A})^\omega$.*

Theorem 2 thus implies that none of the standard risk-averse utility functions can be expressed using scalar, Markovian reward functions. To get an intuition on Theorem 2, consider the fact that the expected cumulative return of a Markovian reward function always is maximised by some stationary (i.e. Markovian) policy. However, a risk-averse objective may require the optimal policy to be non-stationary, because whether or not it is worth taking a particular gamble could depend on how much reward you have accrued in the past. This suggests that there should be instances where risk-sensitive objectives cannot be expressed as Markovian reward functions. Theorem 2 formalises this intuition.

It is also worth remarking on the fact that Theorem 2 considers the value of G_1 and G_2 for *all* trajectories in $(\mathcal{S} \times \mathcal{A})^\omega$. For any particular transition function τ, most of these trajectories are likely to be impossible (unless τ allows you to transition between any two states via any action with non-zero probability). We could therefore alternatively consider the condition where $G_1(\xi_1) \leq G_1(\xi_2) \iff G_2(\xi_1) \leq G_2(\xi_2)$ for those trajectories ξ_1, ξ_2 that are possible in a given environment. In this case, it *can* be possible for G_2 to be non-affine in G_1. For example, consider the case of a tree-shaped MDP, where $\tau(s, a) = s$ and $R_1(s, a) = 0$ for all actions a if s is a leaf-node. In that case, G_2 can be an arbitrary transformation of G_1. However, to construct the corresponding reward function R_2, we would need to have a detailed understanding of the environment (which is against the main tenet in RL), and furthermore the resulting reward function would no longer induce the same behaviour if it were used in a different environment. For this reason, we believe that it is more relevant to consider the set of all trajectories in $(\mathcal{S} \times \mathcal{A})^\omega$. Nonetheless, an interesting direction for further work could be to more extensively study what happens if the set of trajectories under consideration is restricted in various ways.

Finally, note that Theorem 2 assumes that the discount parameter $\gamma \geq 0.5$. It is not clear if this is strictly necessary, so it might be possible to generalise Theorem 2 by removing this requirement. This would, however, require a different proof strategy. Nonetheless, this assumption is not very restrictive, as in practice γ is almost always set to be greater than 0.5 (typically $\gamma \geq 0.9$).

5 MODAL PROBLEMS

The final class of problems that we will examine is a class of tasks that we refer to as *modal* tasks. Before we give a formal definition of this class, we will first provide some intuition. In analytic philosophy, a distinction is made between *categorical* facts and *modal* facts. In short, categorical facts only concern what is true in actuality, whereas modal facts concern what must be true, could have been true, or cannot be true, etc. For example, it is a categorical fact that the Eiffel Tower is brown, and a modal fact that it *could have had* a different colour. It is (arguably) a categorical fact that the number 3 is prime, and a modal fact that it *could not have been* otherwise. To give another example, there is a difference between stating that *nothing can travel faster than light* and that *nothing does travel faster than light* – the former statement, which is modal, is stronger than the latter, which is categorical. One can further distinguish between different kinds of possibility (e.g. logical vs physical possibility, etc), and discussions about modality also involves topics such as *causality* and *counterfactuals*, etc. A complete treatment of this subject is beyond the scope of this paper, but for an overview see Menzel [2021].

Modality does of course relate to *modal logic*, and thus also to *temporal logic*. In particular, computational tree logic (CTL, see e.g. Baier and Katoen [2008]), and its extensions, can express many modal statements. [1]

The intuition behind this section is that a reward function always is expressed in terms of categorical facts, whereas many tasks are naturally expressed in terms of modal facts. For example, consider an instruction such as "you should always be *able* to return to the start state". This instruction seems quite reasonable, but it is not obvious how to translate it into a reward function. Note that this instruction is not telling the agent to *actually* return to the start state, it merely says that it should maintain the *ability* to do so. This illustrates motivation behind modal tasks; they let us reward the agent based on what is *possible* or *impossible* along its trajectory, rather than just in terms of what in fact occurs along that trajectory. Given this background motivation, we can now give a formal definition of modal tasks:

Definition 7. *Given a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$, a* ***modal reward function*** *$R^\Diamond$ is a function $R^\Diamond : \mathcal{S} \times \mathcal{A} \times (\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})) \to \mathbb{R}$ which takes a state $s \in \mathcal{S}$, an*

[1] To avoid a possible confusion, we should emphasise that we here use the term "modal" in a somewhat more narrow sense than the sense of "modal logic". In particular, we use it to mean "pertaining to what is possible or impossible", as in e.g. Kment [2021]. In that sense, Linear Temporal Logic (LTL) does not express modal statements, even though it is a modal logic, because LTL can only make assertions about what in fact occurs. For that reason, not everything that relates to modal logic will be related to the setting we discuss here. The type of possibility we discuss is specifically "possibility according to the transition function".

action $a \in \mathcal{A}$, and a transition function τ over $\mathcal{S}$ and $\mathcal{A}$, and returns a real number.

$R^\Diamond(s, a, \tau)$ is the reward that is obtained when taking action a in state s in an environment whose transition function is τ. Here we allow $R^\Diamond$ an unrestricted dependence on τ, to make our results as general as possible, even if a practical algorithm for solving modal tasks presumably would require restrictions on what this dependence can look like (see Appendix D). Modal reward functions can be used to express instructions such as that we gave above. For example, a simple case might be "you get 1 reward if you reach this goal state, and -1 reward if you ever enter a state from which you cannot reach the initial state". This reward depends on the transition function, because the transition function determines from which states you can reach the initial state. As usual, $R^\Diamond$ then induces a Q-function $Q^\Diamond$, value function $V^\Diamond$, and evaluation function $J^\Diamond$, etc. We say that a modal reward $R^\Diamond$ and an ordinary reward R are *contingently equivalent* given a transition function τ if $J^\Diamond$ and J induce the same ordering of policies given τ, and that they are *robustly equivalent* if $J^\Diamond$ and J induce the same ordering of policies for all τ. We use $R^\Diamond_\tau$ to denote the reward function $R^\Diamond_\tau(s, a) = R^\Diamond(s, a, \tau)$. We will also use the following definition.

Definition 8. *A modal reward function $R^\Diamond$ is **vacuous** if there is a reward function R such that for all τ, R and $R^\Diamond_\tau$ have the same policy ordering under τ.*

The intuition here is that a vacuous modal reward function does not actually depend on τ in any important sense. Note that this is *not* necessarily to say that $R^\Diamond_\tau = R$ for all τ. For example, it could be the case that $R^\Diamond_\tau$ is a *scaled* version of R, or that $R^\Diamond_\tau$ and R differ by *potential shaping* Ng et al. [1999], or that $R^\Diamond_\tau$ is modified in a way such that $\mathbb{E}_{S' \sim \tau(s,a)}[R^\Diamond_\tau(s, a, S')] = \mathbb{E}_{S' \sim \tau(s,a)}[R(s, a, S')]$, since none of these differences affect the policy ordering (for a more in-depth examination, see Skalse et al. [2022a]). From this, we get the following straightforward result:

Theorem 3. *For any modal reward $R^\Diamond$ and any transition function τ, there exists a reward R that is contingently equivalent to $R^\Diamond$ given τ. Moreover, unless $R^\Diamond$ is vacuous, there is no reward that is robustly equivalent to $R^\Diamond$.*

In other words, *every* modal task can be expressed with an ordinary reward function in each particular given environment, but *no* reward function expresses a (non-vacuous) modal task in all environments. Is this enough? We argue that it is not, because the construction of $R^\Diamond_\tau$ will invariably be laborious, and require detailed knowledge of the environment. For example, consider the task "you should always be able to return to the start state"; here, constructing $R^\Diamond_\tau$ would amount to manually enumerating all the states from which the start state is reachable: this would be very much against the spirit of RL, where much of the point is that we want to be able to specify tasks which can be pursued in *unknown* environments. In short, a method which requires a model of the environment is arguably not an RL method. We thus argue that reward functions are largely unable to capture modal tasks in a satisfactory way.

One remaining question might be why one would want to express tasks for RL agents in terms of modal properties. After all, what benefit is there to the instruction "never enter a state from which it is possible to quickly enter an unsafe state" over the instruction "never enter an unsafe state"? One reason is that the former task might lead to behaviour that is more robust to changes in the environment. For example, if an RL agent is trained in a simulated environment, and deployed in the real world, then it seems like it would be preferable to tell the agent to avoid *risky* states, rather than *unsafe* states, since imperfections in the simulation could lead to an underestimation of the risk involved. Another example is the existing work on avoiding side effects (e.g. [Krakovna et al., 2018, 2020, Turner et al., 2020, Griffin et al., 2022]), which it is natural to express in modal terms. This work can be viewed as being aimed at making the behaviour of an RL agent more robust to misspecification of the reward function.

6 SOLVING TASKS THAT ARE INEXPRESSIBLE BY MARKOVIAN REWARDS

We have pointed to three broad classes tasks that cannot be expressed using scalar, Markovian reward functions, namely multi-objective, risk-sensitive, and modal tasks. A natural next question is whether these tasks *can* be solved at all using RL, or whether only tasks corresponding to Markovian reward functions can be effectively learnt. We discuss this issue below. In short, it is indeed possible to design RL algorithms for tasks in each of these categories.

Multi-objective reinforcement learning is well-explored, with many existing algorithms (see Section 7.1). Most of these algorithms are designed to solve a specific MORL objective; for example, Skalse et al. [2022b] solve the **Lex-Max** objective, and Tessler et al. [2019] solve the **ConSat** objective. There is (as far as we know) no algorithm for the solving e.g. the **MaxMin** objective, but there is no good reason to believe that such an algorithm cannot be devised. Similarly, there are existing algorithms for risk-sensitive RL (e.g. Chow et al. [2017]), and even algorithms that solve certain modal tasks [Krakovna et al., 2018, 2020, Turner et al., 2020, Wang et al., 2020, Griffin et al., 2022].

It should also be possible to design algorithms that can flexibly solve many different tasks from the classes we have discussed, instead of having to be designed for just one particular task. For example, suppose a MORL objective can be

represented by a function $U : \mathbb{R}^k \to \mathbb{R}$, such that $\pi_1 \prec \pi_2$ when $U(J_1(\pi_1) \dots J_k(\pi_1)) < U(J_1(\pi_2) \dots J_k(\pi_2))$, and that U is *differentiable*. We give a few examples of such objectives in Appendix C, including e.g. a "soft" version of **MaxMin**. With such an objective, if we have a policy π that is differentiable with respect to some parameters θ, then one could compute the gradient of $U(J_1(\pi) \dots J_k(\pi))$ with respect to θ, and then use a policy gradient method to increase U. This means that it should be possible to design an actor-critic algorithm that can solve any differentiable MORL objective. We consider the development of such methods to be a promising direction for further work.

In Appendix D, we also outline a possible approach for solving a wide class of modal tasks.

7 DISCUSSION

In this paper, we have studied the ability of Markovian reward functions to express different kinds of problems. We have looked at three classes of tasks; multi-objective tasks, risk-sensitive tasks, and modal tasks, and found that Markovian reward functions are unable to express most of the tasks in each of these three classes. In particular, have provided necessary and sufficient conditions for when a single-policy MORL problem can be expressed using a scalar, Markovian reward function, and demonstrated that this only can be done when the MORL objective corresponds to a linear weighting of the individual rewards. Moreover, we have also provided necessary and sufficient conditions for when a monotonic transformation of the return function, G, can be expressed as a Markovian reward function, and demonstrated that this only can be done for affine transformations. Furthermore, we have also also drawn attention to a class of tasks which have just barely been explored previously (namely modal tasks), and shown that most of these tasks cannot be expressed using Markovian reward functions. Finally, we have shown that many of these problems still can be solved with RL, and even outlined some methods for doing this.

There are several conclusions that we can draw from our results. First of all, the fact that the common settings of MORL and risk-sensitive RL indeed are genuine extensions over the standard (scalar, Markovian) setting provides additional motivation for work in these areas. Moreover, the fact that there are intuitive and commonsensical objectives which cannot be expressed using Markovian reward functions, motivates some caution about the indiscriminate use of such reward functions. It means that there are situations in which careful reward specification and reward shaping will not be sufficient to robustly incentivise the desired behaviour. In those cases, we must instead use an alternative policy synthesis method, such as e.g. those offered by MORL. This also means that a reward learning algorithm which assumes that the data is generated from a Markovian reward function runs the risk of being misspecified, for example.

7.1 RELATED WORK

There has been a lot of recent work on the expressivity of Markovian reward functions. Here, we summarise relevant contributions, and detail differences with our work.

Notably, there are three recent papers which provide necessary and sufficient conditions for when a particular type of task can be expressed using a particular type of reward function. The first of these is Pitis [2019], who consider a task to be a preference relation defined over *prospects*, where a prospect is defined as a pair of a state and a policy. Moreover, they generalise the discount function by allowing it to depend on the transition (instead of always being a constant value γ). They then add two axioms (and one assumption) to the famous vNM-axioms (from von Neumann and Morgenstern [1947]), to obtain necessary and sufficient conditions for when a task (as they formalise it) can be expressed as a Markovian reward with transition-dependent discounting. Our work differs from their in several ways.

The next paper is Shakerinava and Ravanbakhsh [2022], who provide an alternative, simpler axiomatisation of the setting considered by Pitis [2019], and also provide further axioms to describe two additional types of environments. They consider environments without any discount factor, but instead use termination probabilities, which can be used to simulate the standard case with exponential discounting.

The third paper is Bowling et al. [2022], who generalise the results of Shakerinava and Ravanbakhsh [2022] even further, and provide an alternative axiom to add to the vNM axioms. They start by considering preference relations over finite trajectories, and then extend this to a preference relation over policies by saying that a policy π_1 is preferred to π_2 if there exists a time t after which the trajectory distribution induced by π_1 is always preferable to the trajectory distribution induced by π_2. This encompasses the setting with exponentially discounted reward, the setting with limit-average reward, and the episodic setting. They consider both the case where the discount function is transition-dependent, and the case when it is constant.

Our work differs from that by Pitis [2019], Shakerinava and Ravanbakhsh [2022], Bowling et al. [2022] in a few ways. First of all, these papers aim to establish general necessary and sufficient conditions for when a task can be formalised as a Markovian reward, whereas we instead focus on three specific classes of tasks that we believe to be especially interesting. It might in principle be possible to derive our results as a special case of theirs. However, doing this would be quite non-trivial, and possibly more difficult than our direct derivations. Secondly, the axiomatisations provided by Pitis [2019], Shakerinava and Ravanbakhsh [2022], Bowling et al. [2022] are quite opaque, and it is difficult to get a sense of all their practical implications. Our results, on the other hand, are very clear and intuitive to understand, and

concern some settings that are both popular and important. Our work could thus be construed as a study on the practical consequences of the work by Pitis [2019], Shakerinava and Ravanbakhsh [2022], Bowling et al. [2022], with results that may be more directly useful to practitioners. There are also several small differences in how we formalise the problem compared to Pitis [2019], Shakerinava and Ravanbakhsh [2022], Bowling et al. [2022]. For example, we consider the case with fixed discount rates, whereas Pitis [2019] and Bowling et al. [2022] consider transition-dependent discount rates. To give another example, Shakerinava and Ravanbakhsh [2022] consider finite trajectories, whereas we consider infinite trajectories (noting that the latter can model the former, but not vice versa). These differences further contribute to distinguishing our results from theirs.

Another notable piece of related work is Abel et al. [2021], who point to three different ways to formalise the notion of a "task" (namely, as a set of acceptable policies, as an ordering over policies, or as an ordering over trajectories). They then demonstrate that each of these classes contains at least one instance which cannot be expressed using a Markovian reward function, and provide algorithms which compute reward functions for these types of tasks. Our work is different from theirs in a few different ways. First of all, we consider three different ways to specify a policy ordering, and then derive necessary and sufficient conditions which can be used to directly determine when the resulting policy ordering can be expressed as a Markovian reward function. Abel et al. [2021] do not provide necessary and sufficient conditions, but instead only provide a counter-example for each type of task, showing that Markovian rewards cannot formalise all tasks of that type.

Another important paper is the work by Vamplew et al. [2022], who argue that there are many important aspects of intelligence which can be captured by MORL, but not by scalar RL. Like them, we also argue that MORL is a genuine extension of scalar RL, but our approach is quite different. They focus on the question of whether MORL or (scalar) RL is a better foundation for the development of general intelligence (considering feasibility, safety, and etc), and they provide qualitative arguments and biological evidence. By contrast, we are more narrowly focused on what incentive structures can be expressed by MORL and scalar RL, and our results are mathematical.

Miura [2022] considers the question of when a task can be expressed as a constrained MDP (CMDP), or as a Markovian reward. They formalise a task as two sets of policies, $\langle \Pi_G, \Pi_B \rangle$, and consider a CMDP to express the task if all policies in Π_G, and none of the policies in Π_B, are feasible, and consider a Markovian reward to express the task if all policies in Π_G, and none of the policies in Π_B, are optimal under that reward. They then derive necessary and sufficient conditions for both of these cases, and show that CMDPs are strictly more expressive than Markovian rewards for these types of tasks. The CMDP framework is a special case of the MORL framework we discuss in Section 3, roughly corresponding to the **MaxSat** objective. On the other hand, we formalise the notion of a task as a policy ordering, whereas Miura [2022] formalises it as a set of feasible policies.

Also relevant is the work by Pitis et al. [2022], who consider a task to consist of multiple Markovian reward functions, each of which may use a different discount parameter, and where the goal is to maximise the sum of these rewards. They then show that this setting may lead to the optimal policy being non-stationary, which demonstrates that it cannot always be expressed using Markovian rewards. Our analysis of the MORL setting allows for more general objectives than the case where the goal is to maximise the sum of the individual rewards. On the other hand, we assume that the same discount parameter is used for each reward. Our analysis is therefore in some ways more general, and in other ways more restrictive, than that of Pitis et al. [2022].

Also related is the work by Skalse et al. [2022c], who demonstrate that if for two rewards R_1, R_2 there are no policies π_1, π_2 such that $J_1(\pi_1) < J_2(\pi_2)$ and $J_2(\pi_1) > J_2(\pi_2)$, then either R_1 and R_2 are equivalent, or one of them is trivial. This means that there are some policy orderings that cannot be expressed using Markovian rewards. We consider different kinds of policy orderings than they do.

There is also other relevant work that is less strongly related. For example, Icarte et al. [2022] point out that there are certain tasks which cannot be expressed using Markovian rewards, and propose a way extend their expressivity by augmenting the reward function with an automaton that they call a *reward machine*. Similar approaches have also been used by Hasanbeig et al. [2020], Hammond et al. [2021], tackling infinite-horizon tasks for single- and multi-agent systems. There are also other ways to extend Markovian rewards to a more general setting, such as *convex RL*, as studied by e.g. Hazan et al. [2019], Zhang et al. [2020], Zahavy et al. [2021], Geist et al. [2022], Mutti et al. [2022], and *vectorial RL*, as studied by e.g. Cheung [2019a,b]. Analysing the expressivity of these problem settings more extensively would be an interesting direction for further work.

There is a large literature on (the overlapping topics of) single-policy MORL, constrained RL, and risk-sensitive RL. Some notable examples of this work include Achiam et al. [2017], Chow et al. [2017], Miryoosefi et al. [2019], Tessler et al. [2019], Skalse et al. [2022b]. This existing literature typically focuses on the creation of algorithms for solving particular MORL problems, rather than on characterising when MORL problems can (or cannot) be reduced to scalar RL. Modal RL has (to the best of our knowledge) never been discussed explicitly in the literature before. However, it relates to some existing work, such as side-effect avoidance [Krakovna et al., 2018, 2020, Turner et al., 2020, Griffin et al., 2022], and the work by Wang et al. [2020].

References

David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael L. Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward, 2021. URL https://arxiv.org/abs/2111.00876.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 22–31. JMLR.org, 2017.

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

Michael Bowling, John D. Martin, David Abel, and Will Dabney. Settling the reward hypothesis, 2022. URL https://arxiv.org/abs/2212.10420.

Wang Chi Cheung. Exploration-exploitation trade-off in reinforcement learning on online Markov decision processes with global concave rewards, 2019a. URL https://arxiv.org/abs/1905.06466.

Wang Chi Cheung. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 489–497, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Charlie Griffin, Joar Max Viktor Skalse, Lewis Hammond, and Alessandro Abate. All's well that ends well: Avoiding side effects with distance-impact penalties. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=3tgegVVh2j6.

L. Hammond, A. Abate, J. Gutierrez, and M. Wooldridge. Multi-agent reinforcement learning with temporal logic specifications. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 583–592, 2021.

M. Hasanbeig, D. Kroening, and A. Abate. Deep reinforcement learning with temporal logics. In *Proceedings of FORMATS, LNCS 12288*, pages 1–22, 2020.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby van Soest. Provably efficient maximum entropy exploration. In *36th International Conference on Machine Learning, ICML 2019*, 36th International Conference on Machine Learning, ICML 2019, pages 4774–4786. International Machine Learning Society (IMLS), January 2019. 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, jan 2022. doi: 10.1613/jair.1.12440.

Boris Kment. Varieties of Modality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability, 2018. URL https://arxiv.org/abs/1806.01186.

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks, 2020. URL https://arxiv.org/abs/2010.07877.

C. Liu, X. Xu, and D. Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2015.

Christopher Menzel. Possible Worlds. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Sobhan Miryoosefi, Kianté Brantley, Hal Daumé III, Miroslav Dudík, and Robert E. Schapire. Reinforcement learning with convex constraints. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 14070–14079, 2019.

Shuwa Miura. On the expressivity of multidimensional Markov reward. *RLDM Workshop on Reinforcement Learning as a Model of Agency*, 2022.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2022.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and

application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287, Bled, Slovenia, 1999. Morgan Kaufmann Publishers Inc.

Silviu Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 7949–7956, Jul. 2019.

Silviu Pitis, Duncan Bailey, and Jimmy Ba. Rational multi-objective agents must admit non-Markov reward representations. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=MNwA4sgzR4W.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 10 2013. ISSN 1076-9757. doi: 10.1613/jair.3987. URL http://dx.doi.org/10.1613/jair.3987.

Mehran Shakerinava and Siamak Ravanbakhsh. Utility theory for sequential decision making. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19616–19625. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shakerinava22a.html.

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287–308, 2000.

Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning, 2022a. URL https://arxiv.org/abs/2203.07475.

Joar Skalse, Lewis Hammond, Charlie Griffin, and Alessandro Abate. Lexicographic multi-objective reinforcement learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3430–3436. International Joint Conferences on Artificial Intelligence Organization, 7 2022b. doi: 10.24963/ijcai.2022/476. URL https://doi.org/10.24963/ijcai.2022/476. Main Track.

Joar Skalse, Niki Howe, Krasheninnikov Dima, and David Krueger. Defining and characterizing reward hacking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2022c.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. Avoiding side effects in complex environments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21406–21415. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

Peter Vamplew, Benjamin J. Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M. Roijers, Conor F. Hayes, Fredrik Heintz, Patrick Mannion, Pieter J. K. Libin, Richard Dazeley, and Cameron Foale. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, Jul 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09575-5. URL https://doi.org/10.1007/s10458-022-09575-5.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

Yu Wang, Nima Roohi, Matthew West, Mahesh Viswanathan, and Geir E. Dullerud. Statistically model checking pctl specifications on Markov decision processes via reinforcement learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1392–1397, 2020. doi: 10.1109/CDC42340.2020.9303982.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex MDPs, 2021. URL https://arxiv.org/abs/2106.00661.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 2020-December, 2020. ISSN 1049-5258.