Misspecification in Inverse Reinforcement Learning

Joar Skalse, Alessandro Abate

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

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). IRL roughly corresponds to the notion of *revealed preferences* in psychology and economics, since it aims to infer *preferences* from *behaviour* (Rothkopf and Dimitrakakis 2011). IRL has many possible applications. For example, it has been used in scientific contexts, as a tool for understanding animal behaviour (Yamaguchi et al. 2018). It can also be used in engineering contexts; many important tasks can be represented as sequential decision problems, where the goal is to maximise a *reward function* over several steps (Sutton and Barto 2018). However, for many complex tasks, it can be very challenging to manually specify a reward function that incentivises the intended behaviour. IRL can then be used to *learn* a good reward function, based on demonstrations of correct behaviour (e.g. Abbeel, Coates, and Ng 2010; Singh et al. 2019). Overall, IRL relates to many fundamental questions about goal-directed behaviour and agent-based modelling.

There are two primary motivations for IRL. The first motivation is to use IRL as a tool for *imitation learning* (e.g. Hussein et al. 2017). For these applications, it is not fundamentally important whether the learnt reward function actually corresponds to the true intentions of the demonstrator, as long as it helps the imitation learning process. The second motivation is to use IRL to understand an agent's preferences and motives (e.g. Hadfield-Menell et al. 2016). From this perspective, the goal is to learn a reward that captures the demonstrator's true intentions. This paper was written with mainly the second motivation in mind.

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.

The motivation behind this work is to provide a theoretically principled understanding of whether and when IRL methods are (or are not) applicable to the problem of inferring a person's (true) preferences and intentions. Human behaviour is very complex, and while a behavioural model can be more or less accurate, it will never be realistically possible to create a behavioural model that is completely free from misspecification (except possibly for in narrow do-

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

mains). Therefore, if we wish to use IRL as a tool for preference elicitation, then it is crucial to have an understanding of how robust the IRL problem is to misspecification. In this paper, we contribute towards building this understanding.

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 (2020) study the effects of *choice set misspecification* in IRL (and reward inference more 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 2021), 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 the question of how much misspecification has to be removed).

Skalse et al. (2022a) study the *partial identifiability* of various reward learning models. Our work uses similar techniques, and can be viewed as an extension of their work. The issue of partial identifiability in IRL has also been studied by Ng and Russell (2000); Dvijotham and Todorov (2010); Cao, Cohen, and Szpruch (2021); Kim et al. (2021).

We will discuss the question of what happens if a reward function is changed or misspecified. This question is also investigated by many previous works, including e.g. Gleave et al. (2021); Skalse et al. (2022b); Jenner, van Hoof, and Gleave (2022); Pan, Bhatia, and Steinhardt (2022).

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*, $\tau : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ 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 *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.

In IRL, it is typically assumed that the preferences of the observed agent are described by a reward function R, that its environment is described by an MDP, and that its behaviour is described by a (stationary) policy π. 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)).

In this paper, 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.

Theoretical Framework

We here introduce the theoretical framework that we will use to analyse how robust various behavioural models are to misspecification. This framework is rather abstract, but it is quite powerful, and makes our analysis easy to carry out.

Definitions and Framework

For a given set of states $\mathcal{S}$ and set of actions $\mathcal{A}$, let $\mathcal{R}$ be the set of all reward functions $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ definable with $\mathcal{S}$ and $\mathcal{A}$. Moreover, if P and Q are partitions of a set X, we write $P \preceq Q$ if $x_1 \equiv_P x_2 \Rightarrow x_1 \equiv_Q x_2$ for $x_1, x_2 \in X$. We will use the following definitions:

1. A *reward object* is a function $f : \mathcal{R} \to X$, where X is any set.
2. The *ambiguity* $\mathrm{Am}(f)$ of f is the partition of $\mathcal{R}$ given by $R_1 \equiv_f R_2 \iff f(R_1) = f(R_2)$.
3. Given a partition P of $\mathcal{R}$, we say that f is *P-admissible* if $\mathrm{Am}(f) \preceq P$, i.e. $f(R_1) = f(R_2) \Rightarrow R_1 \equiv_P R_2$.
4. Given a partition P of $\mathcal{R}$, we say that f is *P-robust to misspecification* with g if f is P-admissible, $f \neq g$, $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, and $f(R_1) = g(R_2) \implies R_1 \equiv_P R_2$.
5. A *reward transformation* is a function $t : \mathcal{R} \to \mathcal{R}$.
6. If F and G are sets of reward transformations, then $F \circ G$ is the set of all transformations that can be obtained by composing transformations in F and G arbitrarily, in any order. Note that $F \circ G = G \circ F$.

We will now explain and justify each of these definitions. 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 $\mathrm{Am}(f) \preceq \mathrm{Am}(g)$ and $f \neq g$, 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. (2022a).

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 $R^\star$ if they are 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 .

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 $f(R) = f(t(R))$.

Fundamental Lemmas

We here give two fundamental lemmas that we will later use to prove our core results. These lemmas can also be used to easily derive the misspecification robustness of new models, beyond those considered in this work. All of our proofs are provided in the supplementary material, which also contains several additional results about our framework.

Lemma 1. *If f is not P-robust to misspecification with g, and $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$, then for any h, $h \circ f$ is not P-robust to misspecification with $h \circ g$.*

This lemma states that if we have an object $h \circ f$ that can be computed from some intermediary object f, and f is not P-robust to some form of misspecification, then $h \circ f$ is likewise not robust to the corresponding misspecification. In other words, any misspecification that f is sensitive to, is "inherited" by all objects that can be computed from f.

Lemma 2. *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 $t \in T$.

Reward Transformations

We here introduce several classes of reward transformations, that we will later use to express our results. First recall *potential shaping* (Ng, Harada, and Russell 1999):

Definition 3 (Potential Shaping)**.** A *potential function* is a function $\Phi : \mathcal{S} \to \mathbb{R}$. 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. (2022a), starting with *S'-redistribution.*

Definition 4 (S'-Redistribution)**.** Given a transition function τ, we say that $R_2 \in \mathcal{R}$ is produced by *S'-redistribution* of $R_1 \in \mathcal{R}$ if

$$\mathbb{E}_{S' \sim \tau(s,a)}\left[R_1(s, a, S')\right] = \mathbb{E}_{S' \sim \tau(s,a)}\left[R_2(s, a, S')\right].$$

If s_1, $s_2 \in \mathrm{Supp}(\tau(s,a))$ then S'-redistribution can increase $R(s,a,s_1)$ if it decreases $R(s,a,s_2)$ proportionally. S'-redistribution can also change R arbitrarily for transitions that occur with probability 0. We next consider *optimality-preserving transformations*:

Definition 5. 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 $a \in \mathrm{argmax}_{a \in \mathcal{A}} A_1^\star(s,a)$.

An optimality preserving transformation of R_1 lets us pick an arbitrary new value function ψ, and then adjust R_2 in any way that respects the new value function and the argmax of $A_1^\star$ — the latter condition ensures that the same actions (and hence the same policies) stay optimal.

Based on these definitions, we can now 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. 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. Note also that $\mathrm{CS} \subseteq \mathrm{PS}_\gamma$ (for all γ), and that all sets are subsets of $\mathrm{OP}_{\tau,\gamma}$ (see Skalse et al. 2022a).

Two Equivalence Classes for Reward Functions

Our definition of misspecification robustness is given relative to an equivalence relation on $\mathcal{R}$. In this section, we define two important equivalence classes, and characterise the transformations that preserve them. Our later results will be given relative to these two 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] Note that if $R_1 \equiv_{\mathrm{ORD}^\mathcal{M}} R_2$ then $R_1 \equiv_{\mathrm{OPT}^\mathcal{M}} R_2$. Skalse et al. (2022a) showed that $R \equiv_{\mathrm{OPT}^\mathcal{M}} t(R)$ for all R if and only if $t \in \mathrm{OP}_{\tau,\gamma}$ (their Theorem 3.2). We characterise the transformations that preserve $\mathrm{ORD}^\mathcal{M}$, which is a novel contribution.

Theorem 6. *$R_1 \equiv_{\mathrm{ORD}^\mathcal{M}} R_2$ if and only if $R_2 = t(R_1)$ for some $t \in S'\mathrm{R}_\tau \circ \mathrm{PS}_\gamma \circ \mathrm{LS}$.*

Stated differently, Theorem 6 is saying that the MDPs $(\mathcal{S}, \mathcal{A}, \tau, \mu_0, R_1, \gamma)$ 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), positive linear scaling, and S'-redistribution (with τ), applied in any order.

$\mathrm{OPT}^\mathcal{M}$ and $\mathrm{ORD}^\mathcal{M}$ are two equivalence relations that should be relevant and informative in almost any context, which is why we have chosen to carry out our analysis in terms of these two relations. However, other partitions could be selected instead. For example, if we know that the learnt reward R_H will be used to compute a reward object f, then $\mathrm{Am}(f)$ would be a natural choice.

We now have results for reasoning about misspecification robustness in IRL. In particular, Lemma 2 tells us that if we want to find the functions that f is P-robust to misspecification with, then all we need to do is find the reward transformations that preserve P, and then compose them with f. $\mathrm{OPT}^\mathcal{M}$ and $\mathrm{ORD}^\mathcal{M}$ are reasonable choices of P, and the transformations that preserve them were just provided.

Misspecification Robustness of IRL Models

We here give our main results on the misspecification robustness of IRL, looking both at misspecification of the behavioural model, as well as of the MDP.

Misspecified Behavioural Models

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 is quite large, and includes e.g. Boltzmann-rational policies (for any β), but it does not include optimal policies (since they do not take all actions with positive probability) or causal entropy maximising policies (since they may take suboptimal actions with high probability).

Theorem 7. *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 7 exactly characterises the misspecification to which the Boltzmann-rational model is $\mathrm{OPT}^\mathcal{M}$-robust.

[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.

Let us briefly comment on the requirement that $\pi(a \mid s) > 0$, which corresponds to the condition that $\mathrm{Im}(g) \subseteq \mathrm{Im}(f)$ in our definition of misspecification robustness. If a learning algorithm $\mathcal{L}$ is based on a model $f : \mathcal{R} \to \Pi^+$ then it assumes that the observed policy takes each action with positive probability in every state. What happens if such an algorithm $\mathcal{L}$ is given data from a policy that takes some action with probability 0? This depends on $\mathcal{L}$, but for most sensible algorithms the result should simply be that $\mathcal{L}$ assumes that those actions are taken with a positive but low probability. This means that it should be possible to drop the requirement that $\pi(a \mid s) > 0$ for most reasonable algorithms.

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 8. *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 policy takes, and ignore the relative probability it assigns to those actions. Formally, we will treat optimal policies as functions $\pi_\star : \mathcal{S} \to \mathcal{P}(\mathrm{argmax}_{a \in \mathcal{A}} A^\star) - \{\varnothing\}$; 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 9. *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 10. *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 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 $t \in \mathrm{OP}_{\tau,\gamma}$. 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.

Misspecified MDPs

A reward object can be parameterised by a γ or τ, implicitly or explicitly. For example, the reward objects in Section 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 7-10 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 11. *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 12. *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. From these lemmas, together with Lemma 1, we get the following result:

Theorem 13. *If $f^{\tau_1} = f^{\tau_1} \circ t$ for all $t \in S'\mathrm{R}_{\tau_1}$ and $f^{\tau_2} = f^{\tau_2} \circ t$ for all $t \in S'\mathrm{R}_{\tau_2}$, 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 $t \in \mathrm{PS}_{\gamma_1}$ and $f^{\gamma_2} = f^{\gamma_2} \circ t$ for all $t \in \mathrm{PS}_{\gamma_2}$, 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.*

In other words, if a behavioural model 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, if the behavioural model 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 R_H has the same optimal policies as the true underlying reward 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.

Generalising the Analysis

In this section, we discuss different ways to generalise our results. We consider what happens if R is restricted to a subset of $\mathcal{R}$, what might happen if R is drawn from a known prior distribution and the learning algorithm has a known inductive bias, and whether we can use stronger equivalence classes to guarantee various forms of transfer learning.

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 $R_1, R_2 \in \hat{\mathcal{R}}$.

All lemmas in Section apply with these more general definitions for any arbitrary subset $\hat{\mathcal{R}} \subseteq \mathcal{R}$. Moreover, the theorems in Section also carry over very directly:

Theorem 14. *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 6. 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 our *negative* results (i.e., those in Section) 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 12, and the corresponding part of Theorem 13, 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.

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 in Section 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

[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 $t \in \mathrm{PS}_\gamma$.

such analysis as a topic for future work.

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 Section 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.

We would also like to remark on a subtlety regarding Theorem 6. One might expect that two reward functions R_1 and R_2 must have the same policy ordering for all τ if and only if they differ by potential shaping and linear scaling. However, this is not the case. To see this, consider the rewards R_1, R_2 where $R_1(s_1, a_1, s_1) = 1$, $R_1(s_1, a_1, s_2) = 0.5$, $R_2(s_1, a_1, s_1) = 0.5$, and $R_2(s_1, a_1, s_2) = 1$, and where R_1 and R_2 are 0 for all other transitions. Now R_1 and R_2 do not differ by potential shaping and linear scaling, yet they have the same policy order for all τ.

Discussion

In this section, we discuss the implications of our results, as well as their limitations.

Conclusions and Implications

We have shown that the misspecification robustness of the behavioural models in IRL can be quantified and understood. Our results show that the Boltzmann-rational model is substantially more robust to misspecification than the optimality model; the optimality model is not robust to any misspecification, whereas the Boltzmann-rationality model is at least $\mathrm{OPT}^{\mathcal{M}}$-robust to many kinds of misspecification. This is not necessarily unexpected, but we now have formal guarantees to back this intuition. We have also quantified the misspecification robustness of the maximal causal entropy model, and found that it lies somewhere between that of the Boltzmann-rational model and the optimality model.

We have shown that none of the standard models are robust to a misspecified τ or γ. Moreover, we need to make very minimal assumptions about how the demonstrator policy is computed to obtain this result, which means that it is likely to generalise to new behavioural models as well. We find this quite surprising; the discount γ is typically selected in a somewhat arbitrary way, and it can often be difficult to establish post-facto which γ was used to compute a given policy. The fact that τ must be specified correctly is somewhat less surprising, yet important to have established.

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 . In particular, if we have a model f, and we wish to use the learnt reward to compute an object g, then we can obtain an expression of the set of all functions to which f is robust in the following way; first, derive $\mathrm{Am}(g)$, and then characterise this partition of $\mathcal{R}$ using a set of reward transformations T. Then, as per Lemma 2, we can obtain the functions that f is robust to misspecification with by simply composing f with each $t \in T$. If we want to know which functions f is robust to misspecification with in a strong sense, then we can obtain an informative answer to this question by composing f with the transformations that preserve the ordering of all policies, which in turn is provided by Theorem 6. Lemma 1 also makes it easier to intuitively reason about the robustness properties of various kinds of behavioural models.

Limitations and Further Work

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 $\mathcal{R}$, where two reward functions are either equivalent or not. It might be fruitful to instead consider *distance metrics* on $\mathcal{R}$: this could make it possible to obtain results such as e.g. bounds on the distance between the true reward function and the learnt reward function, given various forms of misspecification. We believe it would be especially interesting to re-examine Theorem 13 through this lens.

Another notable direction for extensions could be to further develop the analysis in Section , and study the misspecification robustness of different behavioural models in the context where we have particular, known priors concerning R. Our comments on this setting are fairly preliminary, and it might be possible to draw additional, interesting conclusions if this setting is explored more extensively.

Moreover, we have studied the behaviour of algorithms in the limit of infinite data, under the assumption that this is similar to their behaviour in the case of finite but sufficiently large amounts of data. Therefore, another possible extension could be to more rigorously examine the properties of these models in the case of finite data.

Finally, our analysis has of course been limited to the behavioural models that are currently most popular in IRL (optimality, Boltzmann rationality, and causal entropy maximisation) and two particular equivalence relations ($\mathrm{OPT}^{\mathcal{M}}$ and $\mathrm{ORD}^{\mathcal{M}}$). Another direction for extensions would be to broaden our analysis to larger classes of models, and perhaps also to more equivalence relations. In particular, it would be interesting to analyse more realistic behavioural models, which incorporate e.g. prospect theory (Kahneman and Tversky 1979) or hyperbolic discounting.

References

Abbeel, P.; Coates, A.; and Ng, A. Y. 2010. Autonomous Helicopter Aerobatics Through Apprenticeship Learning. *The*

International Journal of Robotics Research, 29(13): 1608–1639.

Armstrong, S.; and Mindermann, S. 2019. Occam's razor is insufficient to infer the preferences of irrational agents. arXiv:1712.05812.

Cao, H.; Cohen, S. N.; and Szpruch, L. 2021. Identifiability in Inverse Reinforcement Learning. *arXiv preprint*, arXiv:2106.03498 [cs.LG].

Chan, L.; Critch, A.; and Dragan, A. 2021. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*.

Dvijotham, K.; and Todorov, E. 2010. Inverse Optimal Control with Linearly-Solvable MDPs. In *Proceedings of the 27th International Conference on Machine Learning*, 335–342. Haifa, Israel: Omnipress, Madison, Wisconsin, USA.

Evans, O.; Stuhlmueller, A.; and Goodman, N. D. 2015. Learning the Preferences of Ignorant, Inconsistent Agents. arXiv:1512.05832.

Freedman, R.; Shah, R.; and Dragan, A. 2020. Choice Set Misspecification in Reward Inference. In *IJCAI-PRICAI-20 Workshop on Artificial Intelligence Safety*.

Gleave, A.; Dennis, M. D.; Legg, S.; Russell, S.; and Leike, J. 2021. Quantifying Differences in Reward Functions. In *International Conference on Learning Representations*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative Inverse Reinforcement Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.*, 50(2).

Jenner, E.; van Hoof, H.; and Gleave, A. 2022. Calculus on MDPs: Potential Shaping as a Gradient. *arXiv preprint arXiv:2208.09570*.

Jeon, H. J.; Milli, S.; and Dragan, A. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 4415–4426. Curran Associates, Inc.

Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–291.

Kim, K.; Garg, S.; Shiragur, K.; and Ermon, S. 2021. Reward Identification in Inverse Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5496–5505. Virtual: PMLR.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 278–287. Bled, Slovenia: Morgan Kaufmann Publishers Inc.

Ng, A. Y.; and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, 663–670. Stanford, California, USA: Morgan Kaufmann Publishers Inc.

Orsini, M.; Raichuk, A.; Hussenot, L.; Vincent, D.; Dadashi, R.; Girgin, S.; Geist, M.; Bachem, O.; Pietquin, O.; and Andrychowicz, M. 2021. What Matters for Adversarial Imitation Learning? *arXiv preprint*, arXiv:2106.00672 [cs.LG]. To appear in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

Pan, A.; Bhatia, K.; and Steinhardt, J. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*.

Ramachandran, D.; and Amir, E. 2007. Bayesian Inverse Reinforcement Learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2586–2591. Hyderabad, India: Morgan Kaufmann Publishers Inc.

Rothkopf, C. A.; and Dimitrakakis, C. 2011. Preference Elicitation and Inverse Reinforcement Learning. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011, Proceedings, Part III*, volume 6913 of *Lecture Notes in Computer Science*, 34–48. Athens, Greece: Springer.

Shah, R.; Gundotra, N.; Abbeel, P.; and Dragan, A. D. 2019. On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference. arXiv:1906.09624.

Singh, A.; Yang, L.; Hartikainen, K.; Finn, C.; and Levine, S. 2019. End-to-End Robotic Reinforcement Learning Without Reward Engineering. In *Proceedings of Robotics: Science and Systems*. Freiburg im Breisgau, Germany.

Skalse, J.; Farrugia-Roberts, M.; Russell, S.; Abate, A.; and Gleave, A. 2022a. Invariance in Policy Optimisation and Partial Identifiability in Reward Learning. *arXiv preprint arXiv:2203.07475*.

Skalse, J.; Howe, N.; Dima, K.; and Krueger, D. 2022b. Defining and Characterizing Reward Hacking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, second edition. ISBN 9780262352703.

White, H. 1994. *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge University Press.

Yamaguchi, S.; Naoki, H.; Ikeda, M.; Tsukada, Y.; Nakano, S.; Mori, I.; and Ishii, S. 2018. Identification of animal behavioral strategies by inverse reinforcement learning. *PLOS Computational Biology*, 14(5): 1–20.

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