QUANTIFYING THE SENSITIVITY OF INVERSE REINFORCEMENT LEARNING TO MISSPECIFICATION

Joar Skalse & Alessandro Abate
Department of Computer Science
Oxford University
{joar.skalse,aabate}@cs.ox.ac.uk

ABSTRACT

Inverse reinforcement learning (IRL) aims to infer an agent's *preferences* (represented as a reward function R) from their *behaviour* (represented as a policy π). To do this, we need a *behavioural model* of how π relates to R. In the current literature, the most common behavioural models are *optimality*, *Boltzmann-rationality*, and *causal entropy maximisation*. However, the true relationship between a human's preferences and their behaviour is much more complex than any of these behavioural models. This means that the behavioural models are *misspecified*, which raises the concern that they may lead to systematic errors if applied to real data. In this paper, we analyse how sensitive the IRL problem is to misspecification of the behavioural model. Specifically, we provide necessary and sufficient conditions that completely characterise how the observed data may differ from the assumed behavioural model without incurring an error above a given threshold. In addition to this, we also characterise the conditions under which a behavioural model is robust to small perturbations of the observed policy, and we analyse how robust many behavioural models are to misspecification of their parameter values (such as e.g. the discount rate). Our analysis suggests that the IRL problem is highly sensitive to misspecification, in the sense that very mild misspecification can lead to very large errors in the inferred reward function.

1 INTRODUCTION

Inverse reinforcement learning (IRL) is a subfield of machine learning that aims to develop techniques for inferring an agent's *preferences* based on their *actions* in a sequential decision-making problem (Ng & Russell, 2000). There are many motivations for IRL. One motivation is to use it as a tool for *imitation learning*, where the objective is to replicate the behaviour of an expert in some task (e.g. Hussein et al., 2017). In this context, it is not essential that the inferred preferences reflect the actual intentions of the expert, as long as they improve the imitation learning process. Another motivation for IRL is to use it as a tool for *preference elicitation*, where the objective is to understand an agent's goals or desires (e.g. Hadfield-Menell et al., 2016). In this context, it is of central importance that the inferred preferences reflect the actual preferences of the observed agent. In this paper, we are primarily concerned with this second motivation.

An IRL algorithm must make assumptions about how the preferences of the observed agent relate to its behaviour. Specifically, in IRL, preferences are typically modelled as a reward function R, and behaviour is typically modelled as a policy π. An IRL algorithm must therefore have a *behavioural model* that describes how π is computed from R; by inverting this model, R can then be deduced based on π. In the current literature, the most common behavioural models are *optimality*, *Boltzmann rationality*, or *causal entropy maximisation*. These behavioural models essentially assume that the observed agent behaves in a way that is (noisily) optimal according to its preferences.

One of the central difficulties in IRL is that the true relationship between a person's preferences and their actions in general is incredibly complex. This means that it typically is very difficult to specify a behavioural model that is perfectly accurate. For example, optimality, Boltzmann-rationality, and causal entropy maximisation are all very simple models that clearly do not capture all the nuances of human behaviour.[1] This means that these behavioural models are *misspecified*, which raises the concern that they might systematically lead to flawed inferences if applied to real data.

In this paper, we study how robust IRL is to misspecification of the behavioural model. If an IRL algorithm that assumes a particular behavioural model is shown data from an agent whose behaviour violates the assumptions behind this model, then will the inferred reward function still be close to the true reward function? We will analyse this question mathematically, and provide several quantitative answers. In particular, we provide necessary and sufficient conditions that completely characterise what types of misspecification a wide class of behavioural models will tolerate. In addition to this, we also study two specific types of misspecification in more detail (namely, perturbation of the observed policy, and misspecification of the parameters in the behavioural model) and provide additional results that describe how they affect the quality of the inferred reward. Our analysis is highly general – it applies to the three behavioural models that are most common in the current IRL literature, but is also directly applicable to a much wider class of models.

The motivation behind this paper is to contribute towards a theoretically principled understanding of when IRL methods are applicable to the problem of preference elicitation. Human behaviour is very complex, and while we can create behavioural models that are more accurate, it will never be realistically possible to create a behavioural model that is totally free from misspecification. It is therefore crucial to have an understanding of how robust the IRL problem is to misspecification, and whether a small amount of misspecification leads to a proportionally small error in the inferred reward function. Our work aims to further our understanding of these question.

1.1 Related Work

There are two previous papers that analyse how robust the IRL problem is to misspecified behavioural models; Hong et al. (2022) and Skalse & Abate (2023). Our work is more complete than these earlier works in several important respects. To start with, our problem setup is both more realistic, and more general. In particular, in order to quantify how robust IRL is to misspecification, we first need a way to formalise what it means for two reward functions to be "close". Skalse & Abate (2023) formalise this in terms of *equivalence relations*, under which two reward functions are either equivalent or not. As such, their analysis is somewhat blunt, and is unable to distinguish between small errors and large errors in the inferred reward. Hong et al. (2022) instead use the ℓ^2-distance between the reward functions. However, this choice is also problematic, because two reward functions can be very dissimilar even though they have a small ℓ^2-distance, and vice versa (cf. Section 2.2). By contrast, our analysis is carried out in terms of specially selected *metrics* on the space of all reward functions, which are backed by strong theoretical guarantees. Moreover, Hong et al. (2022) assume that there is a *unique* reward function that maximises fit to the training data, but this is violated in most real-world cases (Ng & Russell, 2000; Dvijotham & Todorov, 2010; Cao et al., 2021; Kim et al., 2021; Skalse et al., 2022; Schlaginhaufen & Kamgarpour, 2023). In addition to this, many of their results also assume "strong log-concavity", which is a rather opaque condition that is left mostly unexamined. Indeed, Hong et al. (2022) explicitly do not answer if strong log-concavity should be expected to hold under typical circumstances. Both Skalse & Abate (2023) and Hong et al. (2022) recognise these issues as limitations that should be lifted in future work. The analysis we carry out in this paper is not subject to any of these limitations. Moreover, in addition to being based on a more sound problem formulation, our paper also contains several novel results that are not analogous to any results derived by Skalse & Abate (2023) or Hong et al. (2022).

There are also earlier papers that study some specific types of misspecification in IRL. In particular, Freedman et al. (2020) study the effects of misspecified *choice sets* in IRL, and show that such misspecification in some cases can be catastrophic, and Viano et al. (2021) study the effects of misspecified *environment dynamics*, and propose an algorithm for reducing this effect. By contrast, we present a broader analysis that covers *all* forms of misspecification, within a single framework. Also relevant is the work by Armstrong & Mindermann (2018), who show that it is impossible to simultaneously learn a reward function and a behavioural model from a single data set, given an inductive bias towards joint simplicity.

[1]Indeed, there are detectable differences between data collected from human subjects and data synthesised using these standard behavioural models, see Orsini et al. (2021).

1.2 PRELIMINARIES

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*, $\tau : \mathcal{S}{\times}\mathcal{A} \to \Delta(\mathcal{S})$ 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 $\Delta(X)$ denotes the set of all probability distributions over X. In this paper, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite, and that all states are reachable under τ and μ_0. A *policy* is a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. 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* J gives the expected trajectory return given a policy, $J(\pi) = \mathbb{E}_{\xi\sim\pi}[G(\xi)]$. A policy maximising 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, and the Q-function is $Q^\pi(s, a) = \mathbb{E}[R(s, a, S') + \gamma V^\pi(S')]$. $Q^\star$ and $V^\star$ denote the optimal Q- and value-function.

An IRL algorithm needs a *behavioural model* that describes how the observed policy π relates to the underlying reward function R. In the current IRL literature, the most common models are *optimality*, where it is assumed that π is optimal under R (e.g. Ng & Russell, 2000), *Boltzmann-rationality*, where it is assumed that $\mathbb{P}(\pi(s) = a) \propto e^{\beta Q^\star(s,a)}$, where β is a temperature parameter (e.g. Ramachandran & Amir, 2007), and *maximal causal entropy (MCE)*, where it is assumed that π maximises the causal entropy objective, which is given by $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (R(S_t, A_t, S_{t+1}) + \alpha H(\pi(S_t)))]$, where α is a weight and H is the Shannon entropy function (e.g. Ziebart, 2010).

Two reward functions R_1, R_2 are said to differ by *potential shaping* (with γ) if there is a function $\Phi : \mathcal{S} \to \mathbb{R}$ such that $R_2(s, a, s') = R_1(s, a, s') + \gamma\Phi(s') - \Phi(s)$ (Ng et al., 1999), by *S'-redistribution* (with τ) if $\mathbb{E}_{S'\sim\tau(s,a)}[R_1(s, a, S')] = \mathbb{E}_{S'\sim\tau(s,a)}[R_2(s, a, S')]$ (Skalse et al., 2022), and by *positive linear scaling* if there is a positive constant c such that $R_2 = c \cdot R_1$.

Given a set X, a *pseudometric* on X is any function $d : X \times X \to \mathbb{R}$ that satisfies the conditions that $d(x, x) = 0$, $d(x, y) \geq 0$, $d(x, y) = d(y, x)$, and $d(x, z) \leq d(x, y) + d(y, z)$, for all $x, y, z \in X$. A *metric* additionally satisfies that if $d(x, y) = 0$ then $x = y$.

If a reward function R satisfies that $J(\pi_1) = J(\pi_2)$ for all π_1 and π_2, then we say that R is *trivial*. All constant reward functions are trivial, but there are always non-constant trivial rewards as well.[2]

2 THEORETICAL FRAMEWORK

In this section, we will introduce the theoretical definitions and machinery that we will later use to analyse how robust the IRL problem is to different forms of misspecification.

2.1 DEFINING MISSPECIFICATION ROBUSTNESS

Before we can derive formal results about what types of misspecification different IRL algorithms are robust to, we first need to construct an abstract model of the IRL problem. First of all, assume that we have a fixed set of states $\mathcal{S}$ and a fixed set of actions $\mathcal{A}$, that $\mathcal{R}$ is the set of all reward functions $R : \mathcal{S}{\times}\mathcal{A}{\times}\mathcal{S} \to \mathbb{R}$, and that Π is the set of all policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We say that a *behavioural model* is a function $f : \mathcal{R} \to \Pi$, i.e. a function that takes a reward function and returns a policy. For example, the function that, given R, returns the Boltzmann-rational policy of R under transition function τ, discount γ, and temperature β, is an example of a behavioural model. Using this, we can now model the IRL learning problem as follows: first, we assume that there is a true underlying reward function $R^\star$, and that the training data is generated by a behavioural model g, so that the learning algorithm observes the policy π given by $g(R^\star)$. Moreover, we assume that an IRL algorithm $\mathcal{L}$ has a model f of how the observed policy π relates to $R^\star$, where f is also a behavioural model, such that $\mathcal{L}$ converges to a reward function R_h which satisfies $f(R_h) = \pi = g(R^\star)$. If $f \neq g$, then f is *misspecified*, and otherwise f is correctly specified.

For convenience, if $f(R_1) = f(R_2)$ whenever R_1 and R_2 differ by potential shaping, then we say that f is *invariant to potential shaping*, and similarly for S'-redistribution and positive linear scaling.

[2]These are given by applying potential shaping and S'-redistribution to a constant reward function.

In many contexts, we have prior information about $R^\star$, beyond the information provided by $g(R^\star)$. For example, we may know that $R^\star$ is defined in terms of some sparse state features, or we may know that it only depends on the states, and so on. To model this, we give our definitions relative to an arbitrary set of reward functions $\hat{\mathcal{R}} \subseteq \mathcal{R}$. We then assume that the true reward function $R^\star$ is contained in $\hat{\mathcal{R}}$, and that the learning algorithm $\mathcal{L}$ will learn a reward function R_h that is also contained in $\hat{\mathcal{R}}$. Unless otherwise stated, our results apply for any choice of $\hat{\mathcal{R}}$ (including $\hat{\mathcal{R}} = \mathcal{R}$).

Intuitively speaking, we want to say that a behavioural model f is robust to misspecification with g if a learning algorithm that is based on f is guaranteed to learn a reward function that is "close" to the true reward function if it is trained on data generated from g. To make this statement formal, we need a definition of what it means for two reward functions to be "close". In this paper, we assume that this is defined in terms of some pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$. In Section 2.2, we discuss how to choose this pseudometric. Unless otherwise stated, our results apply for any choice of pseudometric. Note also that while $d^{\mathcal{R}}$ of course may be a proper metric, we allow it to be a pseudometric since we may want to consider distinct reward functions to be equivalent. Using this, we can now give our formal definition of misspecification robustness:

Definition 1. Given a set of reward functions $\hat{\mathcal{R}} \subseteq \mathcal{R}$, a pseudometric $d^{\mathcal{R}}$ on $\hat{\mathcal{R}}$, and two behavioural models $f, g : \hat{\mathcal{R}} \to \Pi$, we say that f is ϵ-robust to misspecification with g if each of the following conditions are satisfied:

1. If $f(R_1) = g(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$, for all $R_1, R_2 \in \hat{\mathcal{R}}$.
2. If $f(R_1) = f(R_2)$ then $d^{\mathcal{R}}(R_1, R_2) \leq \epsilon$, for all $R_1, R_2 \in \hat{\mathcal{R}}$.
3. $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ on $\hat{\mathcal{R}}$.
4. There exists $R_1, R_2 \in \hat{\mathcal{R}}$ such that $f(R_1) \neq g(R_2)$.

Before moving on, let us explain each of these conditions intuitively. The first condition is saying that any learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that has a distance of at most ϵ to the true reward function when trained on data generated from g; this is the core property of misspecification robustness. Similarly, the second condition is saying that any learning algorithm $\mathcal{L}$ based on f is guaranteed to learn a reward function that has a distance of at most ϵ to the true reward function *when there is no misspecification*; this condition is included to rule out certain pathological edge cases. The third condition is effectively saying that $\mathcal{L}$ can never observe a policy that is impossible according to its model. Depending on how $\mathcal{L}$ behaves, it may in some cases be possible to drop this condition, but we include it to make our analysis as general as possible. The fourth condition is simply saying that f and g are distinct on $\hat{\mathcal{R}}$ (otherwise f is not misspecified!). More extensive discussion of Definition 1, including more subtle issues, is given in Appendix A.

We are particularly interested in the three behavioural models that are most common in the current IRL literature, and so we will use special notation for these models. Given a transition function τ and a discount parameter γ, let $b_{\tau,\gamma,\beta} : \mathcal{R} \to \Pi$ be the function that returns the Boltzmann-rational policy of R with temperature β, and let $c_{\tau,\gamma,\alpha} : \mathcal{R} \to \Pi$ be the function that returns the MCE policy of R with weight α. Similarly, let $o_{\tau,\gamma} : \mathcal{R} \to \Pi$ be the function that returns the optimal policy of R that takes all optimal actions with equal probability.

2.2 Reward Function Metrics

We wish to obtain results that describe how different the learnt reward function R_h may be compared to the underlying true reward function $R^\star$, given different forms of misspecification. To do this, we need a way to *quantify* the difference between R_h and $R^\star$, in the form of a pseudometric $d^{\mathcal{R}}$ on $\mathcal{R}$. However, finding an appropriate choice of $d^{\mathcal{R}}$ is not straightforward. For example, suppose we simply let $d^{\mathcal{R}}(R_h, R^\star) = ||R_h - R^\star||_2$. In that case, we would have that $d^{\mathcal{R}}(R_h, R^\star)$ can be arbitrarily *large*, even if R_h and $R^\star$ have the *same* ordering of policies, and similarly, that $d^{\mathcal{R}}(R_h, R^\star)$ can be arbitrarily *small*, even if R_h and $R^\star$ have the *opposite* ordering of policies.[3] This means that the ℓ^2-distance between R_h and $R^\star$ does not quantify their qualitative difference in a useful way.

[3]To see this, let R be any nontrivial reward function. Then for any positive c, we have that R and cR have the same ordering of policies, but by making c large, we can make $||R - cR||_2$ arbitrarily large. Similarly, for

Intuitively, we want a pseudometric $d^{\mathcal{R}}$ with the property that $d^{\mathcal{R}}(R_h, R^\star)$ is small if (and only if) it would be safe to optimise R_h instead of $R^\star$. As we have just argued, the ℓ^2-norm does *not* satisfy this condition. Instead, we will use a ***STARC-metric***, introduced by Skalse et al. (2023):

Definition 2. Let τ be a transition function and γ be a discount factor. Given a reward function R, let $V_{\tau,\gamma}(R)$ be the set of all reward functions that differ from R by potential shaping (with γ) and S'-redistribution (with τ). Let $c^{\mathrm{STARC}}_{\tau,\gamma} : \mathcal{R} \to \mathcal{R}$ be the function where $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ returns the element of $V_{\tau,\gamma}(R)$ that has the smallest ℓ^2-norm.[4] Moreover, let $s^{\mathrm{STARC}}_{\tau,\gamma} : \mathcal{R} \to \mathcal{R}$ be the function where $s^{\mathrm{STARC}}_{\tau,\gamma}(R) = c^{\mathrm{STARC}}_{\tau,\gamma}(R)/||c^{\mathrm{STARC}}_{\tau,\gamma}(R)||_2$ if $||c^{\mathrm{STARC}}_{\tau,\gamma}(R)||_2 > 0$, and $c^{\mathrm{STARC}}_{\tau,\gamma}(R)$ otherwise. Finally, let $d^{\mathrm{STARC}}_{\tau,\gamma} : \mathcal{R} \times \mathcal{R} \to [0,1]$ be the function given by

$$d^{\mathrm{STARC}}_{\tau,\gamma}(R_1, R_2) = 0.5 \cdot ||s^{\mathrm{STARC}}_{\tau,\gamma}(R_1) - s^{\mathrm{STARC}}_{\tau,\gamma}(R_2)||_2.$$

There is extensive justification for measuring the distance between reward functions in terms of $d^{\mathrm{STARC}}_{\tau,\gamma}$. However, because the full justification is quite long, and because this justification is not essential to understand most of our results, we have decided to provide it in Appendix B, instead of the main text. Appendix B also contains a less formal, more intuitive explanation of Definition 2. For now, it is sufficient to know that $d^{\mathrm{STARC}}_{\tau,\gamma}(R_1, R_2) = 0$ if and only if R_1 and R_2 induce the same ordering of policies under τ and γ. Moreover, while some of our results are given in terms of $d^{\mathrm{STARC}}_{\tau,\gamma}$, we also have many results that apply for any choice of pseudometric on $\mathcal{R}$. This will be clearly stated in each theorem.

Another prominent pseudometric on $\mathcal{R}$ is EPIC, proposed by Gleave et al. (2021). Appendix C puts forward reasons to not use EPIC for the kind of analysis that we undertake this paper.

2.3 Background Results

In this section, we give a few important results from earlier works that we will rely on throughout this paper, and which will also be helpful for contextualising our results. First, it is useful to know under what conditions two reward functions have the same ordering of policies:

Proposition 1. *$(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma)$ and $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_2, \gamma)$ have the same ordering of policies if and only if R_1 and R_2 differ by potential shaping (with γ), S'-redistribution (with τ), and positive linear scaling.*

For a proof of Proposition 1, see Skalse & Abate (2023) (their Theorem 2.6). Also recall that $d^{\mathrm{STARC}}_{\tau,\gamma}(R_1, R_2) = 0$ if and only if R_1 and R_2 induce the same ordering of policies. Next, it is also useful to know under what conditions $f(R_1) = f(R_2)$ when f is the Boltzmann-rational model or the maximal causal entropy model:

Proposition 2. *For any τ, γ, and β, we have that $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$ are invariant to potential shaping (with γ) and S'-redistribution (with τ), and no other transformations.*

For a proof of Proposition 2, see Skalse et al. (2022) (their Theorem 3.3). Note that Propositions 1 and 2 together imply that any learning algorithm $\mathcal{L}$ that is based on either the Boltzmann-rational model or the MCE-model is guaranteed to learn a reward function R_h that has the same ordering of policies as the true reward function $R^\star$ when there is no misspecification.

3 Misspecification Robustness

In this section, we provide our main results about the misspecification robustness of various behavioural models, including the behavioural models that are most common in the IRL literature. Section 3.1 is quite dense, but 3.2 and 3.3 both provide more intuitive takeaways.

3.1 Necessary and Sufficient Conditions

Given a behavioural model f, it is desirable to have necessary and sufficient conditions that completely characterise when f is robust to misspecification with g. Surprisingly, our first result shows

any positive ϵ, we have that ϵR and $-\epsilon R$ have the opposite ordering of policies. However, by making ϵ small, we can make $||\epsilon R - (-\epsilon R)||_2$ arbitrarily small.

[4] $V_{\tau,\gamma}(R)$ is an affine subspace of $\mathcal{R}$, so there is a unique element of $V_{\tau,\gamma}(R)$ that minimises the ℓ^2-norm.

that if $f(R_1) = f(R_2) \implies d^{\mathcal{R}}(R_1, R_2) = 0$, then we can derive such necessary and sufficient conditions in a relatively straightforward way:

Theorem 1. *Let $\hat{\mathcal{R}}$ be a set of reward functions, let $f : \hat{\mathcal{R}} \to \Pi$ be a behavioural model, and let $d^{\mathcal{R}}$ be a pseudometric on $\hat{\mathcal{R}}$. Assume that $f(R_1) = f(R_2) \implies d^{\mathcal{R}}(R_1, R_2) = 0$ for all $R_1, R_2 \in \hat{\mathcal{R}}$. Then f is ϵ-robust to misspecification with g (as defined by $d^{\mathcal{R}}$) if and only if $g = f \circ t$ for some $t : \hat{\mathcal{R}} \to \hat{\mathcal{R}}$ such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$ for all $R \in \hat{\mathcal{R}}$, and such that $f \neq g$.*

Recall that $d^{\text{STARC}}_{\tau,\gamma}(R_1, R_2) = 0$ if and only if R_1 and R_2 induce the same ordering of policies. Thus, if our reward metric is $d^{\text{STARC}}_{\tau,\gamma}$, then Theorem 1 applies to any behavioural model f for which $f(R_1) = f(R_2)$ implies that R_1 and R_2 have the same policy ordering. Moreover, also recall that both the Boltzmann-rational model and the MCE model satisfy this condition (Propositions 1 and 2). Thus, if we can find the set T_ϵ of all transformations $t : \hat{\mathcal{R}} \to \hat{\mathcal{R}}$ such that $d^{\mathcal{R}}(R, t(R)) \leq \epsilon$, then we can get the set of all behavioural models g to which f is ϵ-robust to misspecification by simply composing f with each element of T_ϵ. We next derive T_ϵ:

Proposition 3. *A transformation $t : \mathcal{R} \to \mathcal{R}$ satisfies that $d^{\text{STARC}}_{\tau,\gamma}(R, t(R)) \leq \epsilon$ for all $R \in \mathcal{R}$ if and only if t can be expressed as $t_1 \circ \cdots \circ t_{n-1} \circ t_n \circ t_{n+1} \circ \cdots \circ t_m$ for some n and m where*

$$||R - t_n(R)||_2 \leq ||c^{\text{STARC}}_{\tau,\gamma}(R)||_2 \cdot \sin(2 \arcsin(\epsilon/2))$$

for all R, and for all $i \neq n$ and all R, we have that R and $t_i(R)$ differ by potential shaping (with γ), S'-redistribution (with τ), or positive linear scaling.

The statement of Proposition 3 is quite terse, so let us briefly unpack it. First of all, $d^{\text{STARC}}_{\tau,\gamma}$ is invariant to any transformation that preserves the policy ordering of the reward function, and these transformations are exactly those that can be expressed as a combination of potential shaping, S'-redistribution, and positive linear scaling. As such, we can apply an arbitrary number of such transformations. Moreover, we can also transform R in any way that does not change the standardised reward function $s^{\text{STARC}}_{\tau,\gamma}(R)$ by more than ϵ; this is equivalent to the stated condition on t_n. Note that $\sin(2 \arcsin(\epsilon/2)) \approx \epsilon$ for small ϵ, so the right-hand side is approximately equal to $\epsilon \cdot ||c^{\text{STARC}}_{\tau,\gamma}(R)||_2$. However, also note that $||c^{\text{STARC}}_{\tau,\gamma}(R)||_2 \leq ||R||_2$. Intuitively speaking, a reward transformation satisfies the conditions given in Proposition 3 if it never changes the policy order of the reward function by a large amount.

Using this, we can now state necessary and sufficient conditions that completely characterise all types of misspecification that the Boltzmann-rational model and the MCE model will tolerate:

Corollary 1. *Let $\hat{\mathcal{R}}$ be a set of reward functions, τ be a transition function, γ a discount factor, β a temperature parameter, and α a weight parameter. Let $\hat{T}_\epsilon$ be the set of all functions $t : \mathcal{R} \to \mathcal{R}$ that satisfy Proposition 3, and additionally satisfy that $t(R) \in \hat{\mathcal{R}}$ for all $R \in \hat{\mathcal{R}}$. Then $b_{\tau,\gamma,\beta} : \hat{\mathcal{R}} \to \Pi$ is ϵ-robust to misspecification with g (as defined by $d^{\text{STARC}}_{\tau,\gamma}$) if and only if $g = b_{\tau,\gamma,\beta} \circ t$ for some $t \in \hat{T}_\epsilon$ such that $b_{\tau,\gamma,\beta} \neq g$, and $c_{\tau,\gamma,\alpha} : \hat{\mathcal{R}} \to \Pi$ is ϵ-robust to misspecification with g (as defined by $d^{\text{STARC}}_{\tau,\gamma}$) if and only if $g = c_{\tau,\gamma,\alpha} \circ t$ for some $t \in \hat{T}_\epsilon$ such that $c_{\tau,\gamma,\alpha} \neq g$.*

In principle, Corollary 1 completely describes the misspecification robustness of the Boltzmann-rational model and of the MCE model (for any $\hat{\mathcal{R}}$). However, the statement of Corollary 1 is rather opaque, and difficult to interpret qualitatively. For this reason, we will in the subsequent sections examine a few important special types of misspecification, and analyse them individually.

We should also briefly comment on the fact that Corollary 1 does not cover $o_{\tau,\gamma}$, i.e. the optimality model. The reason for this is that, unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, there are reward functions R_1, R_2 such that $o_{\tau,\gamma}(R_1) = o_{\tau,\gamma}(R_2)$, but $d^{\text{STARC}}_{\tau,\gamma}(R_1, R_2) > 0$. This ought to be intuitive: two reward functions can have the same optimal policies, but have different policy orderings. This means that Theorem 1 does not apply to $o_{\tau,\gamma}$ when $d^{\mathcal{R}} = d^{\text{STARC}}_{\tau,\gamma}$. Moreover:

Proposition 4. *Unless $|\mathcal{S}| = 1$ and $|\mathcal{A}| = 2$, then for any τ and any γ there exists an $E > 0$ such that for all $\epsilon < E$, there is no behavioural model g such that $o_{\tau,\gamma}$ is ϵ-robust to misspecification with g (as defined by $d^{\text{STARC}}_{\tau,\gamma}$).*

This is simply a consequence of the fact that the second condition of Definition 1 will be violated for any ϵ that is sufficiently small. An analogous result will hold for any behavioural model f and any pseudometric $d^{\mathcal{R}}$ for which $f(R_1) = f(R_2) \not\Longrightarrow d^{\mathcal{R}}(R_1, R_2) = 0$.

3.2 PERTURBATION ROBUSTNESS

It is interesting to know whether or not a behavioural model f is robust to misspecification with any behavioural model g that is "close" to f. But what does it mean for f and g to be "close"? One option is to say that f and g are close if they always produce similar policies. In this section, we will explore under what conditions f is robust to such misspecification, and provide necessary and sufficient conditions. Our results are given relative to a pseudometric d^Π on Π. For example, $d^\Pi(\pi_1, \pi_2)$ may be the ℓ^2-distance between π_1 and π_2, or it may be the KL divergence between their trajectory distributions, or it may be the ℓ^2-distance between their occupancy measures, etc. As usual, our results apply for any choice of d^Π unless otherwise stated. We can now define a notion of a *perturbation* and a notion of *perturbation robustness*:

Definition 3. Let $\hat{\mathcal{R}}$ be a set of reward functions, let $f, g : \hat{\mathcal{R}} \to \Pi$ be two behavioural models, and let d^Π be a pseudometric on Π. Then g is a δ-perturbation of f if $g \neq f$ and for all $R \in \hat{\mathcal{R}}$ we have that $d^\Pi(f(R), g(R)) \leq \delta$.

Definition 4. Let $\hat{\mathcal{R}}$ be a set of reward functions, let $f : \hat{\mathcal{R}} \to \Pi$ be a behavioural model, let $d^{\mathcal{R}}$ be a pseudometric on $\hat{\mathcal{R}}$, and let d^Π be a pseudometric on Π. Then f is ϵ-robust to δ-perturbation if f is ϵ-robust to misspecification with g (as defined by $d^{\mathcal{R}}$) for any behavioural model $g : \hat{\mathcal{R}} \to \Pi$ that is a δ-perturbation of f (as defined by d^Π) with $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$.

A δ-perturbation of f simply is any function that is similar to f on all inputs, and f is ϵ-robust to δ-perturbation if a small perturbation of the observed policy leads to a small error in the inferred reward function. It would be desirable for a behavioural model to be robust in this sense. To start with, this captures any form of misspecification that always leads to a small change in the final policy. Moreover, in practice, we can often not observe the exact policy of the demonstrator, and must instead approximate it from a number of samples. In this case, we should also expect to infer a policy that is a perturbation of the true policy. Before moving on, we need one more definition:

Definition 5. Let $\hat{\mathcal{R}}$ be a set of reward functions, let $f : \hat{\mathcal{R}} \to \Pi$ be a behavioural model, let $d^{\mathcal{R}}$ be a pseudometric on $\hat{\mathcal{R}}$, and let d^Π be a pseudometric on Π. Then f is ϵ/δ-separating if $d^{\mathcal{R}}(R_1, R_2) > \epsilon \implies d^\Pi(f(R_1), f(R_2)) > \delta$ for all $R_1, R_2 \in \hat{\mathcal{R}}$.

Intuitively speaking, f is ϵ/δ-separating if reward functions that are far apart, are sent to policies that are far apart.[5] Using this, we can now state our main result for this section:

Theorem 2. *Let $\hat{\mathcal{R}}$ be a set of reward functions, let $f : \hat{\mathcal{R}} \to \Pi$ be a behavioural model, let $d^{\mathcal{R}}$ be a pseudometric on $\hat{\mathcal{R}}$, and let d^Π be a pseudometric on Π. Then f is ϵ-robust to δ-perturbation (as defined by $d^{\mathcal{R}}$ and d^Π) if and only if f is ϵ/δ-separating (as defined by $d^{\mathcal{R}}$ and d^Π).*

We have thus obtained necessary and sufficient conditions that describe when a behavioural model is robust to perturbations — namely, it has to be the case that this behavioural model sends reward functions that are far apart, to policies that are far apart. This ought to be quite intuitive; if two policies are close, then perturbations may lead us to conflate them. To be sure that the learnt reward function is close to the true reward function, we therefore need it to be the case that policies that are close always correspond to reward functions that are close (or, conversely, that reward functions which are far apart correspond to policies which are far apart).

Our next question is, of course, whether or not the standard behavioural models are ϵ/δ-separating. Surprisingly, we will show that this is *not* the case, when the distance between reward functions is measured using $d^{\mathrm{STARC}}_{\tau,\gamma}$, and the policy metric d^Π is similar to Euclidean distance. Moreover, we only need very mild assumptions about the behavioural model to obtain this result:

Theorem 3. *Let $d^{\mathcal{R}}$ be $d^{\mathrm{STARC}}_{\tau,\gamma}$, and let d^Π be a pseudometric on Π which satisfies the condition that for all δ there exists a δ' such that if $||\pi_1 - \pi_2||_2 < \delta'$ then $d^\Pi(\pi_1, \pi_2) < \delta$. Let c be any positive constant, and let $\hat{\mathcal{R}}$ be a set of reward functions such that if $||R||_2 = c$ then $R \in \hat{\mathcal{R}}$. Let $f : \hat{\mathcal{R}} \to \Pi$ be a behavioural model that is continuous. Then f is not ϵ/δ-separating for any $\epsilon < 1$ or $\delta > 0$.*

[5] Note that this definition is *not* saying that reward functions which are close must be sent to policies which are close. In other words, f being ϵ/δ-separating is *not* a continuity condition. It is also not a local property of f, but rather, a global property. It is, however, a continuity condition on the inverse of f.

This theorem is telling us several things at once. To make things easy, we can begin by noting that we may let $\hat{\mathcal{R}} = \mathcal{R}$, and that we may assume that d^Π simply is the ℓ^2-norm, i.e. $d^\Pi(\pi_1, \pi_2) = ||\pi_1 - \pi_2||_2$. Theorem 3 is then telling us that *no continuous behavioural model* is ϵ/δ-separating for any ϵ or δ (and therefore, by Theorem 2, also not ϵ-robust to δ-perturbation for any ϵ or δ). Note that the Boltzmann-rational model and the maximal causal entropy model (i.e. $b_{\tau,\gamma,\beta}$ and $c_{\tau,\gamma,\alpha}$) both are continuous, and hence subject to Theorem 3. The condition given on d^Π in Theorem 3 is simply a generalisation, that covers other policy-metrics than the ℓ^2-norm.[6] Similarly, the condition on $\hat{\mathcal{R}}$ is also a generalisation to certain restricted reward spaces. We give a more in-depth, intuitive interpretation of Theorem 3, and an explanation of why it is true, in Appendix D.

3.3 Misspecified Parameters

A behavioural model is typically defined relative to some parameters. For example, the Boltzmann-rational model is defined relative to a temperature parameter β and a discount parameter γ, as well as the transition dynamics τ. Moreover, determining the exact values of these parameters ex post facto can often be quite difficult. For example, there is a sizeable literature that attempts to estimate the rate at which humans discount future reward, and there is a fairly large range in the estimates that this literature produces (e.g. Percoco & Nijkamp, 2009). It is therefore interesting to know to what extent a behavioural model is robust to misspecification of its parameters. If π is Boltzmann-rational for discount parameter γ_1, but an IRL algorithm interprets it as being Boltzmann-rational for discount parameter γ_2, where $\gamma_1 \approx \gamma_2$, then are we still guaranteed to learn a reward function that is close to the true reward function? These are the questions that we will study in this section.

We will first consider the case when the discount parameter, γ, is misspecified. Say that a transition function τ is *trivial* if for all states s and all actions a_1, a_2, we have that $\tau(s, a_1) = \tau(s, a_2)$. We now have the following rather surprising result:

Theorem 4. *If $f_\gamma : \mathcal{R} \to \Pi$ is invariant to potential shaping with γ, and $\gamma_1 \neq \gamma_2$, then f_{γ_1} is not ϵ-robust to misspecification with f_{γ_2} under $d^{\mathrm{STARC}}_{\tau,\gamma_3}$ for any non-trivial τ, any γ_3, and any $\epsilon < 0.5$.*

Note that Theorem 4 permits that $\gamma_3 = \gamma_1$ or $\gamma_3 = \gamma_2$. Of course, any interesting environment will have a non-trivial transition function, so this requirement is very mild. Moreover, a $d^{\mathrm{STARC}}_{\tau,\gamma}$-distance of 0.5 is very large; this corresponds to the case where the reward functions are nearly orthogonal. This means that Theorem 4 is saying that if a behavioural model f is invariant to potential shaping, then it is not robust to any misspecification of the discount parameter. Note that this holds even if γ_1 and γ_2 are arbitrarily close! Moreover, optimal policies, Boltzmann-rational policies, and MCE policies are all invariant to potential shaping, and hence $o_{\tau,\gamma}$, $b_{\tau,\gamma,\beta}$, and $c_{\tau,\gamma,\alpha}$ are subject to Theorem 4. In general, we should expect any behavioural model that uses exponential discounting to be invariant to potential shaping, and so Theorem 4 will apply very widely.

We will next consider the case when the transition function, τ, is misspecified. Here, we similarly find that a very wide class of behavioural models are non-robust to any amount of misspecification:

Theorem 5. *If $f_\tau : \mathcal{R} \to \Pi$ is invariant to S'-redistribution with τ, and $\tau_1 \neq \tau_2$, then f_{τ_1} is not ϵ-robust to misspecification with f_{τ_2} under $d^{\mathrm{STARC}}_{\tau_3,\gamma}$ for any τ_3, any γ, and any $\epsilon < 0.5$.*

Note that Theorem 5 permits that $\tau_3 = \tau_1$ or $\tau_3 = \tau_2$. Thus, Theorem 5 is saying that if a behavioural model f is invariant to S'-redistribution, then it is not robust to any degree of misspecification of τ (even if τ_1 and τ_2 are arbitrarily close). Moreover, optimal policies, Boltzmann-rational policies, and maximal causal entropy policies, are all invariant to S'-redistribution, and hence $o_{\tau,\gamma}$, $b_{\tau,\gamma,\beta}$, and $c_{\tau,\gamma,\alpha}$ are subject to Theorem 5. Indeed, since S'-redistribution does not change the expected value of any policy, we should expect almost all sensible behavioural models to be invariant to S'-redistribution. As such, Theorem 5 will also apply very widely.

Theorem 4 and 5 show that a very wide range of behavioural models in principle are highly sensitive to arbitrarily small misspecification of two of their core parameters. To make this result more accessible and easier to understand, we have included two examples in Appendix E that explain the intuition behind these two theorems.

[6] Note that while Theorem 3 uses a "special" pseudometric on $\mathcal{R}$, in the form of $d^{\mathrm{STARC}}_{\tau,\gamma}$, we do not need to use a special (pseudo)metric on Π, because for policies, ℓ^2 does capture the relevant notion of similarity.

Before moving on, we also want to note that the Boltzmann-rational model is robust to arbitrary misspecification of the temperature parameter, β, and that the maximal causal entropy model is robust to arbitrary misspecification of the weight parameter, α. This was shown by Skalse & Abate (2023), in their Theorems 3.2 and 3.4. Specifically, we have that for any τ and γ, any $\epsilon \geq 0$, and any β_1, β_2, α_1, α_2, we have that b_{τ,γ,β_1} is ϵ-robust to misspecification with b_{τ,γ,β_2}, and that c_{τ,γ,α_1} is ϵ-robust to misspecification with c_{τ,γ,α_2}, as defined by $d^{\mathrm{STARC}}_{\tau,\gamma}$. For a detailed description of how to connect the results of Skalse & Abate (2023) to ours, see Appendix G.

4 DISCUSSION

We have quantified how robust IRL is to misspecification of the behavioural model. We first provided necessary and sufficient conditions that fully describe what types of misspecification many behavioural models will tolerate. In principle, these conditions give a complete answer to how tolerant a given behavioural model is to any given type of misspecification. However, these conditions are rather opaque, and difficult to interpret. Therefore, we have also separately provided necessary and sufficient conditions that characterise when a behavioural model is robust to *perturbation*, and we have analysed how robust many behavioural models are to misspecification of the discount factor γ or the environment dynamics τ. Our analysis suggests that the IRL problem is highly sensitive to many plausible forms of misspecification. In particular, a very wide class of behavioural models are unable to guarantee robust inference under arbitrarily small perturbations of the observed policy, or under arbitrarily small misspecification of γ or τ.

Our results present a serious challenge to IRL in the context of preference elicitation. The relationship between human preferences and human behaviour is very complex, and while it is certainly possible to create increasingly accurate models of human behaviour, it will never be realistically possible to create a model that is completely free from all forms of misspecification. Therefore, if IRL is unable to guarantee accurate inferences under even mild misspecification of the behavioural model, as our results suggest, then we should expect it to be very difficult (and perhaps even prohibitively difficult) to guarantee that IRL reliably will produce accurate inferences in real-world situations. This in turn means that IRL should be used cautiously, and that the learned reward functions should be carefully examined and evaluated (as done by e.g. Michaud et al., 2020; Jenner & Gleave, 2022). It also means that we need IRL algorithms that are specifically designed to be more robust under misspecification, such as e.g. that proposed by Viano et al. (2021). It may also be fruitful to combine IRL with other data sources, as done by e.g. Ibarz et al. (2018), or consider policy optimisation algorithms that assume that the reward function may be misspecified, as done by e.g. Krakovna et al. (2018; 2020); Turner et al. (2020); Griffin et al. (2022).

We also need more extensive investigations into the issue of how robust IRL is to misspecification, and there are several ways that our analysis can be extended. First of all, it may in some cases be possible to mitigate some of our negative results if $\mathcal{R}$ is restricted. For example, Theorem 4 and 5 rely on the fact that we for any reward R_1 can find a reward R_2 such that R_1 and R_2 differ by potential shaping or S'-redistribution for a given choice of γ and τ, but such that R_1 and R_2 have a large STARC-distance for other choices of γ and τ. We may thus be able to circumvent these results by restricting $\hat{\mathcal{R}}$ in a way that removes all such reward pairs. However, this is of course not straightforward, not least because we need to ensure that the true reward in fact is contained in $\hat{\mathcal{R}}$. Moreover, some of our results rely on the fact that we use STARC-metrics to quantify the difference between reward functions. While there are compelling theoretical justifications for doing so (cf. Appendix B and C), there may be other relevant options. STARC-metrics are quite strong, and we may be able to derive weaker guarantees using other forms of reward metrics. Furthermore, it may also be fruitful to modify Definition 1, for example by making it more probabilistic, or generalising it in other ways. This topic is discussed in Appendix A. Finally, our analysis can also be extended to other types of behavioural models and other types of misspecification. For example, are policies that use (e.g.) *hyperbolic discounting* subject to a result that is analogous to Theorem 4? Such investigations also present an interesting direction for future work.

REFERENCES

Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31, pp. 5603–5614, Montréal, Canada, 2018. Curran Associates, Inc., Red Hook, NY, USA.

Haoyang Cao, Samuel N. Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *arXiv preprint*, arXiv:2106.03498 [cs.LG], 2021.

Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable MDPs. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 335–342, Haifa, Israel, June 2010. Omnipress, Madison, Wisconsin, USA.

Rachel Freedman, Rohin Shah, and Anca Dragan. Choice set misspecification in reward inference. In *IJCAI-PRICAI-20 Workshop on Artificial Intelligence Safety*, 2020. doi: 10.48550/ARXIV.2101.07691. URL `https://arxiv.org/abs/2101.07691`.

Adam Gleave, Michael D Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=LwEQnp6CYev`.

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

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf`.

Joey Hong, Kush Bhatia, and Anca Dragan. On the sensitivity of reward inference to misspecified human models, 2022.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2), 4 2017. ISSN 0360-0300. doi: 10.1145/3054912. URL `https://doi.org/10.1145/3054912`.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in Atari. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31, pp. 8022–8034, Montréal, Canada, 2018. Curran Associates, Inc., Red Hook, NY, USA.

Erik Jenner and Adam Gleave. Preprocessing reward functions for interpretability, 2022.

Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. Reward identification in inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5496–5505, Virtual, July 2021. PMLR.

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

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

Eric J. Michaud, Adam Gleave, and Stuart Russell. Understanding learned reward functions, 2020.

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, pp. 663–670, Stanford, California, USA, 2000. Morgan Kaufmann Publishers Inc.

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*, pp. 278–287, Bled, Slovenia, 1999. Morgan Kaufmann Publishers Inc.

Manu Orsini, Anton Raichuk, Léonard Hussenot, Damien Vincent, Robert Dadashi, Sertan Girgin, Matthieu Geist, Olivier Bachem, Olivier Pietquin, and Marcin Andrychowicz. What matters for adversarial imitation learning? *arXiv preprint*, arXiv:2106.00672 [cs.LG], 2021. To appear in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

Marco Percoco and Peter Nijkamp. Estimating individual rates of discount: a meta-analysis. *Applied Economics Letters*, 16:1235 – 1239, 2009. URL `https://api.semanticscholar.org/CorpusID:154901975`.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 2586–2591, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc.

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning, 2023.

Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning, 2023.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. *arXiv preprint arXiv:2203.07475*, 2022.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. Starc: A general framework for quantifying differences between reward functions, 2023.

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

Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=t8HduwpoQQv`.

Blake Wulfe, Logan Michael Ellis, Jean Mercat, Rowan Thomas McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=CALFyKVs87`.

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