The Reward Hypothesis is False

Joar Skalse
Department of Computer Science
Oxford University
joar.skalse@cs.ox.ac.uk

Alessandro Abate
Department of Computer Science
Oxford University
aabate@cs.ox.ac.uk

Abstract

The *reward hypothesis* is the hypothesis that “all of what we mean by goals and purposes can be well thought of as the maximisation of the expected value of the cumulative sum of a received scalar signal”(Sutton and Barto 2018). In this paper, we will argue that this hypothesis is false. We will look at three natural classes of reinforcement learning tasks (multi-objective reinforcement learning, risk-averse reinforcement learning, and modal reinforcement learning), and then prove mathematically that these tasks cannot be expressed using any scalar, Markovian reward function. We thus disprove the reward hypothesis by providing many examples of tasks which are both natural and intuitive to describe, but which are nonetheless impossible to express using reward functions. In the process, we provide necessary and sufficient conditions for when a multi-objective reinforcement learning problem can be reduced to ordinary, scalar reward reinforcement learning. We also call attention to a new class of reinforcement learning problems (namely those we call “modal” problems), which have so far not been given any systematic treatment in the reinforcement learning literature.

1 Introduction

To use reinforcement learning (RL) to solve a task, it is necessary to first encode that task using a reward function, i.e. as a function from *state-action-next-state* triples to reals (Sutton and Barto 2018). The *reward hypothesis* states that *any* task which a human might reasonably wish an RL agent to complete can be expressed using such a reward function. In this paper, we analyse the reward hypothesis mathematically. Moreover, we argue that our results *disprove* the reward hypothesis. We will examine three 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 are impossible to express using ordinary reward functions.

The first class of problems we look at, in Section 2, is the single-policy version of multi-objective RL (MORL). In such a problem, the agent receives multiple reward signals, and the aim is to learn a single policy that achieves an optimal trade-off of those rewards according to some criterion (Roijers et al. 2013; Liu, Xu, and Hu 2015). We will look at the question of which MORL problems can be reduced to ordinary RL, and find that this can *only* be done for MORL problems that correspond to a linear weighting of the rewards. This means that it cannot be done for the vast majority of all MORL problems. The next class of problems we look at, in Section 3, is risks-sensitive RL. We show that none of the standard risk-averse utility functions can be expressed using reward functions. The last class of problems we look at, in Section 4, is something we call *modal* tasks. These are tasks where the agent is evaluated not only based on what trajectories it generates, but also based on what it *could have done* along those trajectories. We provide a formalisation of such tasks, and prove that they cannot be formalised using ordinary reward functions. In Section 5, we discuss our results.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1.1 Related Work

There has been a few recent papers which examine the expressivity of Markovian reward functions. The first of these is the work by Abel et al. 2021, who demonstrate that there are tasks which cannot be expressed using Markovian reward functions. We greatly extend their work by providing new results that are significantly stronger. 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, 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.

There is a large literature on single-policy MORL, constrained RL, and risk-sensitive RL. Some notable examples of this work includes Achiam et al. 2017; Chow et al. 2017; Miryoosefi et al. 2019; Tessler, Mankowitz, and Mannor 2019; Skalse et al. 2022c. This existing literature typically focuses on the creation of algorithms for solving particular MORL problems, and has so far not tackled the problem of characterising when MORL problems can 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; Krakovna et al. 2020a; Turner, Ratzlaff, and Tadepalli 2020), and the work by Wang et al. 2020.

1.2 Preliminaries

We will assume that the reader is familiar with the basics of RL, which can be found in Sutton and Barto 2018. An overview of our notation can be found in Appendix B.

MORL problems are formalised using *Multi-Objective MDPs* (MOMDPs), which are tuples $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$. The only place where MOMDPs differ from MDPs are $\vec{R}$, which is a function $\vec{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightsquigarrow \mathbb{R}^k$ that, for each transition s, a, s', returns k different rewards (for some k). We denote the reward function that returns the i'th component of $\vec{R}$ as R_i, and use V_i^π, Q_i^π, J_i, G_i, etc, to refer to its value functions, Q-functions, evaluation function, return function, etc.

2 Multi-Objective Reinforcement Learning

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 $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$ by itself has no one canonical objective to maximise. We therefore introduce the notion of a *MORL objective*:

Definition 2.1. A **MORL objective** over k rewards is a function $\mathcal{O}$ that takes k policy evaluation functions $J_1 \dots J_k$ and returns a (total) ordering $\prec_{\mathcal{O}}$ over the set of all policies Π.

Given a MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$, 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 policy ordering that is obtained when we apply $\mathcal{O}$ to $\mathcal{M}$'s policy evaluation functions. We next give a few examples of some interesting MORL objectives:

Definition 2.2. Given $J_1 \dots J_k$, the **LexMax** objective $\prec_{\texttt{Lex}}$ is given by $\pi_1 \prec_{\texttt{Lex}} \pi_2$ if and only if there is an $i \in \{1 \dots m\}$ such that $J_i(\pi_1) < J_i(\pi_2)$, and $J_j(\pi_1) = J_j(\pi_2)$ for $j < i$.

Definition 2.3. 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 2.4. 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 2.5. 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 if $J_1(\pi_1), J_1(\pi_2) \geq c$ and $J_2(\pi_1) < J_2(\pi_2)$.

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

Definition 2.6. A MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$ with objective $\mathcal{O}$ is **equivalent** to the MDP $\tilde{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$ if and only if $\tilde{M}$'s policy order is $\prec_{\mathcal{O}}^{\mathcal{M}}$.

Note that $\tilde{\mathcal{M}}$ 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 $\tilde{\mathcal{M}}$ if $\tilde{\mathcal{M}}$ is given by replacing $\vec{R} = \langle R_1 \dots R_k \rangle$ with a single reward function $\tilde{R}$, and $\tilde{R}$ induces the same preferences between all policies as $\mathcal{O}(J_1 \dots J_k)$. We can now derive necessary and sufficient conditions for when a MORL problem can be reduced to a scalar-reward RL problem.

Theorem 2.7. *If a MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$ with objective $\mathcal{O}$ is equivalent to an MDP $\tilde{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$, then $\tilde{J}(\pi) = \sum_{i=1}^{k} w_i \cdot J_i(\pi)$ for some $w_1 \dots w_k \in \mathbb{R}$. Moreover, $\mathcal{M}$ with $\mathcal{O}$ is also equivalent to the MDP with reward $R(s,a,s') = \sum_{i=1}^{k} w_i \cdot R_i(s,a,s')$.*

This theorem effectively tells us that *only linear* MORL objectives can be represented using scalar-reward RL! This imposes a harsh limitation on what kinds of tasks can be encoded using scalar rewards. Theorem 2.7 also has the following corollary, which is useful for demonstrating when some MORL objective cannot be expressed using scalar reward functions. Given an ordering $\prec$ over Π dependent 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 f that is strictly monotonic.

Corollary 2.8. *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 not equivalent to any MDP.*

Therefore, we can prove that $\mathcal{M}$ with $\mathcal{O}$ is not equivalent to any MDP by finding a non-linear representation of $\prec_{\mathcal{O}}^{\mathcal{M}}$. In Appendix C, we prove that none of the MORL objectives given in Definition 2.2-2.5 can be expressed using single-objective RL, except in a few degenerate edge cases.

3 Risk-Sensitive Reinforcement Learning

The next area we will look at is that of *risk-sensitive* reinforcement learning. An ordinary RL agent tries to maximise the *expectation* of its reward function. However, there are many cases where it is natural to want the agent to be *risk-averse*. In economics, risk-aversion is typically modelled by using utility functions $U(c)$ that are concave in some relevant quantity c (which might be money, for example). A natural question is then whether a similar trick may be used with reward functions? We will examine this question.

Some of the most common risk-averse utility functions includes *exponential utility*, *isoelastic utility*, and *quadratic utility*. 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 $(-\infty, 1/2\alpha]$. We will examine each of these, and show that none of them can be expressed using reward functions.

In this section, we will consider the domain of G to be the set of all coherent trajectories, *not* the set of trajectories which are possible under some transition function τ. In other words, we consider the set of all trajectories to be $(\mathcal{S} \times \mathcal{A})^\omega$. The reason for this is that we do not want to presume any prior knowledge of the environment. If we restrict the set of trajectories we consider, then some risk-averse utility functions can become possible to express (consider the case of a tree-shaped MDP, for example). Finally, we will say that R is *constant* if it has a constant value for all s, a, s'.

Theorem 3.1. *For any non-constant reward function R_1 and any constant $\alpha \neq 0$, there is no reward function R_2 such that $G_2(\xi) = -e^{\alpha G_1(\xi)}$ for all valid trajectories ξ.*

Theorem 3.2. *For any non-constant reward function R_1 and any constant $\alpha > 0$, $\alpha \neq 1$, there is no reward function R_2 such that $G_2(\xi) = G_1(\xi)^{1-\alpha}$ for all valid trajectories ξ.*

Theorem 3.3. *For any non-constant reward function R_1, there is no reward function R_2 such that $G_2(\xi) = \ln(G_1(\xi))$ for all valid trajectories ξ.*

Theorem 3.4. *For any non-constant reward function R_1 and any $\alpha > 0$ where $\max_\xi G_1(\xi) \leq \frac{1}{2\alpha}$, there is no reward function R_2 such that $G_2(\xi) = G_1(\xi) - \alpha G_1(\xi)^2$ for all ξ.*

It would be desirable to strengthen these results, and provide necessary and sufficient conditions for when it is possible to construct a reward R_2 such that $G_2(\xi) = f(G_1(\xi))$ for some function f and some (non-constant) reward R_1. We consider this to be an important question for further work.

4 Modal Reinforcement Learning

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. To give a few other examples, consider instructions such as "you should never enter a state from which it is *possible* to quickly enter an unsafe state", "you should always be *able* to press the emergency shutdown button", or "you should never enter a state where you would be *unable* to receive a feedback signal". These instructions all seem very reasonable, and they are expressed in terms of what should be *possible* or *impossible* along the trajectory of the agent, rather than 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 4.1. Given a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$, a **modal reward function** $R^\diamondsuit$ is a function $R^\diamondsuit : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \times (\mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}) \to \mathbb{R}$ which takes two states $s, s' \in \mathcal{S}$, an action $a \in \mathcal{A}$, and a transition function τ over $\mathcal{S}$ and $\mathcal{A}$, and returns a real number.

$R^\diamondsuit(s, a, s', \tau)$ is the reward that is obtained when transitioning from state s to s' using action a in an environment whose transition function is τ. Here we allow $R^\diamondsuit$ 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. As usual, $R^\diamondsuit$ then induces a Q-function $Q^\diamondsuit$, value function $V^\diamondsuit$, and evaluation function $J^\diamondsuit$, etc. Modal reward functions can be used to express all the instructions we gave above. We say that a modal reward $R^\diamondsuit$ and an ordinary reward R are *contingently equivalent* given a transition function τ if $J^\diamondsuit$ and J induce the same ordering of policies given τ, and that they are *robustly equivalent* if $J^\diamondsuit$ and J induce the same ordering of policies for all τ. We use $R^\diamondsuit_\tau$ to denote the reward function $R^\diamondsuit_\tau(s, a, s') = R^\diamondsuit(s, a, s', \tau)$. We will also use the following definition:

Definition 4.2. A modal reward function $R^\diamondsuit$ is **trivial** if there is a reward function R such that for all τ, R and $R^\diamondsuit_\tau$ have the same policy ordering under τ.

The intuition here is that a trivial modal reward function does not actually depend on τ in any important sense. Note that this is *not* necessarily to say that $R^\diamondsuit_\tau = R$ for all τ. For example, it could be the case that $R^\diamondsuit_\tau$ is a *scaled* version of R, or that $R^\diamondsuit_\tau$ and R differ by *potential shaping* Ng, Harada, and Russell 1999, or that $R^\diamondsuit_\tau$ is modified in a way such that $\mathbb{E}_{S'\sim\tau(s,a)}[R^\diamondsuit_\tau(s, a, S')] = \mathbb{E}_{S'\sim\tau(s,a)}[R(s, a, S')]$, since none of these differences affect the policy ordering.

Theorem 4.3. *For any modal reward $R^\diamondsuit$ and any transition function τ, there exists a reward function R that is contingently equivalent to $R^\diamondsuit$ given τ. Moreover, unless $R^\diamondsuit$ is trivial, there is no reward function that is robustly equivalent to $R^\diamondsuit$.*

In other words, every modal task can be expressed with ordinary reward function in each particular environment, but no reward function expresses a (non-trivial) modal task in all environments.

5 Discussion

In this paper, we have studied the veracity of the reward hypothesis, by examining 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 reward functions are unable to express most of the tasks in each of these three classes. We have also provided necessary and sufficient conditions for when a multi-objective RL problem can be expressed using a single reward function, and also drawn attention to a class of tasks which have just barely been explored previously (namely modal tasks). Finally, we have also shown that many of these problems still can be solved with RL, and even outlined some methods for how to extend these solutions, which rules out the possibility that only tasks which can be expressed using reward functions can be effectively learnt. We argue that our results show that the reward hypothesis is false – there are tasks, which are natural to state and intuitive to understand, and which can be solved with RL methods, but which cannot be expressed using scalar Markovian reward functions.

References

[1] David Abel et al. *On the Expressivity of Markov Reward*. 2021. DOI: 10.48550/ARXIV.2111.00876. URL: https://arxiv.org/abs/2111.00876.

[2] Joshua Achiam et al. “Constrained Policy Optimization”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, 22–31.

[3] Stephen Adams, Tyler Cody, and Peter A. Beling. “A survey of inverse reinforcement learning”. In: *Artificial Intelligence Review* 55.6 (Aug. 2022), pp. 4307–4346. ISSN: 1573-7462. DOI: 10.1007/s10462-021-10108-x. URL: https://doi.org/10.1007/s10462-021-10108-x.

[4] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 1st. USA: Oxford University Press, Inc., 2014. ISBN: 0199678111.

[5] Yinlam Chow et al. “Risk-constrained Reinforcement Learning with Percentile Risk Criteria”. In: *Journal of Machine Learning Research* 18.1 (2017), pp. 6070–6120.

[6] Paul Christiano et al. *Deep reinforcement learning from human preferences*. 2017. DOI: 10.48550/ARXIV.1706.03741. URL: https://arxiv.org/abs/1706.03741.

[7] Victoria Krakovna et al. *Avoiding Side Effects By Considering Future Tasks*. 2020. DOI: 10.48550/ARXIV.2010.07877. URL: https://arxiv.org/abs/2010.07877.

[8] Victoria Krakovna et al. *Penalizing side effects using stepwise relative reachability*. 2018. DOI: 10.48550/ARXIV.1806.01186. URL: https://arxiv.org/abs/1806.01186.

[9] Victoria Krakovna et al. *Specification gaming: the flip side of AI ingenuity*. Apr. 2020. URL: https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4.

[10] C. Liu, X. Xu, and D. Hu. “Multiobjective Reinforcement Learning: A Comprehensive Overview”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.3 (2015), pp. 385–398.

[11] Sobhan Miryoosefi et al. “Reinforcement Learning with Convex Constraints”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, pp. 14070–14079.

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

[13] 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*. Bled, Slovenia: Morgan Kaufmann Publishers Inc, 1999, pp. 278–287.

[14] Stephen M. Omohundro. “The Basic AI Drives”. In: *AGI*. 2008.

[15] D. M. Roijers et al. “A Survey of Multi-Objective Sequential Decision-Making”. In: *Journal of Artificial Intelligence Research* 48 (Oct. 2013), 67–113. ISSN: 1076-9757. DOI: 10.1613/jair.3987. URL: http://dx.doi.org/10.1613/jair.3987.

[16] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: a modern approach*. 3rd ed. Pearson, 2009.

[17] Stuart J. (Stuart Jonathan) Russell. *Human compatible : artificial intelligence and the problem of control*. eng. New York, New York: Viking, 2019. ISBN: 9780525558613.

[18] Joar Skalse et al. “Defining and Characterizing Reward Gaming”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2022.

[19] Joar Skalse et al. *Invariance in Policy Optimisation and Partial Identifiability in Reward Learning*. 2022. DOI: 10.48550/ARXIV.2203.07475. URL: https://arxiv.org/abs/2203.07475.

[20] Joar Skalse et al. “Lexicographic Multi-Objective Reinforcement Learning”. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Ed. by Lud De Raedt. Main Track. International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 3430–3436. DOI: 10.24963/ijcai.2022/476. URL: https://doi.org/10.24963/ijcai.2022/476.

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

[22] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. “Reward Constrained Policy Optimization”. In: *Proceedings of the 7th International Conference on Learning Representations*. 2019.

[23] Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. “Avoiding Side Effects in Complex Environments”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 21406–21415. URL: https://proceedings.neurips.cc/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

[24] Peter Vamplew et al. “Scalar reward is not enough: a response to Silver, Singh, Precup and Sutton (2021)”. In: *Autonomous Agents and Multi-Agent Systems* 36.2 (June 2022), p. 41. ISSN: 1573-7454. DOI: 10.1007/s10458-022-09575-5. URL: https://doi.org/10.1007/s10458-022-09575-5.

[25] Yu Wang et al. “Statistically Model Checking PCTL Specifications on Markov Decision Processes via Reinforcement Learning”. In: *2020 59th IEEE Conference on Decision and Control (CDC)*. 2020, pp. 1392–1397. DOI: 10.1109/CDC42340.2020.9303982.

[26] Eliezer Yudkowsky. *AGI Ruin: A List of Lethalities*. Accessed: 2022-09-25. June 2022. URL: https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities#Section_A_.

A AI Risk Analysis

One of the central challenges in AGI safety is the question of how to specify instructions for AI systems which are safe to pursue with arbitrary competence. It has been hypothesised that if a sufficiently advanced AI system is directed towards some goal, and this goal does not incorporate all the implicit intentions and preferences of the human user, then this AI system will by default be incentivised to take harmful actions or pursue dangerous plans. For a more complete version of this argument, see e.g. Omohundro 2008; Bostrom 2014; Russell 2019; Yudkowsky 2022. Moreover, the dynamics that these arguments describe can be observed in real systems. It is widely acknowledged that reward functions can be very difficult to specify, and that a misspecified reward function can induce behaviour that is both surprising and undesirable, if not dangerous, even in current AI systems (Krakovna et al. 2020b). It seems likely that this problem will continue to get worse as AI systems become more advanced, unless the fundamental underlying problem is solved. This then implies that we must develop better ways to specify safe instructions for AI systems.

There seems to be a widespread belief that Markovian reward functions must be expressive enough to formalise any task in a satisfactory way. For example, the objective in a sequential decision-making task is almost always formalised as a Markovian reward function (Russell and Norvig 2009; Sutton and Barto 2018). Moreover, current work on specifying tasks which are difficult to formalise typically creates a setup in which a Markovian reward function is learnt from data, in the form of e.g. feedback (e.g. Christiano et al. 2017) or demonstrations (e.g. Adams, Cody, and Beling 2022). It is crucial, therefore, to examine whether or not this assumption is true. As we have seen, there are many natural tasks which cannot be expressed as Markovian reward functions.

The assumption that any natural task can be represented using Markovian reward functions might come from a false analogy to utility functions. In their famous work, Neumann and Morgenstern 1947 show that any preference ordering over lotteries of a finite set of outcomes can be rationalised by a utility function, given some very moderate assumptions. However, reward functions are not utility functions. First, the VNM utility theorem assumes a finite choice set, but the set of possible trajectories is in general uncountable. Second, not all distributions over trajectories can be represented as policies. Third, reward functions have a special linear structure, and cannot express arbitrary functions from trajectories to reals. The VNM utility theorem does therefore not carry over to the reinforcement learning setting.

A better understanding of what can and cannot be expressed as Markovian reward functions will be helpful for developing methods for aligning AI systems. First of all, most reward learning methods will attempt to fit a Markovian reward function to some training data. However, as we have seen, it might be that no Markovian reward function can represent the task that is exemplified in the data. Similarly, if we develop RL methods that can learn tasks which are formalised in more expressive languages, then that might also make it easier to formulate safe tasks for these systems. All in all, we hope that an improved understanding of the expressivity of Markovian reward functions will make it less likely that misspecified rewards will be used in powerful (or otherwise safety critical) AI systems.

We do not anticipate any noteworthy risk of unintended negative externalities from this research.

B Notation

The standard RL setting is formalised using *Markov Decision Processes* (MDPs), 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, $\tau : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ is a transition function, μ_0 is an initial state distribution over $\mathcal{S}$, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightsquigarrow \mathbb{R}$ a reward function, where $R(s, a, s')$ is the reward obtained if the agent moves from state s to s' by taking action a, and $\gamma \in (0, 1)$ is a discount factor. Here, $f : X \rightsquigarrow Y$ denotes a probabilistic mapping f from X to Y. A state is *terminal* if $\tau(s, a) = s$ and $R(s, a, s) = 0$ for all a. A *trajectory* ξ is a path $s_0, a_0, s_1 \dots$ in an MDP that is possible according to μ_0 and τ. We use G to denote the *trajectory return function*, where $G(\xi) = \sum_{t=0}^{\infty} \gamma^t r_t$. A *policy* is a mapping $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$, 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* $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R} = \mathbb{E}_{S' \sim \tau(s,a)}[R(s, a, S') + \gamma \cdot V^\pi(S')]$. 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 induced by $\pi_1 \prec \pi_2 \iff J(\pi_1) < J(\pi_2)$ for all 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

C Inexpressible MORL Objectives

Theorem C.1. *There is no MDP equivalent to $\mathcal{M}$ with $\texttt{LexMax}$, as long as $\mathcal{M}$ has at least two reward functions that are neither trivial, equivalent, or opposites.*

Proof. Suppose $\mathcal{M}$ with LexMax is equivalent to $\tilde{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$. Let i be the smallest number such that R_i is non-trivial, and let j be the smallest number greater than i such that R_j is non-trivial, and not equivalent to or opposite of R_i. Then there are π_1, π_2 such that $J_i(\pi_1) = J_i(\pi_2)$ and $J_j(\pi_1) < J_j(\pi_2)$, which means that $\pi_1 \prec^{\mathcal{M}}_{\texttt{Lex}} \pi_2$. Moreover, since $\tilde{J}$ represents $\prec^{\mathcal{M}}_{\texttt{Lex}}$, it follows that there are no π, π' such that $J_i(\pi) < J_i(\pi')$ and $\tilde{J}(\pi) > \tilde{J}(\pi')$. Then Theorem 1 in Skalse et al. 2022a implies that R_i is equivalent to $\tilde{R}$. However, then $\tilde{J}(\pi_1) = \tilde{J}(\pi_2)$, which means that $\tilde{J}$ cannot represent $\prec^{\mathcal{M}}_{\texttt{Lex}}$. □

Theorem C.2. *There is no MDP equivalent to $\mathcal{M}$ with $\texttt{MaxMin}$, 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.*

Proof. $\mathcal{O}^{\mathcal{M}}_{\texttt{Min}}$ is represented by the function $U(\pi) = \min_i J_i(\pi)$. Moreover, if $\mathcal{M}$ has no reward function R_i such that $J_i(\pi) \leq J_j(\pi)$ for all $j \in \{1 \dots k\}$ and all π then this representation is non-linear. Corollary 2.8 then implies that $\mathcal{M}$ with MaxMin is not equivalent to any MDP. □

Theorem C.3. *There is no MDP equivalent to $\mathcal{M}$ with $\texttt{MaxSat}$, 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 $\pi_1, \pi_2 \in \Pi$.*

Proof. Note that $\texttt{MaxSat}(\mathcal{M})$ is represented by the function $U(\pi) = \sum_{i=1}^{k} \mathbb{1}[J_i(\pi) \geq c_i]$, where $\mathbb{1}[J_i(\pi) \geq c_i]$ is the function that is equal to 1 when $J_i(\pi) \geq c_i$, and 0 otherwise. Moreover, U is not strictly monotonic in any function that is linear in $J_1 \dots J_k$. Corollary 2.8 thus implies that $\mathcal{M}$ with MaxSat is not equivalent to any MDP. □

Theorem C.4. *There is no MDP equivalent to $\mathcal{M}$ with $\texttt{ConSat}$, unless either R_1 and R_2 are equivalent, or $\max_\pi J_1(\pi) \leq c$.*

Proof. $\mathcal{O}^{\mathcal{M}}_{\texttt{Con}}$ is represented by $U(\pi) = \{J_1(\pi)$ if $J_1(\pi) \leq c$, else $J_2(\pi) - \min_\pi J_2(\pi) + c\}$. Moreover, this representation is non-linear, unless either R_1 and R_2 are equivalent, or $\max_\pi J_1(\pi) \leq c$. Corollary 2.8 then implies that $\mathcal{M}$ with `ConSat` is not equivalent to any MDP. □

D Proofs

In this appendix, we provide the proofs of all of our theorems.

D.1 Multi-Objective Reinforcement Learning

Theorem D.1. *If a MOMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \vec{R}, \gamma \rangle$ with objective $\mathcal{O}$ is equivalent to an MDP $\tilde{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$, then $\tilde{J}(\pi) = \sum_{i=1}^k w_i \cdot J_i(\pi)$ for some $w_1 \dots w_k \in \mathbb{R}$. Moreover, $\mathcal{M}$ with $\mathcal{O}$ is also equivalent to the MDP with reward $R(s, a, s') = \sum_{i=1}^k w_i \cdot R_i(s, a, s')$.*

Proof. Suppose $\mathcal{M}$ with $\mathcal{O}$ is equivalent to an MDP $\tilde{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$. First, let $m : \Pi \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the function that maps each policy π to the $|\mathcal{S}||\mathcal{A}|$-dimensional vector where

$$m(\pi)[s, a] = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi \sim \pi}(S_t = s, A_t = a).$$

Moreover, for a reward function R, let $\vec{R} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the $|\mathcal{S}||\mathcal{A}|$-dimensional vector where

$$\vec{R}[s, a] = \mathbb{E}_{S' \sim \tau(s,a)}[R(s, a, S')].$$

Note that we now have that $J(\pi) = m(\pi) \cdot \vec{R}$, for any reward function R. Recall also that multiplication by an $|\mathcal{S}||\mathcal{A}|$-dimensional vector induces a linear function over $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. This means that, for any reward function R, we can express its policy evaluation function $J : \Pi \to \mathbb{R}$ as $L \circ m$, where L is a linear function. In particular, $\tilde{J} = \tilde{L} \circ m$, and $J_i = L_i \circ m$ for each of $R_i \in \vec{R}$.

From the definition of MORL objectives, we have that $\tilde{J}(\pi)$ is a function of $J_1(\pi) \dots J_k(\pi)$. This, in turn, means that $\tilde{L}(v)$ is a function of $L_1(v) \dots L_k(v)$, for any $v \in \mathrm{Im}(m)$. Let M be the $(|S||A| \times k)$-dimensional matrix that maps each vector $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to $\langle L_1(v), \dots, L_k(v) \rangle$ (in other words, the matrix whose rows are $\vec{R}_1 \dots \vec{R}_k$). Since $\tilde{L}(v)$ is a function of $L_1(v) \dots L_k(v)$, we have that $\tilde{L}$ can be expressed as $f \circ M$ for some function f. Since $\tilde{L}$ is a linear function, and since M is a linear transformation, we that f must be a linear function as well. This means that there are $w_1 \dots w_k \in \mathbb{R}^k$ such that $f(x) = \sum_{i=1}^k w_i \cdot x_i$, which implies that $\tilde{L}(v) = \sum_{i=1}^m w_i \cdot L_i(v)$, and further that $\tilde{J}(\pi) = \sum_{i=1}^k w_i \cdot J_i(\pi)$. This completes the first part.

Next, let $R(s, a, s') = \sum i_1{}^k w_i \cdot R_i(s, a, s')$. Straightforward algebra shows that $J(\pi) = \sum_{i=1}^k w_i \cdot J_i(\pi)$. Now, since $J = \tilde{J}$, and since $\mathcal{M}$ with $\mathcal{O}$ is equivalent to $\tilde{\mathcal{M}}$, we have that $\mathcal{M}$ with $\mathcal{O}$ is equivalent to the MDP with reward R. This completes the second part. □

Corollary D.2. *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 not equivalent to any MDP.*

Proof. Assume for contradiction that $\mathcal{M}$ with $\mathcal{O}$ is equivalent the MDP $\tilde{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \tilde{R}, \gamma \rangle$. Then $\tilde{J}$ represents $\mathcal{O}(J_1 \dots J_k)$, and this in turn means that U must be strictly monotonic in $\tilde{J}$. Moreover, Theorem 2.7 implies that $\tilde{J} = \sum_{i=0}^k w_i \cdot J_i$ for some $w_1 \dots w_k \in \mathbb{R}^k$. However, this contradicts our assumptions. □

D.2 Risk-Sensitive Reinforcement Learning

Lemma D.3. *If R is non-constant, then for any state s there exists trajectories $\zeta_1, \zeta_2, \zeta_3$ starting in s such that $G(\zeta_1) \neq G(\zeta_2)$, $G(\zeta_2) \neq G(\zeta_3)$, and $G(\zeta_1) \neq G(\zeta_3)$.*

Proof. First note that if R is non-constant, then there must be *some* state s and some trajectories ξ_1, ξ_2 starting in s such that $G(\xi_1) \neq G(\xi_2)$ (this follows from Theorem 3.8 in Skalse et al. 2022b). We will establish that there is a ξ_3 starting in s such that $G(\xi_3) \neq G(\xi_1)$ and $G(\xi_3) \neq G(\xi_2)$, and then show that this implies that such trajectories exist for *all* states.

Suppose for contradiction that for any ξ_3 starting in s, either $G(\xi_3) = G(\xi_1)$ or $G(\xi_3) = G(\xi_2)$. Consider a transition $\langle s, a, s \rangle$, and let $\zeta_1 = \langle s, a, s \rangle + \xi_1$ and $\zeta_2 = \langle s, a, s \rangle + \xi_2$; we will do a case enumeration, and show that either $G(\zeta_1)$ or $G(\zeta_2)$ must be distinct from both $G(\xi_1)$ and $G(\xi_2)$. Note that $G(\zeta_1) = R(s, a, s) + \gamma G(\xi_1)$ and $G(\zeta_2) = R(s, a, s) + \gamma G(\xi_2)$.

Case 1: $G(\zeta_1) = G(\xi_1)$, $G(\zeta_2) = G(\xi_2)$. If $R(s, a, s) + \gamma G(\xi_1) = G(\xi_1)$ then $R(s, a, s) = (1 - \gamma) G(\xi_1)$, and similarly, if $R(s, a, s) + \gamma G(\xi_2) = G(\xi_2)$ then $R(s, a, s) = (1 - \gamma) G(\xi_2)$. This is a contradiction, since $G(\xi_1) \neq G(\xi_2)$ and $\gamma \neq 1$.

Case 2: $G(\zeta_1) = G(\zeta_2) = G(\xi_1)$. If $R(s, a, s) + \gamma G(\xi_1) = G(\xi_1)$ then $R(s, a, s) = (1 - \gamma) G(\xi_1)$. Using $R(s, a, s) + \gamma G(\xi_2) = G(\xi_1)$, we get $(1 - \gamma) G(\xi_1) + \gamma G(\xi_2) = \gamma G(\xi_1)$. By rearranging, we get $\gamma(G(\xi_1) - G(\xi_2)) = 0$. This is a contradiction, since $G(\xi_1) \neq G(\xi_2)$ and $\gamma \neq 0$.

Case 3: $G(\zeta_1) = G(\zeta_2) = G(\xi_2)$. This is analogous to Case 2.

Case 4: $G(\zeta_1) = G(\xi_2)$, $G(\zeta_2) = G(\xi_1)$. If $R(s, a, s) + \gamma G(\xi_1) = G(\xi_2)$ then $R(s, a, s) = G(\xi_2) - \gamma G(\xi_2)$, and similarly, if $R(s, a, s) + \gamma G(\xi_2) = G(\xi_1)$ then $R(s, a, s) = G(\xi_1) - \gamma G(\xi_2)$. Combining this, and rearranging, gives $(1 + \gamma) G(\xi_1) = (1 + \gamma) G(\xi_2)$. This is a contradiction, since $G(\xi_1) \neq G(\xi_2)$ and $\gamma \neq -1$.

This exhausts all cases, which means that if R is non-constant, then there must be some state s and some trajectories $\zeta_1, \zeta_2, \zeta_3$ starting in s such that $G(\zeta_1) \neq G(\zeta_2)$, $G(\zeta_2) \neq G(\zeta_3)$, and $G(\zeta_1) \neq G(\zeta_3)$. Finally, note that this means that we can construct such trajectories for *any* state s', by simply composing a transition $\langle s', a, s \rangle$ with each of $\zeta_1, \zeta_2, \zeta_3$. □

Lemma D.4. *If $G_2(\xi) = f(G_1(\xi))$ for all ξ and some f, then for any transition $\langle s, a, s' \rangle$ and any trajectory ζ starting in s', $R_2(s, a, s') = f(R_1(s, a, s') + \gamma G_1(\zeta)) - \gamma f(G_1(\zeta))$.*

Proof. Suppose that $G_2(\xi) = f(G_1(\xi))$ for all trajectories ξ. Let $\langle s, a, s' \rangle$ be an arbitrary transition, let ζ be an arbitrary trajectory starting in s', and let $\xi = \langle s, a, s' \rangle + \zeta$. We have that $G_2(\xi) = R_2(s, a, s') + \gamma G_2(\zeta)$, and also that $G_2(\xi) = f(G_1(\xi))$, which implies that

$$R_2(s, a, s') + \gamma G_2(\zeta) = f(G_1(\xi)).$$

Since $G_1(\xi) = R_1(s, a, s') + \gamma G_1(\zeta)$, this implies that

$$R_2(s, a, s') + \gamma G_2(\zeta) = f(R_1(s, a, s') + \gamma G_1(\zeta)).$$

By using the fact that $G_2(\zeta) = f(G_1(\zeta))$, and rearranging, we get that

$$R_2(s, a, s') = f(R_1(s, a, s') + \gamma G_1(\zeta)) - \gamma f(G_1(\zeta)).$$

Since $\langle s, a, s' \rangle$ and ζ were chosen arbitrarily, this completes the proof. □

Lemma D.5. *For any non-constant reward R_1 and any f that is injective on $\mathrm{range}(G_1)$, if for any $y \in \mathrm{range}(R_1)$ and any $\gamma \in (0, 1)$ there are at most two distinct x_1, x_2 such that $f(y + \gamma x_1) - \gamma f(x_1) = f(y + \gamma x_2) - \gamma f(x_2)$ then there is no reward R_2 such that $G_2(\xi) = f(G_1(\xi))$ for all ξ.*

Proof. Suppose for contradiction that $G_2(\xi) = f(G_1(\xi))$ for all ξ. Let $\langle s, a, s' \rangle$ be an arbitrary transition. Applying Lemma D.4, we get that

$$R_2(s, a, s') = f(R_1(s, a, s') + \gamma G_1(\zeta)) - \gamma f(G_1(\zeta))$$

for all trajectories ζ starting in s'. For clarity, let $x = G_1(\zeta)$ and $y = R_1(s, a, s')$, so that $f(y + \gamma x) - \gamma f(x)$. By assumption, there can be at most two distinct values x_1, x_2 such that $f(y + \gamma x_1) - \gamma f(x_1) = f(y + \gamma x_2) - \gamma f(x_2)$. However, Lemma D.3 implies that there are at least three $\zeta_1, \zeta_2, \zeta_3$ starting in s' with distinct values of G_1. Since f is injective on $\mathrm{range}(G_1)$, this means that there are at least three distinct values of x for which $f(y + \gamma x) - \gamma f(x)$ must be constant (and equal to $R_2(s, a, s')$), which is a contradiction. □

Theorem D.6. *For any non-constant reward function R_1 and any constant $\alpha \neq 0$, there is no reward function R_2 such that $G_2(\xi) = -e^{\alpha G_1(\xi)}$ for all valid trajectories ξ.*

Proof. With $f(x) = -e^{\alpha x}$, the expression in Lemma D.5 becomes $-e^{\alpha(y+\gamma x)}+\gamma e^{\alpha x}$. The derivative of this expression with respect to x is $\gamma\alpha(-e^{\alpha(y+\gamma x)} + e^{\alpha x})$, which has only one root when $\gamma \neq 0$ and $\alpha \neq 0$. This means that there can be at most two distinct values x_1, x_2 such that $-e^{\alpha(y+\gamma x_1)} + \gamma e^{\alpha x_1} = -e^{\alpha(y+\gamma x_2)} + \gamma e^{\alpha x_2}$. Since $-e^{\alpha x}$ is injective, we can thus apply Lemma D.5, which completes the proof. □

Theorem D.7. *For any non-constant reward function R_1 and any constant $\alpha > 0$, $\alpha \neq 1$, there is no reward function R_2 such that $G_2(\xi) = G_1(\xi)^{1-\alpha}$ for all valid trajectories ξ.*

Proof. With $f(x) = x^{1-\alpha}$, the expression in Lemma D.5 becomes $(y + \gamma x)^{(1-\alpha)} - \gamma x^{1-\alpha}$. The derivative of this expression with respect to x is $\gamma(\alpha - 1)(x^{-\alpha} - (\gamma x + y)^{-\alpha})$, which has only one root when $\gamma \neq 0$ and $\alpha \notin \{0, 1\}$. This means that there can be at most two distinct values x_1, x_2 such that $(y + \gamma x_1)^{(1-\alpha)} - \gamma x_1^{1-\alpha} = (y + \gamma x_2)^{(1-\alpha)} - \gamma x_2^{1-\alpha}$. Since $x^{1-\alpha}$ is injective, we can thus apply Lemma D.5, which completes the proof. □

Theorem D.8. *For any non-constant reward function R_1, there is no reward function R_2 such that $G_2(\xi) = \ln(G_1(\xi))$ for all valid trajectories ξ.*

Proof. With $f(x) = \ln(x)$, the expression in Lemma D.5 becomes $\ln(y + \gamma x) - \gamma \ln(x)$. The derivative of this expression with respect to x is $\gamma(1/(y + \gamma x) - 1/x)$, which has only one root when $\gamma \neq 0$. Since $\ln(x)$ is injective, we can thus apply Lemma D.5, which completes the proof. □

Theorem D.9. *For any non-constant reward function R_1 and any $\alpha > 0$ where $\max_\xi G_1(\xi) \leq \frac{1}{2\alpha}$, there is no reward function R_2 such that $G_2(\xi) = G_1(\xi) - \alpha G_1(\xi)^2$ for all ξ.*

Proof. With $f(x) = x - \alpha x^2$, the expression in Lemma D.5 becomes $y + \gamma x - \alpha(y + \gamma x)^2$. This is a second-degree polynomial, which means that there can be at most two distinct values x_1, x_2 such that $y + \gamma x_1 - \alpha(y + \gamma x_1)^2 = y + \gamma x_2 - \alpha(y + \gamma x_2)^2$. Moreover, if $\max_\xi G_1(\xi) \leq \frac{1}{2\alpha}$ then $f(x) = x - \alpha x^2$ is injective on $\text{range}(G_1)$. We can thus apply Lemma D.5. □

D.3 Modal Reinforcement Learning

Theorem D.10. *For any modal reward $R^\Diamond$ and any transition function τ, there exists a reward function R that is contingently equivalent to $R^\Diamond$ given τ. Moreover, unless $R^\Diamond$ is trivial, there is no reward function that is robustly equivalent to $R^\Diamond$.*

Proof. This is straightforward. For the first part, simply let $R(s, a, s') = R^\Diamond(s, a, s', \tau)$. The second part is immediate from the definition of trivial modal reward functions. □