Misspecification in Inverse Reinforcement Learning

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 aim of Inverse Reinforcement Learning (IRL) is to infer a reward function R from a policy π. To do this, we need a model of how π relates to R. In the current literature, the most common models are *optimality*, *Boltzmann rationality*, and *causal entropy maximisation*. One of the primary motivations behind IRL is to infer human preferences from human behaviour. However, the true relationship between human preferences and human behaviour is much more complex than any of the models currently used in IRL. This means that they are *misspecified*, which raises the worry that they might lead to unsound inferences if applied to real-world data. In this paper, we provide a mathematical analysis of how robust different IRL models are to misspecification, and answer precisely how the demonstrator policy may differ from each of the standard models before that model leads to faulty inferences about the reward function R. We also introduce a framework for reasoning about misspecification in IRL, together with formal tools that can be used to easily derive the misspecification robustness of new IRL models.

1 Introduction

Inverse Reinforcement Learning (IRL) is an area of machine learning concerned with inferring what objective an agent is pursuing based on the actions taken by that agent (Ng and Russell 2000). An IRL algorithm must make assumptions about how the preferences of an agent relate to its behaviour. Most IRL algorithms are based on one of three models; *optimality*, *Boltzmann rationality*, or *causal entropy maximisation*. These behavioural models are very simple, whereas the true relationship between a person's preferences and their actions of course is incredibly complex. In fact, there are observable differences between human data and data synthesised using these standard assumptions (Orsini et al. 2021). This means that the behavioural models are *misspecified*, which raises the concern that they might systematically lead to flawed inferences if applied to real-world data.

In this paper, we study how robust the behavioural models in IRL are to misspecification. To do this, we first introduce a theoretical framework for analysing misspecification robustness in IRL. We then derive a number of formal tools for inferring the misspecification robustness of IRL models, and apply these tools to exactly characterise what forms of misspecification the standard IRL models are (or are not) robust to. Our analysis is general, as it is carried out in terms of *behavioural models*, rather than *algorithms*, which means that our results will apply to any algorithm based on these models. Moreover, the tools we introduce can also be used to easily derive the misspecification robustness of new behavioural models, beyond those we consider in this work.

1.1 Related Work

It is well-known that the standard behavioural models of IRL are misspecified in most applications. However, there has nonetheless so far not been much research on this topic. Freedman, Shah, and Dragan 2021 study the effects of *choice set misspecification* in IRL (and reward inference more

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

broadly), following the formalism of Jeon, Milli, and Dragan 2020. Our work is wider in scope, and aims to provide necessary and sufficient conditions which fully describe the kinds of misspecification to which each behavioural model is robust. In the field of statistics more broadly, misspecification is a widely studied issue White 1994.

There has been a lot of work on *reducing* misspecification in IRL. One approach to this is to manually add more detail to the models (Evans, Stuhlmueller, and Goodman 2015; Chan, Critch, and Dragan 2019), and another approach is to try to *learn* the behavioural model from data (Armstrong and Mindermann 2019; Shah et al. 2019). In contrast, our work aims to understand how sensitive IRL is to misspecification (and thus to answer how much misspecification has to be removed).

1.2 Preliminaries

We assume the reader to be familiar with the basics of reinforcement learning, which can be found in Sutton and Barto 2018. A summary of our choice of notation can be found in Appendix C. In this paper, we assume that all states are reachable, and that the set of states and actions both are finite.

An IRL algorithm also needs a *behavioural model* of how π relates to R. In the current IRL literature, the most common models are:

1. *Optimality*: We assume that π is optimal under R (e.g. Ng and Russell 2000).
2. *Boltzmann Rationality*: We assume that $\mathbb{P}(\pi(s) = a) \propto e^{\beta Q^\star(s,a)}$, where β is a temperature parameter (e.g. Ramachandran and Amir 2007).
3. *Maximal Causal Entropy*: We assume that π maximises the causal entropy objective, which is given by $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t, s_{t+1}) + \alpha H(\pi(s_{t+1})))]$, where α is a weight and H is the Shannon entropy function (e.g. Ziebart 2010).

Finally, we will also refer to several ways to transform reward functions. First recall *potential shaping* Ng, Harada, and Russell 1999:

Definition 1.1 (Potential Shaping)**.** A *potential function* is a function $\Phi : \mathcal{S} \to \mathbb{R}$, where $\Phi(s) = 0$ if s is a terminal state. Given a discount γ, we say that $R_2 \in \mathcal{R}$ is produced by *potential shaping* of $R_1 \in \mathcal{R}$ if for some potential Φ, $R_2(s, a, s') = R_1(s, a, s') + \gamma \cdot \Phi(s') - \Phi(s)$.

Potential shaping is widely used for reward shaping. We next define two classes of transformations that were used by Skalse et al. 2022, starting with S'*-redistribution*.

Definition 1.2 (S'-Redistribution)**.** Given a transition function τ, $R_2 \in \mathcal{R}$ is produced by S'-*redistribution* of $R_1 \in \mathcal{R}$ if $\mathbb{E}_{S' \sim \tau(s,a)} [R_1(s, a, S')] = \mathbb{E}_{S' \sim \tau(s,a)} [R_2(s, a, S')]$.

We next consider *optimality-preserving transformations*:

Definition 1.3. Given a transition function τ and a discount γ, we say that $R_2 \in \mathcal{R}$ is produced by an *optimality-preserving transformation* of $R_1 \in \mathcal{R}$ if there exists a function $\psi : \mathcal{S} \to \mathbb{R}$ such that $\mathbb{E}_{S' \sim \tau(s,a)}[R_2(s, a, S') + \gamma \cdot \psi(S')] \leq \psi(s)$, with equality if and only if $a \in \text{argmax}_{a \in \mathcal{A}} A_1^\star(s, a)$.

2 Theoretical Framework

We here introduce the theoretical framework that we will use to analyse how robust various behavioural models are to misspecification. For a given set of states $\mathcal{S}$ and a given set of actions $\mathcal{A}$, let $\mathcal{R}$ be the set of all reward functions $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. We will use the following definitions:

1. A *reward object* is a function $f : \mathcal{R} \to X$, where X is any set.
2. Given partition P of $\mathcal{R}$, we say that f is P*-admissible* if $f(R_1) = f(R_2) \implies R_1 \equiv_P R_2$.
3. Given a partition P of $\mathcal{R}$, we say that f is P*-robust to misspecification* with g if f is P-admissible, $f \neq g$, $\text{Im}(g) \subseteq \text{Im}(f)$, and $f(R_1) = g(R_2) \implies R_1 \equiv_P R_2$.
4. A *reward transformation* is a function $t : \mathcal{R} \to \mathcal{R}$.

These definitions give us a way to analyse misspecification robustness in the limit of infinite data. We provide some justification and intuition for these definitions in Appendix B. Next, we give a fundamental lemma that we will later use to prove our core results. All of our proofs are provided in the appendix, which also contains several additional results about our framework.

Lemma 2.1. *If f is P-admissible, and T is the set of all reward transformations that preserve P, then f is P-robust to misspecification with g if and only if $g = f \circ t$ for some $t \in T$ where $f \circ t \neq f$.*

This lemma gives us a very powerful tool for characterising the misspecification robustness of reward objects. Specifically, we can derive the set of objects to which f is P-robust by first deriving the set T of all transformations that preserve P, and then composing f with each $t \in T$.

3 Reward Function Equivalence Classes

Our definition of misspecification robustness is given relative to an equivalence relation on $\mathcal{R}$. In this section, we characterise two important equivalence classes. Given an environment $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma \rangle$ and two reward functions R_1, R_2, we say that $R_1 \equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$ if $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma \rangle$ and $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma \rangle$ have the same *optimal* policies, and that $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ if they have the same *ordering* of policies. [1] Skalse et al. 2022 showed that $R_1 \equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$ if and only if R_1 and R_2 differ by an optimality-preserving transformation (their Theorem 3.16). We characterise the transformations that preserve $\mathrm{ORD}^{\mathcal{M}}$, which is a novel contribution.

Theorem 3.1. *$R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ if and only if R_1 and R_2 differ by potential shaping, S'-redistribution, and positive linear scaling (applied in any order).*

This theorem fully characterises when R_1 and R_2 have the same ordering of policies in a given MDP. It is also worth noting that this directly implies that R_1 and R_2 have the same ordering of policies for all τ if and only if they differ by potential shaping and positive linear scaling.

4 Misspecification Robustness of Behavioural Policies

We here give our main results on the misspecification robustness of IRL, starting with the Boltzmann-rational model. Let Π^+ be the set of all policies such that $\pi(a \mid s) > 0$ for all s, a, let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma \rangle$, and let $F^{\mathcal{M}}$ be the set of all functions $f^{\mathcal{M}} : \mathcal{R} \to \Pi^+$ that, given R, returns a policy π which satisfies $\mathrm{argmax}_{a \in \mathcal{A}} \pi(a \mid s) = \mathrm{argmax}_{a \in \mathcal{A}} Q^\star(s, a)$, where $Q^\star$ is the optimal Q-function in $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$. In other words, $F^{\mathcal{M}}$ is the set of functions that generate policies which take each action with positive probability, and that take the optimal actions with the highest probability. This class includes e.g. Boltzmann-rational policies (for any β).

Theorem 4.1. *Let $f^{\mathcal{M}} \in F^{\mathcal{M}}$ be surjective onto Π^+. Then f^M is $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with g if and only if $g \in F^{\mathcal{M}}$ and $g \neq f^{\mathcal{M}}$.*

Boltzmann-rational policies are surjective onto Π^+,[2] so Theorem 4.1 exactly characterises the misspecification to which the Boltzmann-rational model is $\mathrm{OPT}^{\mathcal{M}}$-robust.

We next turn our attention to the misspecification to which the Boltzmann-rational model is $\mathrm{ORD}^{\mathcal{M}}$-robust. Let $\psi : \mathcal{R} \to \mathbb{R}^+$ be any function from reward functions to positive real numbers, and let $b_\psi^{\mathcal{M}} : \mathcal{R} \to \Pi^+$ be the function that, given R, returns the Boltzmann-rational policy with temperature $\psi(R)$ in $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$. Moreover, let $B^{\mathcal{M}} = \{b_\psi^{\mathcal{M}} : \psi \in \mathcal{R} \to \mathbb{R}^+\}$ be the set of all such functions $b_\psi^{\mathcal{M}}$. This set includes Boltzmann-rational policies; just let ψ return a constant β for all R.

Theorem 4.2. *If $b_\psi^{\mathcal{M}} \in B^{\mathcal{M}}$ then $b_\psi^{\mathcal{M}}$ is $\mathrm{ORD}^{\mathcal{M}}$-robust to misspecification with g if and only if $g \in B^{\mathcal{M}}$ and $g \neq b_\psi^{\mathcal{M}}$.*

This means that the Boltzmann-rational model is $\mathrm{ORD}^{\mathcal{M}}$-robust to misspecification of the temperature parameter β, but not to any other form of misspecification.

We next turn our attention to optimal policies. First of all, a policy is optimal if and only if it only gives support to optimal actions, and if an optimal policy gives support to multiple actions in some state, then we would normally not expect the exact probability it assigns to each action to convey any information about the reward function. We will therefore only look at the actions that the optimal

[1] By this, we mean that $\mathcal{J}_1(\pi) > \mathcal{J}_1(\pi')$ if and only if $\mathcal{J}_2(\pi) > \mathcal{J}_2(\pi')$, for all pairs of policies π, π'.

[2] If a policy π takes each action with positive probability, then its action probabilities are always the softmax of some Q-function, and any Q-function corresponds to some reward function.

policy takes, and ignore the relative probability it assigns to those actions. Formally, we will treat optimal policies as functions $\pi_\star : \mathcal{S} \to \mathcal{P}(\text{argmax}_{a \in \mathcal{A}} A^\star) - \{\varnothing\}$; i.e. as functions that for each state return a non-empty subset of the set of all actions that are optimal in that state. Let $\mathcal{O}^\mathcal{M}$ be the set of all functions that return such policies, and let $o_m^\mathcal{M} \in \mathcal{O}^\mathcal{M}$ be the function that, given R, returns the function that maps each state to the set of *all* actions which are optimal in that state. Intuitively, $o_m^\mathcal{M}$ corresponds to optimal policies that take all optimal actions with positive probability.

Theorem 4.3. *No function in $\mathcal{O}^\mathcal{M}$ is $\mathrm{ORD}^\mathcal{M}$-admissible. The only function in $\mathcal{O}^\mathcal{M}$ that is $\mathrm{OPT}^\mathcal{M}$-admissible is $o_m^\mathcal{M}$, but $o_m^\mathcal{M}$ is not $\mathrm{OPT}^\mathcal{M}$-robust to any misspecification.*

This essentially means that the optimality model is not robust to any form of misspecification. We finally turn our attention to causal entropy maximising policies. As before, let $\psi : \mathcal{R} \to \mathbb{R}^+$ be any function from reward functions to positive real numbers, and let $c_\psi^\mathcal{M} : \mathcal{R} \to \Pi^+$ be the function that, given R, returns the causal entropy maximising policy with weight $\psi(R)$ in $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$. Furthermore, let $C^\mathcal{M} = \{c_\psi^M : \psi \in \mathcal{R} \to \mathbb{R}^+\}$ be the set of all such functions $c_\psi^\mathcal{M}$. This set includes causal entropy maximising policies; just let ψ return a constant α for all R.

Theorem 4.4. *If $c_\psi^\mathcal{M} \in C^\mathcal{M}$ then $c_\psi^\mathcal{M}$ is $\mathrm{ORD}^\mathcal{M}$-robust to misspecification with g if and only if $g \in C^\mathcal{M}$ and $g \neq c_\psi^\mathcal{M}$.*

In other words, the maximal causal entropy model is $\mathrm{ORD}^\mathcal{M}$-robust to misspecification of the weight α, but not to any other kind of misspecification.

Finally, let us briefly discuss the misspecification to which the maximal causal entropy model is $\mathrm{OPT}^\mathcal{M}$-robust. Lemma 2.1 tells us that $c_\psi^\mathcal{M} \in C^\mathcal{M}$ is $\mathrm{OPT}^\mathcal{M}$-robust to misspecification with g if $g = c_\psi^\mathcal{M} \circ t$ for some transformation $t : \mathcal{R} \to \mathcal{R}$ that preserves optimal policies in $\mathcal{M}$. In other words, if $g(R_1) = \pi$ then there must exist an R_2 such that π maximises causal entropy with respect to R_2, and such that R_1 and R_2 have the same optimal policies. It seems hard to express this as an intuitive property of g, so we have refrained from stating this result as a theorem.

5 Extensions

The analysis in Section 4 can be extended in several ways. In Appendix D.2, we analyse what happens if $\mathcal{R}$ is restricted to some subset of all possible reward functions; we find that this does not fundamentally change any of the results in Section 4. In Appendix D.1, we analyse misspecification of the MDP dynamics. In Appendix D.3, we discuss how to analyse the case when we have known priors concerning R, and in Appendix D.4, we discuss the issue of transfer to new environments.

6 Discussion

We have shown how to exactly characterise the misspecification robustness of the behavioural models in IRL. First, for $\mathrm{ORD}^\mathcal{M}$-robustness, we find that the Boltzmann-rational model is robust to misspecification of the temperature parameter, that the maximal causal entropy model is robust to misspecification of the weight parameter, and that the optimality model lacks any misspecification robustness. Next, for $\mathrm{OPT}^\mathcal{M}$-robustness, we find that the Boltzmann-rational model only requires that the observed policy always takes the most valuable action with the highest probability, that the optimality model again lacks any misspecification robustness, and that the maximal causal entropy model is robust to some misspecification, but that it is hard to define it in an intuitive way. It is noteworthy that no model is robust to misspecification of the discount parameter, γ. In addition to these contributions, we have also provided several formal tools for deriving the misspecification robustness of new behavioural models, in the form of the lemmas in Section 2.

Our analysis makes a few simplifying assumptions, that could be ideally lifted in future work. First of all, we have been working with *equivalence relations* on $\mathcal{R}$: it might be fruitful to instead consider *distance metrics* on $\mathcal{R}$. Another notable direction for extensions could be to study the misspecification robustness in the context where we have particular priors concerning R. Finally, we have studied the behaviour of algorithms in the limit of infinite data. Another possible extension could be to more rigorously examine the properties of these models in the case of finite data.

References

[1] Stuart Armstrong and Sören Mindermann. *Occam's razor is insufficient to infer the preferences of irrational agents*. 2019. arXiv: 1712.05812 [cs.AI].

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

[3] Lawrence Chan, Andrew Critch, and Anca Dragan. *Irrationality can help reward inference*. 2019. URL: https://openreview.net/pdf?id=BJlo91BYPr.

[4] Owain Evans, Andreas Stuhlmueller, and Noah D. Goodman. *Learning the Preferences of Ignorant, Inconsistent Agents*. 2015. arXiv: 1512.05832 [cs.AI].

[5] Rachel Freedman, Rohin Shah, and Anca Dragan. *Choice Set Misspecification in Reward Inference*. 2021. DOI: 10.48550/ARXIV.2101.07691. URL: https://arxiv.org/abs/2101.07691.

[6] Evan Hubinger et al. *Risks from Learned Optimization in Advanced Machine Learning Systems*. 2019. DOI: 10.48550/ARXIV.1906.01820. URL: https://arxiv.org/abs/1906.01820.

[7] Hong Jun Jeon, Smitha Milli, and Anca Dragan. "Reward-rational (implicit) choice: A unifying formalism for reward learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 4415–4426. URL: https://proceedings.neurips.cc/paper/2020/file/2f10c1578a0706e06b6d7db6f0b4a6af-Paper.pdf.

[8] 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.

[9] Andrew Y Ng and Stuart Russell. "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Vol. 1. Stanford, California, USA: Morgan Kaufmann Publishers Inc, 2000, pp. 663–670.

[10] Manu Orsini et al. "What Matters for Adversarial Imitation Learning?" In: *arXiv preprint* arXiv:2106.00672 [cs.LG] (2021). To appear in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[11] Deepak Ramachandran and Eyal Amir. "Bayesian Inverse Reinforcement Learning". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. Hyderabad, India: Morgan Kaufmann Publishers Inc, 2007, pp. 2586–2591.

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

[13] Rohin Shah et al. *On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference*. 2019. arXiv: 1906.09624 [cs.LG].

[14] Joar Skalse et al. "Invariance in Policy Optimisation and Partial Identifiability in Reward Learning". In: *arXiv preprint arXiv:2203.07475* (2022).

[15] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. second. MIT Press, 2018. ISBN: 9780262352703.

[16] Halbert White. *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press, 1994. DOI: 10.1017/CCOL0521252806.

[17] 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_.

[18] Brian D Ziebart. "Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy". PhD thesis. Carnegie Mellon University, 2010.

A X-Risk Analysis

We believe that the work presented in this paper could contribute to decreasing the risk of an existential threat from advanced AI. In this section, we aim to provide all the core assumptions behind this belief. We also discuss whether our research might fail to decrease, or even increase, this risk.

A simple arguments for why AI could be a source of existential risk can be laid out as follows:

1. An advanced AI can be well-modelled as optimising the world according to some criterion.
2. If some criterion other than human values were to be optimised with sufficiently great optimisation power, then that would lead to catastrophic consequences.
3. Human values are very difficult to specify.

Each of these points seem reasonably likely, and if one accepts all three, then it follows that we must develop methods for aligning AI systems with human values, before we develop AI systems which are too advanced. The version of the argument which we give above is greatly simplified, but a more extensive and nuanced version can be found in e.g. Bostrom 2014 and Russell 2019. Another noteworthy recent reference is Yudkowsky 2022.

The next question is, of course, how we can develop methods for aligning AI systems with human values. Call this the *alignment problem*. It seems reasonably likely that a solution to the alignment problem must include a solution to the problem of how to specify human values. Of course, the alignment problem also includes other problems, see e.g. Hubinger et al. 2019. Moreover, there might be ways to solve the alignment problem without figuring out how to specify human values – we will discuss this possibility further down. But, nonetheless, it seems at least reasonably likely that solving the alignment problem requires first solving the problem of how to specify human values.

The next question is how we can specify human values. It seems incredibly (and probably prohibitively) difficult to specify this directly. A hope is therefore that we might be able to *learn* human values from data, using ML. In particular, there is hope that this problem could be solved using IRL (e.g. Russell 2019). The next issue which presents itself is that human values are *unobservable*, which means that they only can be a latent variable in any kinds of data we could collect. This is true of IRL, and also of any other learning method we might device. This, in turn, means that an ML method designed to learn human values must make assumptions about the relationship between human values and the observed data. For example, an IRL algorithm must make assumptions about how human values relate to human behaviour, and so on.

The next question is then how to specify a model of the relationship between human values and some appropriate source of data. At first, it was hoped that this model might itself be learnable using RL. However, Armstrong and Mindermann 2019 demonstrated that the task of simultaneously learning both a model of a persons preferences, and a model of their rationality, from a single stream of data, is impossible. Their paper was written with a focus on IRL, and assumed that the learning algorithm would use a joint simplicity prior, but the proof strategy generalises far beyond these two assumptions. Another possibility might be to first learn a model of human rationality, and then in a separate step learn a model of human preferences. This approach might be possible, but also faces many difficulties in practice, see Shah et al. 2019. In short, in order to learn a model of how a persons preferences relate to their behaviour, you would have to find situations where you are sure of their preferences, and then measure their behaviour. However, those situations are rare, if they even exist at all. This means that this approach will put a lot of pressure on the ability of the rationality model to generalise correctly far outside the training distribution, which seems unreliable.

We thus have two options; try to correctly specify a model of human rationality, or try to learn a correct model, in spite of the difficulties we just described. Neither option seems very promising. However, there is a crucial question which the argument has so far overlooked, namely, *how robust is the value learning problem to misspecification of the relationship between human values and the observed data*? In particular, how robust is IRL to misspecification of the behavioural model? If the answer is that this inference problem is very sensitive to such misspecification, then it seems unlikely that we will be able to create a behavioural model of sufficient quality, regardless of whether we try to specify it directly, or learn it using ML. In that case, we should give up on value learning, and any approach to solving the alignment problem which relies on it. If, on the other hand, it turns out that IRL is very robust to misspecification of the behavioural model, then the situation would instead be very hopeful. In the more extreme case, it might even be enough to just use a broadly plausible model of bounded rationality. In that case, we should expect to be able to construct such a model, and to able to learn human values with IRL.

This paper is a first step towards answering this question, and determining how robust the value learning problem is to misspecification. It is not yet a solution to that question, but it is progress on the way. An answer to this question will be advantageous for guiding future research, and determining what strategies for value alignment are more promising or less promising. Moreover, it will also be necessary in order to *trust* the result of any value learning method — misspecification is ubiquitous in reward learning, and while it can be decreased, it cannot be eliminated. Therefore, how reliable a reward learning method is will invariably in part depend on how robust the value learning problem is to misspecification. This explains the context of this paper within the broader research landscape on risk from advanced AI.

The next question is whether our research might fail to contribute to reducing existential risk from advanced AI, or even increase it. We will start with the latter question, by providing a story of how our research could be net harmful. First, the notion of "human values" is incredibly subtle, as is the question of what makes an outcome *good* or *bad*. After all, this is why Ethics is a field of study. One could worry that the existing field of value learning conflates the question of *what is good* with the question of *what is in accordance with a person's current desires*. That is, one could worry that it conflates *human values* and *human preferences*. Moreover, one could then also worry that aligning an advanced AI with the latter essentially would amount to misalignment. If this is true, then the risk of this outcome is increased by any research that improves the capabilities of reward learning methods, and by any research that instils false confidence in these methods. This paper could plausibly belong to the latter category. At least, this is the most plausible scenario we can see in which this paper would end up being net harmful. Then, there is of course also opportunity cost. For example, it is possible that value alignment is nearly impossible, and that the best way to solve the alignment problem is to try to develop "docile" AI systems which do not attempt to optimise the world (i.e., an AI which defies assumption 1 in the argument above). In that case, any research which draws attention away from this strategy, and towards e.g. value learning, could be seen as net bad. However, on the whole, we still expect our research to be a positive contribution to the task of solving the alignment problem.

B Intuition

In this appendix, we explain and justify each of the definitions in Section 2. First of all, anything that can be computed from a reward function can be seen as a reward object. For example, we could consider a function b that, given a reward R, returns the Boltzmann-rational policy with temperature β in the MDP $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$, or a function r that, from R, gives the return function G in the MDP $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma \rangle$. This makes reward objects a versatile abstract building block for more complex constructions. We will mainly, but not exclusively, consider reward objects with the type $\mathcal{R} \to \Pi$, i.e. functions that compute policies from rewards.

We can use reward objects to create an abstract model of a reward learning algorithm $\mathcal{L}$ as follows; first, we assume, as reasonable, that there is a true underlying reward function $R^\star$, and that the observed training data is generated by a reward object g, so that $\mathcal{L}$ observes $g(R^\star)$. Here $g(R^\star)$ could be a *distribution*, which models the case where $\mathcal{L}$ observes a sequence of random samples from some source, but it could also be a single, finite object. Next, we suppose that $\mathcal{L}$ has a model f of how the observed data relates to $R^\star$, where f is also a reward object, and that $\mathcal{L}$ learns (or converges to) a reward function R_H such that $f(R_H) = g(R^\star)$. If $f \neq g$ then f is *misspecified*, otherwise f is correctly specified. Note that this primarily is a model of the *asymptotic* behaviour of learning algorithms, in the limit of *infinite data*.

There are two ways to interpret $\mathrm{Am}(f)$. First, we can see it as a bound on the amount of information we can get about $R^\star$ by observing (samples from) $f(R^\star)$. For example, multiple reward functions might result in the same Boltzmann-rational policy, thus observing trajectories from that policy could never let us distinguish between them: this ambiguity is described by $\mathrm{Am}(b)$. We can also see $\mathrm{Am}(f)$ as the amount of information we need to have about $R^\star$ to construct $f(R^\star)$. Next, if $f \preceq g$, this means that we get less information about $R^\star$ by observing $g(R^\star)$ than $f(R^\star)$, and that we would need more information to construct $f(R^\star)$ than $g(R^\star)$. For an extensive discussion about these notions, see Skalse et al. 2022.

Intuitively, we want to say that a behavioural model is robust to some type of misspecification if an algorithm based on that model will learn a reward function that is "close enough" to the true reward function when subject to that misspecification. To formalise this intuitive statement, we first need a definition of what it should mean for two reward functions to be "close enough". In this work, we have chosen to define this in terms of *equivalence classes*. Specifically, we assume that we have a partition P of $\mathcal{R}$ (which, of course, corresponds to an equivalence relation), and that the learnt reward function R_H is "close enough" to the true reward function $R^\star$ if they're in the same class,

$R_H \equiv_P R^\star$. We will for now leave open the question of which partition P of $\mathcal{R}$ to pick, and later revisit this question in Section 3.

Given this, we can now see that our definition of P-admissibility is equivalent to stating that a learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that is P-equivalent to the true reward function when there is no misspecification. Furthermore, our definition of P-robustness says that f is P-robust to misspecification with g if any learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that is P-equivalent to the true reward function when trained on data generated from g. The requirement that $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ ensures that the learning algorithm $\mathcal{L}$ is never given data that is impossible according to its model. Depending on how $\mathcal{L}$ reacts to such data, it may be possible to drop this requirement. We include it, since we want our analysis to apply to all algorithms. The requirement that f is P-admissible is included to rule out some uninteresting edge cases.

Reward transformations can be used to characterise the ambiguity of reward objects, or define other partitions of $\mathcal{R}$. Specifically, we say that a partition P corresponds to a set of reward transformations T_P if T_P contains all reward transformations t that satisfy $t(R) \equiv_P R$. If P is the ambiguity of f then T_P would be the set of all reward transformations that satisfy $f(R) = f(t(R))$. Note that if T corresponds to a partition of $\mathcal{R}$ then T must contain the identity map, be closed under composition, and contain inverses.

C Notation

In this appendix, we list all of our notation. A *Markov Decision Processes* (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R, \gamma)$ 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*, $\mu_0 \in \Delta(\mathcal{S})$ is an *initial state distribution*, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a *reward function* and $\gamma \in (0, 1]$ is a *discount rate*. Here $f : X \rightsquigarrow Y$ denotes a probabilistic mapping from X to Y. In this paper, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite. A state s is *terminal* if $\tau(s, a) = s$ and $R(s, a, s') = 0$ for all a, s'. A *policy* is a function $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$. A *trajectory* $\xi = \langle s_0, a_0, s_1, a_1 \dots \rangle$ is a possible path in an MDP. The *return function* G gives the cumulative discounted reward of a trajectory, $G(\xi) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})$, and the *evaluation function* $\mathcal{J}$ gives the expected trajectory return given a policy, $\mathcal{J}(\pi) = \mathbb{E}_{\xi \sim \pi}[G(\xi)]$. A policy maximising $\mathcal{J}$ is an *optimal policy*. The *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ of a policy encodes the expected future discounted reward from each state when following that policy. The Q-function is $Q^\pi(s, a) = \mathbb{E}[R(s, a, S') + \gamma V^\pi(S')]$, and the *advantage function* is $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$. $Q^\star$, $V^\star$, and $A^\star$ denote the Q-, value, and advantage functions of the optimal policies. In this paper, we assume that all states in S are reachable under τ and μ_0. Moreover, we will often talk about pairs or sets of reward functions. In these cases, we will give each reward function a subscript R_i, and use $\mathcal{J}_i$, $V_i^\star$, and V_i^π, and so on, to denote R_i's evaluation function, optimal value function, and π value function, and so on.

D Generalising the Analysis

In this appendix, we discuss different ways to generalise our results.

D.1 Misspecified MDPs

A reward object can be parameterised by a γ or τ, implicitly or explicitly. For example, the reward objects in Section 4 are parameterised by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma \rangle$. In this section, we explore what happens if these parameters are misspecified. We show that nearly all behavioural models are sensitive to this type of misspecification.

Theorems 4.1-4.4 already tell us that the standard behavioural models are not ($\mathrm{ORD}^{\mathcal{M}}$ or $\mathrm{OPT}^{\mathcal{M}}$) robust to misspecified γ or τ, since the sets $F^{\mathcal{M}}$, $B^{\mathcal{M}}$, and $C^{\mathcal{M}}$, all are parameterised by γ and τ. We will generalise this further. To do this, we first derive two lemmas. We say that τ is *trivial* if for each $s \in \mathcal{S}$, $\tau(s, a) = \tau(s, a')$ for all $a, a' \in \mathcal{A}$.

Lemma D.1. *If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ then f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible for $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau_2, \mu_0, \_, \gamma \rangle$ unless $\tau_1 = \tau_2$.*

Lemma D.2. *If $f^{\gamma_1} = f^{\gamma_1} \circ t$ for all $t \in \mathrm{PS}_{\gamma_1}$ then f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible for $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma_2 \rangle$ unless $\gamma_1 = \gamma_2$ or τ is trivial.*

Note that if f is not $\mathrm{OPT}^{\mathcal{M}}$-admissible then f is also not $\mathrm{ORD}^{\mathcal{M}}$-admissible, and similarly for misspecification robustness. From these lemmas, we get the following result:

Theorem D.3. *If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ then f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{τ_2} for any $\mathcal{M}$. Moreover, if $f^{\gamma_1} = f^{\gamma_1} \circ t$ for all $t \in \mathrm{PS}_{\gamma_1}$ then f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{γ_2} for any $\mathcal{M}$ whose transition function τ is non-trivial.*

The statement of this theorem can be explained in words as follows. The first part shows that if a behavioural model says that the policy is insensitive to S'-redistribution, then that model is not $\mathrm{OPT}^{\mathcal{M}}$-robust (and therefore also not $\mathrm{ORD}^{\mathcal{M}}$-robust) to misspecification of the transition function τ. Similarly, the second part shows that if the behavioural model says that the policy is insensitive to potential shaping, then that model is not $\mathrm{OPT}^{\mathcal{M}}$-robust (and therefore also not $\mathrm{ORD}^{\mathcal{M}}$-robust) to misspecification of the discount parameter γ. Note that all transformations in $S'\mathrm{R}_\tau$ and PS_γ preserve the ordering of policies. This means that an IRL algorithm must specify τ and γ correctly in order to guarantee that the learnt reward function R_H has the same optimal policies as the true underlying reward function R^*, unless the algorithm is based on a behavioural model which says that the observed policy depends on features of R which do not affect its policy ordering. This should encompass most natural behavioural models.

That being said, we note that this result relies on the requirement that the learnt reward function should have *exactly* the same optimal policies, or ordering of policies, as the true reward function. If $\gamma_1 \approx \gamma_2$ and $\tau_1 \approx \tau_2$, then the learnt reward function's optimal policies and policy ordering will presumably be *similar* to that of the true reward function. Analysing this case is beyond the scope of this paper, but we consider it to be an important topic for further work.

D.2 Restricted Reward Functions

Here, we discuss what happens if the reward function is restricted to belong to some subset of $\mathcal{R}$, i.e. if we know that $R \in \hat{\mathcal{R}}$ for some $\hat{\mathcal{R}} \subseteq \mathcal{R}$. For example, it is common to consider reward functions that are linear in some state features. It is also common to define the reward function over a restricted domain, such as $\mathcal{S}\times\mathcal{A}$; this would correspond to restricting $\mathcal{R}$ to the set of reward functions such that $R(s, a, s') = R(s, a, s'')$ for all s, a, s', s''. As we will see, our results are largely unaffected by such restrictions.

We first need to generalise the framework, which is straightforward. Given partitions P, Q of $\mathcal{R}$, reward objects f, g, and set $\hat{\mathcal{R}} \subseteq \mathcal{R}$, we say that $P \preceq Q$ on $\hat{\mathcal{R}}$ if $R_1 \equiv_P R_2$ implies $R_1 \equiv_Q R_2$ for all $R_1, R_2 \in \hat{\mathcal{R}}$, that f is P-admissible on $\hat{\mathcal{R}}$ if $\mathrm{Am}(f) \preceq P$ on $\hat{\mathcal{R}}$, and that f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ if f is P-admissible on $\hat{\mathcal{R}}$, $f|_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$, $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$, and $f(R_1) = g(R_2) \implies R_1 \equiv_P R_2$ for all $R_1, R_2 \in \hat{\mathcal{R}}$.

The theorems in Section 4 also carry over very directly:

Theorem D.4. *If f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ then f is P-robust to misspecification with g' on $\mathcal{R}$ for some g' where $g'|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, unless f is not P-admissible on $\mathcal{R}$. If f is P-robust to misspecification with g on $\mathcal{R}$ then f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, unless $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$.*

The intuition for this theorem is that if f is P-robust to misspecification with g if and only if $g \in G$, then f is P-robust to misspecification with g' on $\hat{\mathcal{R}}$ if and only if g' behaves like some $g \in G$ for all $R \in \hat{\mathcal{R}}$. Restricting $\mathcal{R}$ does therefore not change the problem in any significant way.

If an equivalence relation P of $\mathcal{R}$ is characterised by a set of reward transformations T, then the corresponding equivalence relation on $\hat{\mathcal{R}}$ is characterised by the set of reward transformations $\{t \in T : \mathrm{Im}(t|_{\hat{\mathcal{R}}}) \subseteq \hat{\mathcal{R}}\}$; this can be used to generalise Theorem 3.1. However, here there is a minor subtlety to be mindful of: $(A \circ B) - C$ is not necessarily equal to $(A - C) \circ (B - C)$. This means that if we wish to specify $\{t \in A \circ B : \mathrm{Im}(t|_{\hat{\mathcal{R}}}) \subseteq \hat{\mathcal{R}}\}$, then we cannot do this by simply removing the transformations where $\mathrm{Im}(t|_{\hat{\mathcal{R}}}) \not\subseteq \hat{\mathcal{R}}$ from each of A and B. For example, consider the transformations $S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma$ restricted to the space $\hat{\mathcal{R}}$ of reward functions where $R(s, a, s') = R(s, a, s'')$, i.e. to reward functions over the domain $\mathcal{S}\times\mathcal{A}$. The only transformation in $S'\mathrm{R}_\tau$ on $\hat{\mathcal{R}}$ is the identity mapping, and the only transformations in PS_γ on $\hat{\mathcal{R}}$ are those where Φ is constant over all states. However, $S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma$ on $\hat{\mathcal{R}}$ contains all transformations where Φ is selected arbitrarily, and $t(R)(s, a, s')$ is set to $R(s, a, s') + \gamma\mathbb{E}[\Phi(S')] - \Phi(s)$. This means that there probably are no general shortcuts for deriving $\{t \in T : \mathrm{Im}(t|_{\hat{\mathcal{R}}}) \subseteq \hat{\mathcal{R}}\}$ for arbitrary $\hat{\mathcal{R}}$.

It should be noted that *negative* results might *not* hold if $\mathcal{R}$ is restricted. Recall that f is not P-robust to misspecification with g if there exist R_1, R_2 such that $g(R_1) = f(R_2)$, but $R_1 \not\equiv_P R_2$. If $\mathcal{R}$ is restricted, it could be the case that all such counterexamples are removed. For example, if we restrict $\mathcal{R}$ to e.g. the set $\hat{\mathcal{R}}$ of reward functions that only reward a single transition, then Lemma D.2, and the corresponding part of Theorem D.3, no longer apply.[3] This means that, if the reward function is guaranteed to lie in this set $\hat{\mathcal{R}}$, then a behavioural model may still be $\mathrm{OPT}^{\mathcal{M}}$-robust to a misspecified discount parameter. However, the reason for this is simply that the discount parameter no longer affects which policies are optimal if there is only a single transition that has non-zero reward.

D.3 Known Prior and Inductive Bias

So far, we have assumed that we do not know which distribution R is sampled from, or which inductive bias the learning algorithm $\mathcal{L}$ has. In this section, we discuss what might happen if we lift these assumptions.

To some extent, our results from Appendix D.2 can be used to understand this setting as well. Suppose we have a set $\hat{\mathcal{R}} \subseteq \mathcal{R}$ of "likely" reward functions, such that $\mathbb{P}(R^\star \in \hat{\mathcal{R}}) = 1 - \delta$, and such that the learning algorithm $\mathcal{L}$ returns a reward function R_H in $\hat{\mathcal{R}}$ if there exists an $R_H \in \hat{\mathcal{R}}$ such that $f(R_H) = g(R^\star)$. Then if f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, it follows that $\mathcal{L}$ returns an R_H such that $R_H \equiv_P R^\star$ with probability at least $1 - \delta$.

So, for example, suppose $\hat{\mathcal{R}}$ is the set of all reward functions that are "sparse", for some way of formalising that property. Then this tells us, informally, that if the underlying reward function is likely to be sparse, and if $\mathcal{L}$ will attempt to fit a sparse reward function to its training data, then it is sufficient that f is P-robust to misspecification with g on the set of all sparse reward functions, to ensure that the learnt reward function R_H is P-equivalent to the true reward function with high probability. It seems likely that more specific claims could be made about this setting, but we leave such analysis as a topic for future work.

D.4 Transfer to New Environments

The equivalence relations we have worked with ($\mathrm{OPT}^{\mathcal{M}}$ and $\mathrm{ORD}^{\mathcal{M}}$) only guarantee that the learnt reward function R_H has the same optimal policies, or ordering of policies, as the true reward $R^\star$ in a given environment $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma \rangle$. A natural question is what happens if we strengthen this requirement, and demand that R_H has the same optimal policies, or ordering of policies, as $R^\star$, for any choice of τ, μ_0, or γ. We discuss this setting here.

In short, it is impossible to guarantee transfer to any τ or γ within our framework, and trivial to guarantee transfer to any μ_0. First, the lemmas provided in Appendix D.1 tell us that none of the standard behavioural models are $\mathrm{OPT}^{\mathcal{M}}$-admissible when τ or γ is different from that of the training environment. This means that none of them can guarantee that R_H has the same optimal policies (or ordering of policies) as $R^\star$ if τ or γ is changed, with or without misspecification. Second, if $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ or $R_1 \equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$, then this remains the case if μ_0 is changed. We can thus trivially guarantee transfer to arbitrary μ_0.

E proofs

In this Appendix, we provide the proofs of all our results, as well as of some additional lemmas.

Before giving the proofs, we must first specify several sets of reward transformations:

1. Let PS_γ be the set of all reward transformations t such that $t(R)$ is given by potential shaping of R relative to the discount γ.
2. Let $S'\mathrm{R}_\tau$ be the set of all reward transformations t such that $t(R)$ is given by S'-redistribution of R relative to the transition function τ.

[3] The reason for this is that there are no $R_1, R_2 \in \hat{\mathcal{R}}$ where $R_1 = t(R_2)$ for some $t \in \mathrm{PS}_\gamma$.

3. Let LS be the set of all reward transformations t that scale each reward function by some positive constant, i.e. for each R there is a $c \in \mathbb{R}^+$ such that $t(R)(s, a, s') = c \cdot R(s, a, s')$.
4. Let CS be the set of all reward transformations t that shift each reward function by some constant, i.e. for each R there is a $c \in \mathbb{R}$ such that $t(R)(s, a, s') = R(s, a, s') + c$.
5. Let $\mathrm{OP}_{\tau,\gamma}$ be the set of all reward transformations t such that $t(R)$ is given by an optimality-preserving transformation of R relative to τ and γ.

Note that these sets are defined in a way that allows their transformations to be "sensitive" to the reward function it takes as input. For example, a transformation $t \in \mathrm{PS}_\gamma$ might apply one potential function Φ_1 to R_1, and a different potential function Φ_2 to R_2. Similarly, a transformation $t \in \mathrm{LS}$ might scale R_1 by a positive constant c_1, and R_2 by a different constant c_2, etc.

E.1 Fundamental Lemmas

We here prove a number of results about our framework. Our proofs in this section are given relative to the somewhat more general definitions of P-robustness and refinement given in Appendix D.2, rather than those given in Section 2.

Lemma E.1. *For any f and h, if f is not P-admissible on $\hat{\mathcal{R}}$ then $h \circ f$ is not P-admissible on $\hat{\mathcal{R}}$.*

Proof. If f is not P-admissible on $\hat{\mathcal{R}}$ then there are $R_1, R_2 \in \hat{\mathcal{R}}$ such that $f(R_1) = f(R_2)$, but $R_1 \not\equiv_P R_2$. But if $f(R_1) = f(R_2)$ then $h \circ f(R_1) = h \circ f(R_2)$, so there are $R_1, R_2 \in \hat{\mathcal{R}}$ such that $h \circ f(R_1) = h \circ f(R_2)$, but $R_1 \not\equiv_P R_2$. Thus $h \circ f$ is not P-admissible on $\hat{\mathcal{R}}$. □

Lemma E.2. *If f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ then g is P-admissible on $\hat{\mathcal{R}}$.*

Proof. Suppose for contradiction that f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, but that g is not P-admissible on $\hat{\mathcal{R}}$. Since g is not P-admissible on $\hat{\mathcal{R}}$, there are $R_1, R_2 \in \hat{\mathcal{R}}$ where $g(R_1) = g(R_2)$ but $R_1 \not\equiv_P R_2$. Since $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$ there is an $R_3 \in \hat{\mathcal{R}}$ such that $f(R_3) = g(R_1) = g(R_2)$. But then either $R_3 \not\equiv_P R_1$ or $R_3 \not\equiv_P R_2$, which is a contradiction, since f is P-robust to misspecification with g on $\hat{\mathcal{R}}$. □

Lemma E.3. *If f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ and $\mathrm{Im}(f|_{\hat{\mathcal{R}}}) = \mathrm{Im}(g|_{\hat{\mathcal{R}}})$ then g is P-robust to misspecification with f on $\hat{\mathcal{R}}$.*

Proof. If f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ then this immediately implies that $f_{\hat{\mathcal{R}}} \neq g|_{\hat{\mathcal{R}}}$, and that if $f(R_1) = g(R_2)$ for some $R_1, R_2 \in \hat{\mathcal{R}}$ then $R_1 \equiv_P R_2$. Lemma E.2 implies that g is P-admissible on $\hat{\mathcal{R}}$, and if $\mathrm{Im}(f|_{\hat{\mathcal{R}}}) = \mathrm{Im}(g|_{\hat{\mathcal{R}}})$ then $\mathrm{Im}(f|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(g|_{\hat{\mathcal{R}}})$. This means that g is P-robust to misspecification with f on $\hat{\mathcal{R}}$. □

Lemma E.4. *f is P-admissible on $\hat{\mathcal{R}}$ but not P-robust to any misspecification on $\hat{\mathcal{R}}$ if and only if $\mathrm{Am}(f) = P$ on $\hat{\mathcal{R}}$.*

Proof. First suppose $\mathrm{Am}(f) = P$ on $\hat{\mathcal{R}}$. This immediately implies that f is P-admissible on $\hat{\mathcal{R}}$. Next, assume that f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, let R_1 be any element of $\hat{\mathcal{R}}$, and consider $g(R_1)$. Since $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$, there is an $R_2 \in \hat{\mathcal{R}}$ such that $f(R_2) = g(R_1)$. Since f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, this implies that $R_2 \equiv_P R_1$. Moreover, if $\mathrm{Am}(f) = P$ then $R_2 \equiv_P R_1$ if and only if $f(R_2) = f(R_1)$, so it must be the case that $f(R_2) = f(R_1)$. Now, since $f(R_2) = f(R_1)$ and $f(R_2) = g(R_1)$, we have that $g(R_1) = f(R_1)$. Since R_1 was chosen arbitrarily, this implies that $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, which is a contradiction. Hence, if $\mathrm{Am}(f) = P$ on $\hat{\mathcal{R}}$ then f is P-admissible on $\hat{\mathcal{R}}$ but not P-robust to any misspecification on $\hat{\mathcal{R}}$.

For the other direction, suppose that f is P-admissible on $\hat{\mathcal{R}}$ and that $\mathrm{Am}(f) \neq P$ on $\hat{\mathcal{R}}$. If $\mathrm{Am}(f) \neq P$ on $\hat{\mathcal{R}}$ then there are $R_1, R_2 \in \hat{\mathcal{R}}$ such that $R_1 \equiv_P R_2$ but $f(R_1) \neq f(R_2)$. We can then construct a g as follows; let $g(R_1) = f(R_2)$, $g(R_2) = f(R_1)$, and $g(R) = f(R)$ for all

$R \neq R_1, R_2$. Now f is P-robust to misspecification with g on $\hat{\mathcal{R}}$. Hence, if f is P-admissible on $\hat{\mathcal{R}}$ but not P-robust to any misspecification on $\hat{\mathcal{R}}$ then $\mathrm{Am}(f) = P$ on $\hat{\mathcal{R}}$. □

Lemma E.5. *If f is not P-robust to misspecification with g on $\hat{\mathcal{R}}$, and $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$, then for any h, $h \circ f$ is not P-robust to misspecification with $h \circ g$ on $\hat{\mathcal{R}}$.*

Proof. If f is not P-robust to misspecification with g on $\hat{\mathcal{R}}$, and $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$, then either f is not P-admissible on $\hat{\mathcal{R}}$, or $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, or $f(R_1) = g(R_2)$ but $R_1 \not\equiv_P R_2$ for some $R_1, R_2 \in \hat{\mathcal{R}}$.

In the first case, if f is not P-admissible on $\hat{\mathcal{R}}$ then $h \circ f$ is not P-admissible on $\hat{\mathcal{R}}$, as per Lemma E.1. This implies that $h \circ f$ is not P-robust to any misspecification (including with $h \circ g$) on $\hat{\mathcal{R}}$.

In the second case, if $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$ then $h \circ f|_{\hat{\mathcal{R}}} = h \circ g|_{\hat{\mathcal{R}}}$. This implies that $h \circ f$ is not P-robust to misspecification with $h \circ g$ on $\hat{\mathcal{R}}$.

In the last case, suppose $f(R_1) = g(R_2)$ but $R_1 \not\equiv_P R_2$ for some $R_1, R_2 \in \hat{\mathcal{R}}$. If $f(R_1) = g(R_2)$ then $h \circ f(R_1) = h \circ g(R_2)$, so there are $R_1, R_2 \in \hat{\mathcal{R}}$ such that $h \circ f(R_1) = h \circ g(R_2)$, but $R_1 \not\equiv_P R_2$. This implies that $h \circ f$ is not P-robust to misspecification with $h \circ g$ on $\hat{\mathcal{R}}$. □

Lemma E.6. *Let f be P-admissible on $\hat{\mathcal{R}}$, and let T be the set of all reward transformations that preserve P on $\hat{\mathcal{R}}$. Then f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ if and only if $g = f \circ t$ for some $t \in T$ such that $f \circ t|_{\hat{\mathcal{R}}} \neq f|_{\hat{\mathcal{R}}}$.*

Proof. First suppose that f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ — we will construct a t that fits our description. For each $y \in \mathrm{Im}(g|_{\hat{\mathcal{R}}})$, let $R_y \in \hat{\mathcal{R}}$ be some reward function such that $f(R_y) = y$; since $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$, such an $R_y \in \hat{\mathcal{R}}$ always exists. Now let t be the function that maps each $R \in \hat{\mathcal{R}}$ to $R_{g(R)}$. Since by construction $g(R) = f(R_{g(R)})$, and since f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, we have that $R \equiv_P R_{g(R)}$. This in turn means that $t \in T$, since t preserves P on $\hat{\mathcal{R}}$. Finally, note that $g = f \circ t$, which means that we are done.

For the other direction, suppose $g = f \circ t$ for some $t \in T$ where $f \circ t|_{\hat{\mathcal{R}}} \neq f|_{\hat{\mathcal{R}}}$. By assumption we have that f is P-admissible on $\hat{\mathcal{R}}$, and that $g|_{\hat{\mathcal{R}}} \neq f|_{\hat{\mathcal{R}}}$. Moreover, we clearly have that $\mathrm{Im}(g|_{\hat{\mathcal{R}}}) \subseteq \mathrm{Im}(f|_{\hat{\mathcal{R}}})$. Finally, if $g(R_1) = f(R_2)$ then $f \circ t(R_1) = f(R_2)$, which means that $R_1 \equiv_P R_3$ for some $R_3 \in \hat{\mathcal{R}}$ such that $f(R_3) = f(R_2)$. Since f is P-admissible on $\hat{\mathcal{R}}$ it follows that $R_3 \equiv_P R_2$, which then implies that $R_1 \equiv_P R_2$. Thus f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, so we are done. □

E.2 Reward Function Equivalence Classes

In this section we will prove Theorem 3.1, which turns out to be quite involved. We start by proving several lemmas, which we will need for the main proof.

E.2.1 Lemmas Concerning Reward Transformations

Here, we provide some lemmas concerning reward transformations.

Lemma E.7. *If $\gamma_1 \neq \gamma_2$ then* $\mathrm{PS}_{\gamma_1} \cap \mathrm{PS}_{\gamma_2} = \mathrm{CS}$.

Proof. First, it is straightforward that $\mathrm{CS} \subseteq \mathrm{PS}_\gamma$ for all γ. Suppose $t \in \mathrm{CS}$, and let $\Phi_R(s) = c_R/(1-\gamma)$ for all s, where c_R is the constant that t shifts R by. Then $t(R)$ is given by potential shaping R with Φ_R, which means that $t \in \mathrm{PS}_\gamma$. Hence $\mathrm{CS} \subseteq \mathrm{PS}_{\gamma_1} \cap \mathrm{PS}_{\gamma_2}$.

For the other direction, suppose $t \in \mathrm{PS}_{\gamma_1} \cap \mathrm{PS}_{\gamma_2}$, and let $s_1, s_2 \in \mathcal{S}$. We have that:

$$\begin{aligned}
\Phi_1(s_1) - \gamma_1\Phi_1(s_2) &= \Phi_2(s_1) - \gamma_2\Phi_2(s_2)\\
\Phi_1(s_2) - \gamma_1\Phi_1(s_1) &= \Phi_2(s_2) - \gamma_2\Phi_2(s_1)\\
\Phi_1(s_1) - \gamma_1\Phi_1(s_1) &= \Phi_2(s_1) - \gamma_2\Phi_2(s_1)\\
\Phi_1(s_2) - \gamma_1\Phi_1(s_2) &= \Phi_2(s_2) - \gamma_2\Phi_2(s_2)
\end{aligned}$$

By substituting, we can show that

$$\begin{aligned}
&\Phi_1(s_1) - \gamma_1\Phi_1(s_2) = \Phi_2(s_1) - \gamma_2\Phi_2(s_2) \\
&= \left(\frac{1-\gamma_1}{1-\gamma_2}\right)\Phi_1(s_1) - \gamma_1\left(\frac{1-\gamma_1}{1-\gamma_2}\right)\Phi_1(s_2) \\
&\implies \left(1 - \frac{1-\gamma_1}{1-\gamma_2}\right)\Phi(s_1) = \left(\gamma_1 - \gamma_2\frac{1-\gamma_1}{1-\gamma_2}\right)\Phi_1(s_2) \\
&\implies \Phi_1(s_1) = \Phi_1(s_2).
\end{aligned}$$

By induction, we get that Φ_1 has a constant value for all s (and by symmetry, that this is true of Φ_2 as well). Hence $t \in \mathrm{CS}$, and so we have proven that $\mathrm{PS}_{\gamma_1} \cap \mathrm{PS}_{\gamma_2} = \mathrm{CS}$. □

Lemma E.8. $(F \circ H) \cap (G \circ H) = (F \cap G) \circ H$

Proof. If $(F \circ H) \cap (G \circ H)$ then t can be expressed as a finite sequence $t_1 \circ \dots \circ t_n$, where each t_i is in F or H, and each t_i is in G or H. Equivalently, each t_i is in both F and G, or in H, which implies that $t \in (F \cap G) \circ H$, and so $(F \circ H) \cap (G \circ H) \subseteq (F \cap G) \circ H$. An analogous argument shows that $(F \cap G) \circ H \subseteq (F \circ H) \cap (G \circ H)$, and so $(F \circ H) \cap (G \circ H) = (F \cap G) \circ H$. □

E.2.2 Lemmas Concerning State-Action Visit Counts

Here we provide some lemmas about the topological structure of MDPs. Recall that we assume that all states in S are reachable.

Let Π be the set of all policies. Moreover, given τ and μ_0, let $m_{\tau,\mu_0} : \Pi \to \mathbb{R}^{|S||A|}$ be a map that sends each policy π to a vector d_π, such that

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

In other words, let $m_{\tau,\mu_0}(\pi)$ be a vector that records the expected discounted "density" of π's trajectories in each state-action pair (under τ and μ_0).

Given a reward function R and a transition function τ, let $\vec{R}_\tau \in \mathbb{R}^{|S||A|}$ be the vector where $\vec{R}_\tau[s,a] = \mathbb{E}_{S'\sim\tau(s,a)}[R(s,a,S')]$. Moreover, note that $\mathcal{J}(\pi) = m_{\tau,\mu_0}(\pi) \cdot \vec{R}_\tau$.

Let $\bar{\Pi} \subset \Pi$ be the set of all policies that visit each state with positive probability.

Lemma E.9. *m_{τ,μ_0} is injective on $\bar{\Pi}$.*

Proof. Suppose $m_{\tau,\mu_0}(\pi) = m_{\tau,\mu_0}(\pi')$ for some $\pi, \pi' \in \bar{\Pi}$. Next, given τ, μ_0, define w_π as

$$w_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\xi\sim\pi}(S_t = s).$$

Note that if $m_{\tau,\mu_0}(\pi) = m_{\tau,\mu_0}(\pi')$ then $w_\pi = w_{\pi'}$, and moreover that

$$m_{\tau,\mu_0}(\pi)[s,a] = w_\pi(s)\pi(a \mid s).$$

This means that if $w_\pi(s) \neq 0$ for all s, which is the case for all $\pi \in \bar{\Pi}$, then we can express π as

$$\pi(a \mid s) = \frac{m_{\tau,\mu_0}(\pi)[s,a]}{w_\pi(s)}.$$

This means that if $m_{\tau,\mu_0}(\pi) = m_{\tau,\mu_0}(\pi')$ for some $\pi, \pi' \in \bar{\Pi}$ then $\pi = \pi'$. □

Note that m_{τ,μ_0} is *not* injective on Π; if there is some state s that π reaches with probability 0, then we can alter the behaviour of π at s without changing $m_{\tau,\mu_0}(\pi)$.

Lemma E.10. $\mathrm{Im}(m_{\tau,\mu_0})$ *is located in an affine space with $|S|(|A|-1)$ dimensions.*

Proof. To show that $\mathrm{Im}(m_{\tau,\mu_0})$ is located in an affine space with $|S|(|A|-1)$ dimensions, first note there is no π such that $m(\pi)$ is the zero vector. This means that the smallest affine space which contains $\mathrm{Im}(m_{\tau,\mu_0})$ does not contain the origin.

Next, recall if R_2 is produced by shaping R_1 with Φ, and $\mathbb{E}_{S_0\sim\mu_0}[\Phi(S_0)] = 0$, then $\mathcal{J}_1(\pi) = \mathcal{J}_2(\pi)$ for all π. This means that knowing the value of $\mathcal{J}(\pi)$ for all π determines $\vec{R}$ modulo at least $|S|-1$ free variables, which means that $\mathrm{Im}(m_{\tau,\mu_0})$ contains at most $|S|(|A|-1)+1$ linearly independent vectors. Since the smallest affine space that contains $\mathrm{Im}(m_{\tau,\mu_0})$ does not contain the origin, this means that $\mathrm{Im}(m_{\tau,\mu_0})$ is located in an affine space with $|S|(|A|-1)$ dimensions. □

For the next lemma, let $\tilde{\Pi} \subset \Pi$ be the set of all policies that take all actions with positive probability in each state, and note that $\tilde{\Pi} \subset \bar{\Pi}$ (i.e., a policy that takes every action with positive probability in each state visits every state with positive probability).

Lemma E.11. *$m_{\tau,\mu_0}(\tilde{\Pi})$ is open in $\mathbb{R}^{|S|(|A|-1)}$, and m_{τ,μ_0} is a homeomorphism between $\tilde{\Pi}$ and $m_{\tau,\mu_0}(\tilde{\Pi})$.*

Proof. By the Invariance of Domain Theorem (Brouwer, 1912), if

1. U is an open subset of $\mathbb{R}^n$, and
2. $f : U \to \mathbb{R}^n$ is an injective continuous map,

then $f(U)$ is open in $\mathbb{R}^n$ and f is a homeomorphism between U and $f(U)$. We will show that m and $\tilde{\Pi}$ satisfy the requirements of this theorem.

We begin by noting that Π can be represented as a set of points in $\mathbb{R}^{|S|(|A|-1)}$. We do this by considering each policy π as a vector $\vec{\pi}$ of length $|S||A|$, where $\vec{\pi}[s,a] = \pi(a \mid s)$. Moreover, since $\sum_{a\in A} \pi(a \mid s) = 1$ for all s, we can remove one dimension, and embed Π in $\mathbb{R}^{|S|(|A|-1)}$.

$\tilde{\Pi}$ is an open set in $\mathbb{R}^{|S|(|A|-1)}$. By Lemma E.10, we have that m_{τ,μ_0} is a mapping $m_{\tau,\mu_0} : \tilde{\Pi} \to \mathbb{R}^{|S|(|A|-1)}$. By Lemma E.9, we have that m_{τ,μ_0} is injective on $\tilde{\Pi}$. Finally, m_{τ,μ_0} is continuous. We can therefore apply the Invariance of Domain Theorem, and conclude that $m_{\tau,\mu_0}(\tilde{\Pi})$ is open in $\mathbb{R}^{|S|(|A|-1)}$, and that m_{τ,μ_0} is a homeomorphism between $\tilde{\Pi}$ and $m_{\tau,\mu_0}(\tilde{\Pi})$. □

Note that lemma E.11 holds for all τ and μ_0.

E.2.3 Results Concerning the Policy Order

In this section, we prove our results concerning the policy orderings. First, we need to define a new set of transformations. Let $\mathrm{PS}^k_{\gamma,\mu_0}$ be the set of all potential shaping transformations t that, for each R, apply a potential function Φ such that $\mathbb{E}_{S_0\sim\mu_0}[\Phi(S_0)] = k$.

Lemma E.12. *$\mathcal{J}_1 = \mathcal{J}_2$ if and only if $R_1 = t(R_2)$ for some $t \in \mathrm{PS}^0_{\gamma,\mu_0} \circ S'\mathrm{R}_\tau$.*

Proof. First suppose $R_1 = t(R_2)$ for some $t \in \mathrm{PS}^0_{\gamma,\mu_0} \circ S'\mathrm{R}_\tau$. Then $V_1^\pi(s) = V_2^\pi(s) - \Phi(s)$, where Φ is the potential shaping function applied by t (see e.g. Lemma B1 in Skalse et al. 2022). Hence $\mathcal{J}_1(\pi) = \mathcal{J}_2(\pi) - \mathbb{E}_{s_0\sim\mu_0}[\Phi(s_0)] = \mathcal{J}_2(\pi)$, and so we have proven the first direction.

For the other direction, recall that since $\mathcal{J}(\pi) = m_{\tau,\mu_0}(\pi) \cdot \vec{R}_\tau$, we have that $\mathcal{J}$ determines the value of $m_{\tau,\mu_0}(\pi) \cdot \vec{R}_\tau$ for each $m_{\tau,\mu_0}(\pi) \in \mathrm{Im}(m_{\tau,\mu_0})$. This means that if we can pick n linearly independent vectors from $\mathrm{Im}(m_{\tau,\mu_0})$, then that determines that $\vec{R}_\tau$ is located in some affine space with $(|\mathcal{S}||\mathcal{A}| - n)$ dimensions. Lemma E.11 says that $\mathrm{Im}(m_{\tau,\mu_0})$ is open in $\mathbb{R}^{|S|(|A|-1)}$, which means that $\mathrm{Im}(m_{\tau,\mu_0})$ contains $|S|-1$ linearly independent vectors. We know that $\mathcal{J}$ is preserved by transformations in $\mathrm{PS}^0_{\gamma,\mu_0}$, and each such transformation is specified by $|S|-1$ variables (corresponding to the value of Φ in each state, which is determined for one of the initial states). This means that $\mathrm{Im}(m_{\tau,\mu_0}) \cdot \vec{R}_\tau$ determines $\vec{R}_\tau$ modulo exactly $|\mathcal{S}|-1$ degrees of freedom, which we can identify with the values of Φ for the transformations in $\mathrm{PS}^0_{\gamma,\mu_0}$. Hence $\mathrm{Im}(m_{\tau,\mu_0}) \cdot \vec{R}_\tau$ (and therefore

$\mathcal{J}$) are preserved by transformations in $\mathrm{PS}^0_{\gamma,\mu_0}$ and transformations that preserve $\vec{R}_\tau$, and no other transformations. $\vec{R}_\tau$ is of course preserved by S'-redistribution, and no other transformations. We have hence proven the other direction. □

We can now finally prove (a somewhat generalised version of) Theorem 3.1.

Theorem E.13. *The invariance of* $\mathrm{ORD}^{\mathcal{M}}$ *is exactly characterised by the following transformations:*

1. $R \equiv_{\mathrm{ORD}^{\mathcal{M}}} t(R)$ *for all* R *if and only if* $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$.
2. $R \equiv_{\mathrm{ORD}^{\mathcal{M}}} t(R)$ *for all* R *and* τ *if and only if* $t \in \mathrm{PS}_\gamma \circ \mathrm{LS}$.
3. $R \equiv_{\mathrm{ORD}^{\mathcal{M}}} t(R)$ *for all* R *and* γ *if and only if* $t \in S'\mathrm{R}_\tau \circ \mathrm{CS} \circ \mathrm{LS}$.
4. $R \equiv_{\mathrm{ORD}^{\mathcal{M}}} t(R)$ *for all* R, τ, *and* γ *if and only if* $t \in \mathrm{CS} \circ \mathrm{LS}$.

Moreover, if $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$ *then* $R \equiv_{\mathrm{ORD}^{\mathcal{M}}} t(R)$ *for all* R *and* μ_0.

Proof. We begin by proving Claim 1. First, $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ if and only if $\mathcal{J}_1$ is a monotonic transformation of $\mathcal{J}_2$. Next, since $\mathcal{J}(\pi) = m_{\tau,\mu_0}(\pi) \cdot \vec{R}_\tau$ we have that $\mathcal{J}$ has a "hidden linear structure" which implies that affine transformations are the only possible monotonic transformations of $\mathcal{J}$. Hence $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ if and only if $\mathcal{J}_1 = a \cdot \mathcal{J}_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$.

We next show that $\mathcal{J}_1 = a \cdot \mathcal{J}_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$ if and only if $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$. The first direction is straightforward. First, if $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau$ then $\mathcal{J}_1 = \mathcal{J}_2$. Next, if $R_1 = t(R_2)$ for some $t \in \mathrm{PS}_\gamma$ then $\mathcal{J}_1 = \mathcal{J}_2 - \mathbb{E}_{S_0 \sim \mu_0}[\Phi_t(S_0)]$ (see e.g. Lemma B1 in Skalse et al. 2022). Finally, if $R_1 = t(R_2)$ for some $t \in \mathrm{LS}$ then $\mathcal{J}_1 = c \cdot \mathcal{J}_2$ for some $c \in \mathbb{R}^+$. Hence if $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$ then $\mathcal{J}_1 = a \cdot \mathcal{J}_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$, which means that $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$.

For the other direction, suppose $\mathcal{J}_1 = a \cdot \mathcal{J}_2 + b$ for some $a \in \mathbb{R}^+, b \in \mathbb{R}$. Consider the reward function R_3 given by first scaling R_2 by a, and then shape the resulting reward with the potential function Φ that is equal to $-b$ for all initial states, and equal to 0 elsewhere. Now $\mathcal{J}_3 = \mathcal{J}_1$, so (by Lemma E.12) there is a $t' \in \mathrm{PS}^0_{\gamma,\mu_0} \circ S'\mathrm{R}_\tau$ such that $R_1 = t'(R_3)$. By composing t' with the transformation that produced R_3 from R_2, we obtain a $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$ such that $R_1 = t(R_2)$. Hence if $R_1 \equiv_{\mathrm{ORD}^{\mathcal{M}}} R_2$ then $R_1 = t(R_2)$ for some $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$. We have thus proven both directions, and hence Claim 1.

Claim 2-4 follows from Claim 1 and Lemma E.7 and E.8. The remark at the end is straightforward. □

E.3 Misspecified Behavioural Models

In this section, we prove our results from Section 4.

Theorem E.14. *Let* $f^{\mathcal{M}} \in F^{\mathcal{M}}$ *be surjective onto* Π^+. *Then* f^M *is* $\mathrm{OPT}^{\mathcal{M}}$*-robust to misspecification with* g *if and only if* $g \in F^{\mathcal{M}}$ *and* $g \neq f^{\mathcal{M}}$.

Proof. $f^{\mathcal{M}}$ is $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with g in $\mathcal{M}$ if and only if $f^{\mathcal{M}}$ is $\mathrm{OPT}^{\mathcal{M}}$-admissible, $g \neq f^{\mathcal{M}}$, $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, and for all $\pi \in \mathrm{Im}(g)$ we have that all $R \in \mathrm{Am}_g(\pi)$ and all $R \in \mathrm{Am}_f(\pi)$ have the same optimal policies in $\mathcal{M}$.

For all $f \in F^{\mathcal{M}}$ and all R, $\mathrm{argmax}_{a \in \mathcal{A}} f(R)(a \mid s) = \mathrm{argmax}_{a \in \mathcal{A}} Q^\star(s,a)$. Since $f^{\mathcal{M}} \in F^{\mathcal{M}}$, this means that if $f^{\mathcal{M}}(R_1) = f^{\mathcal{M}}(R_2)$ then $\mathrm{argmax}_{a \in \mathcal{A}} Q_1^\star(s,a) = \mathrm{argmax}_{a \in \mathcal{A}} Q_2^\star(s,a)$ in $\mathcal{M}$. Moreover, R_1 and R_2 have the same optimal policies in $\mathcal{M}$ if and only if $\mathrm{argmax}_{a \in \mathcal{A}} Q_1^\star(s,a) = \mathrm{argmax}_{a \in \mathcal{A}} Q_2^\star(s,a)$ in $\mathcal{M}$. Therefore, if $f^{\mathcal{M}}(R_1) = f^{\mathcal{M}}(R_2)$ then $R_1 \equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$, and so $f^{\mathcal{M}}$ is $\mathrm{OPT}^{\mathcal{M}}$-admissible.

Let $g \in F^{\mathcal{M}}$ and $g \neq f^{\mathcal{M}}$. Since g is a function $\mathcal{R} \to \Pi^+$, and since $f^{\mathcal{M}}$ is surjective onto Π^+, we have that $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$. Next, by the same argument as above, if $f^{\mathcal{M}}(R_1) = g(R_2)$ then $\mathrm{argmax}_{a \in \mathcal{A}} Q_1^\star(s,a) = \mathrm{argmax}_{a \in \mathcal{A}} Q_2^\star(s,a)$, which implies that $R_1 \equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$. This means that f^M is $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with g.

Next, suppose f^M is $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with g. This means that $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ and that if $f^{\mathcal{M}}(R_1) = g(R_2)$ then $\mathrm{argmax}_{a\in\mathcal{A}} Q_1^\star(s,a) = \mathrm{argmax}_{a\in\mathcal{A}} Q_2^\star(s,a)$. Since $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ implies that g is a function $\mathcal{R} \to \Pi^+$, and since $f^{\mathcal{M}}(R_1) = g(R_2)$ implies that $\mathrm{argmax}_{a\in\mathcal{A}} f^{\mathcal{M}}(R)(a \mid s) = \mathrm{argmax}_{a\in\mathcal{A}} g(R)(a \mid s)$, this implies that $g \in F^{\mathcal{M}}$. □

Theorem E.15. *Let $b_\psi^{\mathcal{M}} \in B^{\mathcal{M}}$. Then $b_\psi^{\mathcal{M}}$ is $\mathrm{ORD}^{\mathcal{M}}$-robust to misspecification with g if and only if $g \in B^{\mathcal{M}}$ and $g \neq b_\psi^{\mathcal{M}}$.*

Proof. As per Theorem 3.3 in Skalse et al. 2022, $\mathrm{Am}(b_\psi^{\mathcal{M}})$ is characterised by $\mathrm{PS}_\gamma \circ S'\mathrm{R}_\tau$, and as per Theorem 3.1, $\mathrm{ORD}_{\mathcal{M}}$ is characterised by $\mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$. Hence $b_\psi^{\mathcal{M}}$ is $\mathrm{ORD}_{\mathcal{M}}$-admissible, which means that Lemma 2.1 implies that $b_\psi^{\mathcal{M}}$ is $\mathrm{ORD}_{\mathcal{M}}$-robust to misspecification with g if and only if $g \neq b_\psi^{\mathcal{M}}$, and there exists a $t \in \mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$ such that $g = b_\psi^{\mathcal{M}} \circ t$. Recall that $b_\psi^{\mathcal{M}}(R)$ is given by

$$b_\psi^{\mathcal{M}}(R)(a \mid s) = \frac{\exp \psi(R) A_R(s,a)}{\sum_{a\in\mathcal{A}} \exp \psi(R) A_R(s,a)}.$$

where A_R is the optimal advantage function of R in $\mathcal{M}$. If $g(R) = b_\psi^{\mathcal{M}} \circ t(R)$ for some $t \in \mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$, then we have that

$$\begin{aligned} g(R)(a \mid s) &= \frac{\exp \psi(t(R)) A_{t(R)}(s,a)}{\sum_{a\in\mathcal{A}} \exp \psi(t(R)) A_{t(R)}(s,a)} \\ &= \frac{\exp \psi(t(R)) c_R A_R(s,a)}{\sum_{a\in\mathcal{A}} \exp \psi(t(R)) c_R A_R(s,a)}, \end{aligned}$$

where c_R is the linear scaling factor that t applies to R. Note that the advantage function A is preserved by both potential shaping and S'-redistribution. Now let $\psi'(R) = \psi(t(R)) \cdot c_R$, and we can see that $g = b_{\psi'}^{\mathcal{M}} \in B^{\mathcal{M}}$. We have hence shown that $b_\psi^{\mathcal{M}}$ is strongly robust to misspecification with g in $\mathcal{M}$ if and only if $g \in B^{\mathcal{M}}$ and $g \neq b_\psi^{\mathcal{M}}$. □

Theorem E.16. *No function in $\mathcal{O}^{\mathcal{M}}$ is $\mathrm{ORD}^{\mathcal{M}}$-admissible. The only function in $\mathcal{O}^{\mathcal{M}}$ that is $\mathrm{OPT}^{\mathcal{M}}$-admissible is $o_m^{\mathcal{M}}$, and this function is not $\mathrm{OPT}^{\mathcal{M}}$-robust to any misspecification.*

Proof. This Theorem largely follows from Lemma E.4. First, if $o^{\mathcal{M}} \in \mathcal{O}^{\mathcal{M}}$ then $o^{\mathcal{M}}$ has a finite codomain, whereas there is an uncountable number of $\mathrm{ORD}^{\mathcal{M}}$-equivalence classes. This means that $o^{\mathcal{M}}$ cannot be $\mathrm{ORD}^{\mathcal{M}}$-admissible. Moreover, $\mathrm{Am}(o_m^{\mathcal{M}}) = \mathrm{OPT}^{\mathcal{M}}$. Therefore, by Lemma E.4, $o_m^{\mathcal{M}}$ is $\mathrm{OPT}^{\mathcal{M}}$-admissible, but not $\mathrm{OPT}^{\mathcal{M}}$-robust to any misspecification. Finally, if $o^{\mathcal{M}} \in \mathcal{O}^{\mathcal{M}}$ but $o^{\mathcal{M}} \neq o_m^{\mathcal{M}}$, then there is a pigeonhole argument to show that there must be at least two R_1, R_2 such that $o^{\mathcal{M}}(R_1) = o^{\mathcal{M}}(R_2)$ but $R_1 \not\equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$. This means that $o^{\mathcal{M}}$ is not $\mathrm{OPT}^{\mathcal{M}}$-admissible.

The pigeonhole argument goes like this: the codomain of each $o^{\mathcal{M}} \in O^{\mathcal{M}}$ has $(2^{|\mathcal{A}|} - 1)^{|\mathcal{S}|}$ elements, and there are $(2^{|\mathcal{A}|} - 1)^{|\mathcal{S}|}$ $\mathrm{OPT}^{\mathcal{M}}$-equivalence classes. This means that if $o^{\mathcal{M}}$ is $\mathrm{OPT}^{\mathcal{M}}$-admissible, then there must be a one-to-one correspondence between $\mathrm{OPT}^{\mathcal{M}}$-equivalence classes and elements of $o^{\mathcal{M}}$'s codomain, so that there for each equivalence class C is a y_C such that $o^{\mathcal{M}}(R) = y_C$ if and only if $R \in C$. Further, say that if $f, g : X \to \mathcal{P}(Y)$ are set-valued functions, then $f \subseteq g$ if $f(x) \subseteq g(x)$ for all $x \in X$, and $f \subset g$ if $f \subseteq g$ but $g \not\subseteq f$. Then if $o^{\mathcal{M}} \in \mathcal{O}^{\mathcal{M}}$ we have that $o^{\mathcal{M}}(R) \subseteq o_m^{\mathcal{M}}(R)$ for all R — a policy is optimal if and only if it takes only optimal actions, but it need not take all optimal actions. Moreover, if $o^{\mathcal{M}} \neq o_m^{\mathcal{M}}$ then there is an R_1 such that $o^{\mathcal{M}}(R_1) \subset o_m^{\mathcal{M}}(R_1)$. Let R_2 be a reward function so that $o_m^{\mathcal{M}}(R_2) = o^{\mathcal{M}}(R_1)$ — for any function $\mathcal{S} \to \mathcal{P}(\mathcal{A}) - \varnothing$, there is a reward function for which those are the optimal actions, so there is always some R_2 such that $o_m^{\mathcal{M}}(R_2) = o^{\mathcal{M}}(R_1)$. Now either $o^{\mathcal{M}}(R_2) = o^{\mathcal{M}}(R_1)$ or $o^{\mathcal{M}}(R_2) \subset o^{\mathcal{M}}(R_1)$, since all actions that are optimal under R_2 are optimal under R_1. In the first case, since $o^{\mathcal{M}}(R_1) = o^{\mathcal{M}}(R_2)$ but $R_1 \not\equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$, we have that $o^{\mathcal{M}}$ is not $\mathrm{OPT}^{\mathcal{M}}$-admissible. In the second case, let R_3 be a reward function so that $o_m^{\mathcal{M}}(R_3) = o^{\mathcal{M}}(R_2)$, and repeat the same argument. Since there can only be a finite sequence $o^{\mathcal{M}}(R_n) \subset \cdots \subset o^{\mathcal{M}}(R_2) \subset o^{\mathcal{M}}(R_1)$, we have that we must eventually find two R_n, R_{n-1} such that $o^{\mathcal{M}}(R_n) = o^{\mathcal{M}}(R_{n-1})$ but $R_n \not\equiv_{\mathrm{OPT}^{\mathcal{M}}} R_{n-1}$. This means that $o^{\mathcal{M}}$ cannot be $\mathrm{OPT}^{\mathcal{M}}$-admissible. □

Theorem E.17. *Let $c^{\mathcal{M}}_\psi \in C^{\mathcal{M}}$. Then $c^{\mathcal{M}}_\psi$ is $\mathrm{ORD}^{\mathcal{M}}$-robust to misspecification with g if and only if $g \in C^{\mathcal{M}}$ and $g \neq c^{\mathcal{M}}_\psi$.*

Proof. As per Theorem 3.4 in Skalse et al. 2022, $\mathrm{Am}(c^{\mathcal{M}}_\psi)$ is characterised by $\mathrm{PS}_\gamma \circ S'\mathrm{R}_\tau$, and as per Theorem 3.1, $\mathrm{ORD}_{\mathcal{M}}$ is characterised by $\mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$. Hence $c^{\mathcal{M}}_\psi$ is $\mathrm{ORD}_{\mathcal{M}}$-admissible, which means that Lemma 2.1 implies that $c^{\mathcal{M}}_\psi$ is $\mathrm{ORD}_{\mathcal{M}}$-robust to misspecification with g if and only if $g \neq c^{\mathcal{M}}_\psi$, and there exists a $t \in \mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$ such that $g = c^{\mathcal{M}}_\psi \circ t$. Moreover, $c^{\mathcal{M}}_\psi(R)$ is the unique policy that maximises the maximal causal entropy objective;

$$\mathcal{J}^{\mathrm{MCE}}_{\psi(R)}(\pi) = \mathcal{J}_R(\pi) - \psi(R)\mathbb{E}_{S_t \sim \pi, \tau, \mu_0}[\gamma^t \mathcal{H}(\pi(S_t))].$$

Therefore, if $g(R) = c^{\mathcal{M}}_\psi \circ t(R)$ then $g(R)$ is the policy that maximises the objective

$$\begin{aligned}
&\mathcal{J}^{\mathrm{MCE}}_{\psi(t(R))}(\pi)\\
=&\mathcal{J}_{t(R)}(\pi) - \psi(t(R))\mathbb{E}_{S_t \sim \pi, \tau, \mu_0}[\gamma^t \mathcal{H}(\pi(S_t))]\\
=&c_R \cdot \mathcal{J}_R(\pi) - \psi(t(R))\mathbb{E}_{S_t \sim \pi, \tau, \mu_0}[\gamma^t \mathcal{H}(\pi(S_t))]
\end{aligned}$$

where c_R is the linear scaling factor that t applies to R. Note that $\mathcal{J}_R$ is preserved by S'-redistribution, and potential shaping can only change $\mathcal{J}_R$ by inducing a uniform constant shift of $\mathcal{J}_R$ for all policies. This means that linear scaling is the only transformation in $\mathrm{PS}_\gamma \circ \mathrm{LS} \circ S'\mathrm{R}_\tau$ that could affect the maximal causal entropy objective. Finally, let ψ' be the function $\psi'(R) = \psi(t(R)) \cdot c_R$, and we can see that $g = c^{\mathcal{M}}_{\psi'} \in C^{\mathcal{M}}$. We have hence shown that $c^{\mathcal{M}}_\psi$ is $\mathrm{ORD}^{\mathcal{M}}$-robust to misspecification with g in $\mathcal{M}$ if and only if $g \in C^{\mathcal{M}}$ and $g \neq c^{\mathcal{M}}_\psi$. □

E.4 Misspecified MDPs

We here prove our results from Appendix D.1. The first of these proofs is straightforward.

Lemma E.18. *If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ then f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible for $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau_2, \mu_0, \_, \gamma \rangle$ unless $\tau_1 = \tau_2$.*

Proof. This follows directly from Theorem 4.2 in Skalse et al. 2022. □

To prove the next result, we first need a supporting lemma. We say that a state s is *controllable* relative to a transition function τ, initial state distribution μ_0, and discount γ, if there exist two policies π, π' such that

$$\sum_{t=1}^{\infty} \gamma^t \mathbb{P}_{\xi \sim \pi}(s_t = s) \neq \sum_{t=1}^{\infty} \gamma^t \mathbb{P}_{\xi \sim \pi'}(s_t = s).$$

Note that the sum starts from $t = 1$. It can therefore be viewed as summing the discounted probability that π and π' enter s at each time step. Recall also that τ is trivial if for all $s \in \mathcal{S}$ and $a, a' \in \mathcal{A}$, we have $\tau(s, a) = \tau(s, a')$.

Lemma E.19. *For any μ_0, γ, and τ, there exists a controllable state if and only if τ is non-trivial.*

Proof. It is straightforward to see that if τ is trivial then there are no controllable states.

For the other direction, suppose there are no controllable states. This in turn implies that every policy is optimal under any reward function defined over the domain $\mathcal{S}$. Formally, if R is a reward function such that for each $s \in \mathcal{S}$, we have that $R(s, a_1, s_1) = R(s, a_2, s_2)$ for all $s_1, s_2 \in \mathcal{S}, a_1, a_2 \in \mathcal{A}$, and if there are no controllable states, then every policy is optimal under R. In particular, every deterministic policy is optimal under all such reward functions.

Given a reward function defined over the domain $\mathcal{S}$, let $\vec{R} \in \mathbb{R}^{|S|}$ be the vector such that $\vec{R}[s]$ is the reward that R assigns to transitions leaving s. Moreover, given a deterministic policy π, let T^π be the $|\mathcal{S}| \times |\mathcal{S}|$-dimensional transition matrix that describes the transitions of π under τ. Then if all deterministic policies are optimal under R, we can apply Theorem 3 from Ng and Russell 2000 and conclude that

$$(T^\pi - T^{\pi'})(I - \gamma T^\pi)^{-1}\vec{R} = 0$$

for all deterministic policies π, π'. If this holds for all $\vec{R}$, we then have that $(T^\pi - T^{\pi'})(I - \gamma T^\pi)^{-1}$ is the zero matrix for all deterministic policies π, π'. Moreover, since $(I - \gamma T^\pi)^{-1}$ has no zero eigenvalues, this then means that $(T^\pi - T^{\pi'})$ must be the zero matrix for all pairs of deterministic policies π, π'. This, in turn, implies that τ must be trivial. □

We can now prove the second lemma from Appendix D.1.

Lemma E.20. *If $f^{\gamma_1} = f^{\gamma_1} \circ t$ for all $t \in \mathrm{PS}_{\gamma_1}$ then f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible for $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma_2 \rangle$ unless $\gamma_1 = \gamma_2$ or τ is trivial.*

Proof. As per Lemma E.19, if τ is non-trivial then there is a state s that is controllable relative to τ, μ_0, and γ_2. Let R_1 be any reward function, and let R_2 be the reward that is obtained by potential shaping R_1 with the discount γ_1 and the potential function that is equal to X on s (where $X \neq 0$), and 0 on all other states. Note that there is a $t \in \mathrm{PS}_{\gamma_1}$ such that $R_2 = t(R_1)$, which means that $f^{\gamma_1}(R_1) = f^{\gamma_1}(R_2)$. Next, let $\Delta^\pi = \mathcal{J}_2(\pi) - \mathcal{J}_1(\pi)$, evaluated in $\mathcal{M}$. Moreover, given a policy π, let

$$n_2^\pi = \sum_{t=0}^{\infty} \gamma_2^t \mathbb{P}(\pi \text{ enters } s \text{ at time } t),$$
$$x_2^\pi = \sum_{t=0}^{\infty} \gamma_2^t \mathbb{P}(\pi \text{ exits } s \text{ at time } t).$$

We then have that $\Delta^\pi = X \cdot (\gamma_1 n_2^\pi - x_2^\pi)$. We will use p to denote $\mu_0(s)$. If $\gamma_1 = \gamma_2$ then we know that $\Delta^\pi = -X \cdot p$, which gives that

$$\begin{aligned} X \cdot (\gamma_2 n_2^\pi - x_2^\pi) &= -X \cdot p \\ \gamma_2 n_2^\pi - x_2^\pi &= -p \\ x_2^\pi &= \gamma_2 n_2^\pi + p \end{aligned}$$

By plugging this into the above, and rearranging, we obtain

$$\Delta^\pi = X n_2^\pi (\gamma_1 - \gamma_2) + pX.$$

Moreover, if s is controllable then there are π, π' such that $n_2^\pi \neq n_2^{\pi'}$, which means that $\Delta^\pi \neq \Delta^{\pi'}$. In particular, there are π, π' such that $\Delta^\pi \neq \Delta^{\pi'}$, and π is optimal under R_1, but π' is not. Now, if $\gamma_1 \neq \gamma_2$ then by making X sufficiently large or sufficiently small, we can make it so that π' is optimal under R_2, but π is not. Hence $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma_2 \rangle$ and $\langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma_2 \rangle$ have different optimal policies. This means that there are two reward functions R_1, R_2, such that $f^{\gamma_1}(R_1) = f^{\gamma_1}(R_2)$, but $R_1 \not\equiv_{\mathrm{OPT}^{\mathcal{M}}} R_2$. Therefore, if $\gamma_1 \neq \gamma_2$ and τ is non-trivial then f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible. □

It is worth noting that the above lemma works even if f^{γ_1} is only invariant to γ_1-based potential shaping whose potential is 0 for all initial states, provided that τ gives control over some non-initial state. This can be used to generalise the lemma somewhat, since there are some reward objects which are invariant only to such potential shaping; see Skalse et al. 2022.

Using these two lemmas, we can now prove the theorem:

Theorem E.21. *If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ then f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{τ_2} for any $\mathcal{M}$. Moreover, if $f^{\gamma_1} = f^{\gamma_1} \circ t$ for all $t \in \mathrm{PS}_{\gamma_1}$ then f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{γ_2} for any $\mathcal{M}$ whose transition function τ is non-trivial.*

Proof. Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \tau, \mu_0, \_, \gamma \rangle$. If f is $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with g then f must by definition be $\mathrm{OPT}^{\mathcal{M}}$-admissible. Moreover, Lemma E.2 says that g must be $\mathrm{OPT}^{\mathcal{M}}$-admissible as well. This proof will proceed by showing that in each case, at least one of the relevant reward objects fails to be $\mathrm{OPT}^{\mathcal{M}}$-admissible.

If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ then Lemma D.1 says that f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible unless $\tau_1 = \tau$, and similarly for f^{τ_2}. If $\tau_1 \neq \tau_2$ then either $\tau_1 \neq \tau$ or $\tau_2 \neq \tau$. Hence either f^{τ_1} or f^{τ_2} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible, which means that f^{τ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{τ_2}.

Similarly, if $f^{\gamma_1} = f^{\gamma_1} \circ t$ for all $t \in \mathrm{PS}_{\gamma_1}$ then Lemma D.2 says that f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible unless $\gamma_1 = \gamma$ or τ is trivial, and similarly for f^{γ_2}. If $\gamma_1 \neq \gamma_2$ then either $\gamma_1 \neq \gamma$ or $\gamma_2 \neq \gamma$. Hence either f^{γ_1} or f^{γ_2} is not $\mathrm{OPT}^{\mathcal{M}}$-admissible, unless τ is trivial, which means that f^{γ_1} is not $\mathrm{OPT}^{\mathcal{M}}$-robust to misspecification with f^{γ_2}, unless τ is trivial. □

E.5 Restrictions on the Reward Function

Here, we prove our results from Appendix D.2.

Theorem E.22. *If f is P-robust to misspecification with g on $\hat{\mathcal{R}}$ then f is P-robust to misspecification with g' on $\mathcal{R}$ for some g' where $g'|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$, unless f is not P-admissible on $\mathcal{R}$, and if f is P-robust to misspecification with g on $\mathcal{R}$ then f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, unless $f|_{\hat{\mathcal{R}}} = g|_{\hat{\mathcal{R}}}$.*

Proof. Suppose f is P-robust to misspecification with g on $\hat{\mathcal{R}}$, and that f is P-admissible on $\mathcal{R}$. We construct a g' as follows; let $g'(R) = g(R)$ for all $R \in \hat{\mathcal{R}}$, and let $g'(R) = f(R)$ for all $R \not\in \hat{\mathcal{R}}$. Now f is P-robust to misspecification with g' on $\mathcal{R}$, and $g(R) = g'(R)$ for all $R \in \hat{\mathcal{R}}$. The other direction is straightforward. □