Failure Modes of Learning Reward Models for LLMs and other Sequence Models

Silviu Pitis [1]

Abstract

To align large language models (LLMs) and other sequence-based models with human values, we typically assume that human preferences can be well represented using a "reward model". We infer the parameters of this reward model from data, and then train our models to maximize reward. Effective alignment with this approach relies on a strong reward model, and reward modeling becomes increasingly important as the dominion of deployed models grows. Yet in practice, we often assume the existence of a particular reward model, without regard to its potential shortcomings. In this preliminary work, I survey several failure modes of learned reward models, which may be organized into three broad categories: model misspecification, ambiguous preferences, and reward misgeneralization. Several avenues for future work are identified. It is likely that I have missed several points and related works; to that end, I greatly appreciate your correspondence.

1. Introduction

The alignment of large language models (LLMs) (Bommasani et al., 2021; Brown et al., 2020; OpenAI, 2023) with human needs and values typically involves the use of a "reward model" to perform reinforcement learning (RL) from human feedback (RLHF) (Christiano et al., 2017; Leike et al., 2018). Although this has proven effective in training language models that follow instruction and behave as helpful and harmless assistants (Ouyang et al., 2022; Bai et al., 2022), numerous researchers have commented on certain weaknesses of reward models, and the challenges involved in using them to accurately represent human preferences (Armstrong & Mindermann, 2018b; Pitis, 2019; Abel et al., 2021; Ziegler et al., 2019; Freedman et al., 2021; Skalse & Abate, 2022; Knox et al., 2022; Tien et al., 2023). This preliminary work surveys certain shortcomings of learned reward models that have been identified by the author and others. They may be broadly categorized as follows:

Model misspecification: one or more models in the reward modeling process is misspecified, leading to inaccurate reward inference.

Ambiguous preferences: the modeled preferences depend on exogenous variables that are unknown or may change between reward inference and policy deployment.

Reward misgeneralization (unidentifiability): the empirical data is insufficient to determine the reward in such a way that it generalizes out-of-distribution.

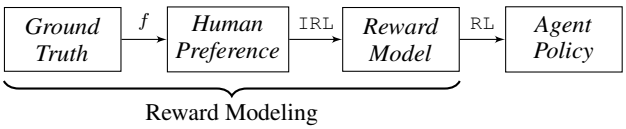
Many of the failure modes identified are hard problems. But recognition is the first step toward finding solutions and implementing adequate protections. We should strive to find solutions to these problems before strong AI systems are granted sufficient agency to pose a real threat to society.

I will start by outlining a general framework for reward modeling in Section 2, and then proceed in Section 3 to outline the several failure modes.

2. Reward Models

2.1. A birds-eye view of the reward modeling process

I propose the following high level framework:



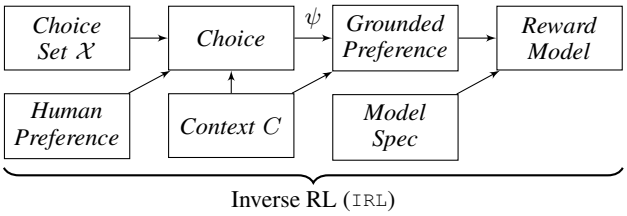
Here, the ground truth consists of the preferences that a human principal with omniscience would possess regarding the agent's policy. We can understand the ground truth as "objective" preferences. A noisy function f maps them to the subjective (or in context) human preferences. The human preferences are then used to train a reward model using some reward learning algorithm `IRL` (for Inverse RL, but see below). Finally, the learned reward model is used to train an agent policy using algorithm `RL`, which is what we typically think of as the RL problem. For the most part,

[1]University of Toronto. Vector Institute. Toronto, Canada. Correspondence to: <spitis@cs.toronto.edu>.

The Many Facets of Preference-based Learning, Workshop at the International Conference on Machine Learning (ICML) 2023.

RL is a solved problem, or at least a highly attended to one: our algorithms work, sometimes too well; cf. reward hacking (Pan et al., 2022; Skalse et al., 2022). This paper is concerned only with the accuracy of the reward model and not the final agent policy; i.e., only with the mapping $\texttt{IRL} \circ f$ from ground truth preferences to reward model.

Traditionally, Inverse RL maps from policy to reward, but I will lean on the Reward Rational Implicit Choice framework (Jeon et al., 2020) and consider the broader view of mapping arbitrary (possibly implicit) choices to rewards:



This figure decomposes the IRL arrow from the first figure into three steps. First, given some context C, human preference is expressed by choosing one or more options from a choice set $\mathcal{X}$, which may be a set of trajectories, natural language utterances, demonstrations, or any other modality that provides signal about human preference (see Jeon et al. (2020) for other possibilities). Second, the human choice is grounded into a preference over trajectories using grounding function $\psi : x \mapsto \tilde{\tau}$, where $x \in \mathcal{X}$, τ is a trajectory, and ~ marks a distribution so that $\tilde{\tau}$ is a distribution over trajectories. Finally, the parameters of a reward model are optimized so as to represent the grounded preferences.

2.2. Three comments on reward modeling

C1. Rewards compress utility

Perhaps due to the historical prevalence of the Markov assumption (Sutton & Barto, 2018), our field focuses a lot on the notion of "reward" and very little on the notion of "utility". Abbeel & Ng (2004) offered this strong statement:

> [T]he entire field of RL is founded on the presupposition that the reward function, rather than [preferences or utility], is the most succinct, robust, and transferable *definition* of [a] task. *(emphasis mine)* (1)

The reward function maps states and actions to a scalar signal, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, allowing value (a model of utility) to be determined as the cumulative sum of future rewards (Bowling et al., 2022). This depends critically on the assumption that the state is Markovian: it captures all information necessary to determine future preference. This conditional independence assumption allows us to determine values for a very large space (all possible futures) using a relatively small space ($\mathcal{S} \times \mathcal{A}$) (cf. Pitis et al. (2022)), thereby offering significant *compression* of utility, and enabling efficient dynamic programming algorithms (Paster et al., 2022).

We should be wary, however, of treating rewards as primitive objects (i.e., as the *definition* of a task, as in (1)), simply because most work in RL begins by assuming the existence of a Markovian reward function. Reward specification is difficult (Shah et al., 2022) because individual rewards, on their own, carry little meaning. I argue that step-wise rewards are best understood as a difference of utilities (Pitis, 2019), but this only applies under certain assumptions that may not always be tenable (Pitis, 2023). For example, ff the tasks we care about are not Markovian in the state, Markovian rewards are no longer sufficient (Abel et al., 2021).

There has been a trend in recent work toward defining a trajectory-wise reward function/model $R : \tau \mapsto r$ (Ziegler et al., 2019; Jeon et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022). Although this diverges from the traditional step-wise definition, I believe this is wholly appropriate under a broader view of reward, not as a step-wise function, but as a "view" or abstraction of utility. What I mean by this will become more clear below, particularly in Subsection 3.2, but the point I want to make here is that utility can be decomposed into several reward components. These components may be causally related in various ways (Barreto et al., 2017; Haarnoja et al., 2018; Icarte et al., 2018; Colas et al., 2019). The accuracy of this decomposition, together with the causal independencies it leverages, determines the succinctness, robustness, and transferability of rewards (1).

C2. Utility represents in-context, rational preference

Utilities (hence rewards) are numerical representations of a *rational* preference relation ($\succ$ / $\succeq$) over some set of alternatives $\mathcal{O}$ given some *context* C.

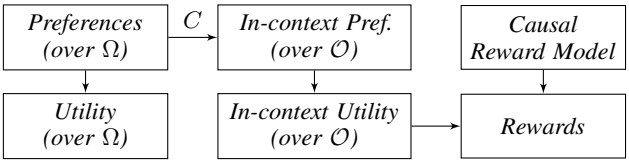
Rational can mean different things, but the weakest notions of rationality typically require preferences to be asymmetric ($x \succ y \Rightarrow y \not\succ x$) and transitive ($x \succeq y, y \succeq z \Rightarrow x \succeq z$). Together, these imply the existence of an ordinal utility representation, $U : \mathcal{O} \to \mathbb{R}$, where $U(x) > U(y) \Leftrightarrow x \succ y$. With some additional consistency assumptions about rationality, we can obtain a cardinal utility function, where utility differences are meaningful (Pitis, 2019).

Given omniscience (infinite compute / time for consideration), we imagine that preferences could be defined in a context-free manner, over the space of exhaustively specified alternatives, Ω. This is, of course, impractical, and so preferences are usually restricted to a more manageable space $\mathcal{O}$, with everything left unspecified in some context C, so that $\Omega = \mathcal{O} \times C$. Sometimes we extract an additional conditioning variable from the context, $g \in \mathcal{G}$, which represents a particular goal, task or skill, and may be specified via a partial trajectory (prompt), explicit goal (instruction), or latent conditioning variable (adapter), so that $\Omega = \mathcal{O} \times \mathcal{G} \times C$.

Further to comment C1, we can obtain rewards over some sub-components of $\mathcal{O}$ (e.g. state-action (s, a) is a com-

ponent of trajectory τ) by decomposing utility using a set of causal mechanisms (reward components). To achieve compression, we exploit (possibly local) conditional independencies between reward components (Pitis et al., 2020).

We can summarize this as follows:



C3. Preferences may be represented numerically in many ways, one way, or none at all

Let $\Omega = \{x, y\}$, and consider the preference relation where $x \succ y$. Then any pair of real numbers $a, b \in \mathbb{R}$ with $a > b$ is a valid utility representation of $\succ$ so long as $a > b$. Alternatively, let $\Omega = \{x, y, z\}$ and consider the preference relation where $x \succ y, y \succ z, z \succ x$. Then no triplet of real numbers $a, b, c \in \mathbb{R}$ is a valid utility representation of $\succ$.

While simplistic, these exemplify two general solution classes to the preference representation problem that we must be wary of. Since different rewards may yield equivalent preferences, we should be careful about the statements we make regarding rewards, as well as how we integrate different information sources. See Skalse et al. (2023).

We should be particularly wary of cases where rational preferences fail to possess a reward representation. For instance, the composition of in-context preferences with respect to two objectives, each defined with respect to the same alternatives $\mathcal{O}$ and context $\mathcal{C}$, may produce preferences that appear irrational with respect to $\mathcal{O}$, even though they are rational with respect to some expanded $\mathcal{O}' \supset \mathcal{O}$ (Pitis, 2023).

3. Failure Modes of Reward Models

Having contextualized the meaning and purpose of rewards and reward modeling in the previous Section, let us now enumerate several potential failures that may arise.

3.1. Model misspecification

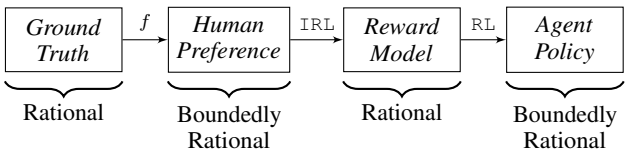
At each step of the reward modeling process presented in Subsection 2.1, there are one or more models involved. A misspecification of any of these models gives rise to potentially bad outcomes. Let us examine in each turn.

M1. Misspecified human preferences and choice rule

In C2-C3, we assumed that utility (and hence rewards) should be "rational," which humans are evidently not: our preferences and choice behavior deviate, often systematically, from what rationality would prescribe (Simon, 1972; Tversky & Kahneman, 1986). Accurate reward inference requires both prescriptive and descriptive modeling. We would like our agents to to correct for human failures in rationality, which we can do by prescribing a rational reward model. But to infer its parameters, we need a descriptive model of how human choice relates to the underlying "ground truth" preferences we are trying to model; using the wrong model can lead to policies that perform no better than chance, or worse (Armstrong & Mindermann, 2018a; Laidlaw & Dragan, 2022; Skalse & Abate, 2022).

As a first approximation, which can help account for the empirical noise we observe in human preferences, we can assert that humans operate according to a "Boltzmann rationality" (Ramachandran & Amir, 2007) choice rule, but several recent works have suggested that richer models are needed to accurately capture human behavior and have explored ways of learning the human model (Shah et al., 2019; Knox et al., 2022; Hong et al., 2022; Ghosal et al., 2022).

To summarize, if we understand human choice behavior as a noisy representation of an underlying ground truth (Condorcet, 1785; Pitis & Zhang, 2020), an accurate model of the noisy channel is essential for aligning the learned reward with ground truth preferences:



M2. Misspecified choice set $\mathcal{X}$ or grounding function ψ

Freedman et al. (2021) consider the risks of making the incorrect assumption about the choice set $\mathcal{X}$ from which human choices are made during IRL. They identify several ways in which the choice set can be misaligned, and show that in the worst case—when the robot incorrectly assumes that $\mathcal{X}$ contains certainly alternatives that are actually preferred to the empirically observed choices—misspecification can lead to significant regret.

Recall that IRL applies grounding function ψ to map observed human choices to distributions over the set of alternatives $\mathcal{O}$. For example, if $a, b \in \mathcal{X}$, and a choice of a is observed, we assume that the human assigns higher utility to the distribution $\psi(a)$ of alternatives in $\mathcal{O}$ than the distribution $\psi(b)$. To my present knowledge, this type of misspecification has not been explored in the literature.

M3. Misspecified reward model

Even if all other models are correct, and the reward model is well optimized during IRL, an underexpressive reward model can lead to incorrect inferences about preferences. The most commonly used form of reward model assumes

a fixed discount factor γ and only infers the parameters of function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (Ng & Russell, 2000; Christiano et al., 2017). Christiano et al. (2017) (footnote 3) notes that it would be reasonable to adopt a reward model that uses explicit or inferred discounting, but to my present knowledge, nobody actually does this (the closest within RL is perhaps Schultheis et al. (2022)). But prior work has shown that a reward model using fixed γ cannot represent all rational preferences for a reasonably specific definition of rationality (White, 2017; Pitis, 2019). Pitis (2019) shows that although `IRL` with fixed γ can recover a reward function that correctly implies the optimal policy when the `RL` process is perfect, it may incorrectly predict preferences with respect to suboptimal policies. This may be problematic, as in practical problems, only suboptimal policies are achievable.

A closely related modeling failure has to do with feature selection for reward models. Non-Markovian rewards are the prime example of this: if the features do not contain all relevant historical information, so that the true reward function is non-Markovian with respect to state features, `IRL` will obviously fail to produce a strong reward function (Abel et al., 2021). If LLM reward models condition on a long enough context, this is not likely to be a large problem.

3.2. Ambiguous preferences

As noted in C2, asymmetry ($x \succ y \Rightarrow y \not\succ x$) is a basic requirements for preference $\succ$. Yet, particularly when the space of alternatives $\mathcal{O}$ is underspecified, $\succ$ may depend critically on certain unstable elements of the context C—if C changes, we may observe preference reversal (see Pitis (2019), Section 4.2). Humans providing preference feedback may be so uncertain about C (they might ask, "should we assume c when choosing?") that they cannot express a clear preference one way or the other (see Ziegler et al. (2019), Section 4.3). Below are three ways this type of ambiguity might arise, even if all models are correctly specified.

A1. Future value

Christiano et al. (2017) and derivative works (Stiennon et al., 2020; Ziegler et al., 2019; Ouyang et al., 2022) use trajectory *segments* as the objects of preferences $\mathcal{O}$, leaving future behaviors to the context C. Even if their model accurate reflects preferences about trajectory segments in-distribution, a change in assumed background policy or dynamics could change the future value of actions within the observed trajectory segments. For example, if it assumed that an LLM will guard private information about the user outside a given trajectory segment, then segments where private information is obtained will not be penalized. If this background assumption changes, so that private information collected during a trajectory segment could be leaked in the future, such segments should be less preferred. To account for this, the assumptions made by RLHF—or RLAIF (Bai et al., 2022)—feedback providers should be informed or elucidated, and the reward model should contain some information about the future; e.g., a regret-based model (Knox et al., 2022).

A2. Resolved stochasticity and counterfactual outcomes

As noted in Pitis (2019) (Section 4.2), to obtain asymmetric preferences over trajectories, we must make assumptions about how we treat resolved stochasticity, which is why in Pitis (2019) I argue that preferences should be expressed with respect to state-policy pairs, which do not involved any unresolved stochasticity (cf. Kreps & Porteus (1978)). As an example, suppose we are expressing preferences over Texas Hold'em trajectories. Standard reward modeling is strictly outcome based, and would prefer a trajectory where 72o calls a significant bet preflop and wins, to one where AA makes a significant bet preflop and loses. But very rational humans (myself included) would express preference for the latter in absence of an explicit instruction to ignore counterfactual outcomes. While providing human feedback on trajectories involving LLM tool usage, the author has observed similar preference ambiguity due to resolved stochasticity. Should trajectories where an the LLM takes a risky action that ends up having the correct outcomes be preferred? Or should the potentially bad counterfactual outcomes be included? These questions have not been explored, and it is unclear what impact this has on reward modeling.

A3. Base rates and background assumptions

More broadly, any shift in base rates or background assumptions can cause preference reversal. As a straightforward example: given prompt p, Alice may prefer LLM completion a to LLM completion b, whereas Bob may prefer LLM completion b to LLM completion a. Thus, a reward model learned from Bob may not be suitable for Alice. But similar differences may arise from subtler differences in context, such as an LLM's capability. For example, in context of a debate, we might prefer an LLM to give highly opinionated responses, even if that is not the general preference for a chat model. To combat this problem, we should aim to give human annotators as much instruction as possible in order to disambiguate the context C, and consider collecting annotations for a variety of different tasks from $\mathcal{G}$ (see C2). The latter may be necessary to determine the causal mechanisms responsible for human preference (Arjovsky et al., 2019).

3.3. Reward misgeneralization (unidentifiability

Even if all models are correctly specified, and the context C is held fixed, so that a policy optimizing a learned reward model correctly represents the ground truth preference on the data distribution it was trained on, we may still find that learned rewards generalize poorly when that data distribution changes. This is known as reward misgeneralization

(Di Langosco et al., 2022), and may arise for a few reasons.

U1. Equivalence Classes of Reward Functions

As noted in C3, there may be multiple solutions to the reward inference problem, whether we are adopting a deterministic choice rule (Ng & Russell, 2000) or, as is the preferred approach in Inverse RL, a stochastic one (Fu et al., 2018). This is referred to as the unidentifiability problem in reward modeling (Kim et al., 2021). Unfortunately, while all solutions may be valid for purposes of imitating behavior in-distribution, they may not generalize well out-of-distribution (Tien et al., 2023). Cao et al. (2021) provide an analysis of this problem and show that under certain conditions, observing two experts acting under different environment dynamics or discount factors is sufficient for identification. This solution might be applied to language model reward inference by using the system prompt to change how concise the LLM's responses are (changing γ) or by finetuning a chat model to two different users (changing dynamics).

U2. Limitations of Training Data

Another failure mode is caused by imbalances in or limited coverage of the training data used to learn the reward model. When a reward model is trained on data from a specific distribution, it may perform poorly when exposed to examples that are out-of-distribution (OOD). This effect is most visible in RLHF-trained language models that do not have sufficient KL regularization, so that RLHF drives them OOD (Ziegler et al., 2019). It has also been observed in the traditional RL setting, where reward functions learned via RLHF that are useful for training the current online agent fail to be useful for training a new agent (McKinney et al., 2022). This failure mode is well recognized and has at least two solutions that are currently used in practice: the KL regularization mentioned above (Gaon & Brafman, 2020), and online RLHF training / iterative development and deployment (Leike et al., 2018; Brundage et al., 2022).

U3. Consistency and Strength of Preference

When we are learning certain reward components independently, we may incorrectly gauge the relative strength of preference, leading to poor predictions about preferences between their combinations, *even if the components are independent*. Although consistency correlates with strength of preference (Alós-Ferrer & Garagnani, 2021), strength of preference cannot be determined from consistency alone. Consider the following example. It is perfectly to rational to (a) strongly prefer steak to salad, yet only choose to eat steak 2/3 of the time, and (b) weakly prefer watching TV before doing the dishes to doing the dishes before watching TV, yet consistently choose to watch TV before doing the dishes. But if standard RLHF is only presented with comparisons (a) and (b) separately, then it will learn strong preference for (b) and weak preference for (a), since it equates consistency with strength. But suppose we concatenate the trajectory segments and compare the combination (steak, dishes then TV) to (salad, TV then dishes). It is natural to think that (steak, dishes then TV) will preferred approximately 2/3 of the time, since the strong preference dominates. But standard RLHF (Christiano et al., 2017) would have us predict that (salad, TV and dishes) is consistently preferred. To resolve this, we need not only a more expressive model (M1), but also a data distribution that allows us to infer the relative strength of preference, by including bundles like (steak, dishes then TV) in the comparison queries.

4. Conclusion

In this survey, I have provided an opinionated view of reward modeling, and outlined several ways in which it might fail. Although some of these failure modes have received, and are continuing to receive, significant attention (M1, U1, U2), others have received little to no attention (M3, A1-A3, U3). Not all failure modes are equally problematic. For instance, I would not consider a misspecified reward *model* (M3) to be a big problem for LLMs, because the long context length allows for reward functions that are highly history dependent (of course, this works against C1, and may compromise generalizability). But this view may change, depending on future approaches to LLM rewards. On the other hand, I consider A1-A3 to be very important problems. Although ambiguity of preference has been noted as a challenge of RLHF (Ziegler et al., 2019; Gao et al., 2022), to my knowledge little has been done to address it. Despite the attention it has received, primarily from Anca Dragan's group, I believe much more work is needed with regards to descriptive modeling of human behavior (M1-M2). As noted in the abstract, this work is largely preliminary, and I welcome any feedback or correspondence from interested parties.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *The Twenty-first International Conference on Machine learning*, pp. 1. ACM, 2004.

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., and Singh, S. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021.

Alós-Ferrer, C. and Garagnani, M. Choice consistency and strength of preference. *Economics Letters*, 198:109672, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint*

arXiv:1907.02893, 2019.

Armstrong, S. and Mindermann, S. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, 2018a.

Armstrong, S. and Mindermann, S. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in neural information processing systems*, 31, 2018b.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bowling, M., Martin, J. D., Abel, D., and Dabney, W. Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Brundage, M., Mayer, K., Eloundou, T., Agarwal, S., Adler, S., Krueger, G., Leike, J., and Mishkin, P. Lessons learned on language model safety and misuse, 2022.

Cao, H., Cohen, S., and Szpruch, L. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Colas, C., Fournier, P., Chetouani, M., Sigaud, O., and Oudeyer, P.-Y. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning*, pp. 1331–1340. PMLR, 2019.

Condorcet, J. A. M. N. C. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, volume 252. 1785.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Freedman, R., Shah, R., and Dragan, A. Choice set misspecification in reward inference. *arXiv preprint arXiv:2101.07691*, 2021.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

Gaon, M. and Brafman, R. Reinforcement learning with non-markovian rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3980–3987, 2020.

Ghosal, G. R., Zurek, M., Brown, D. S., and Dragan, A. D. The effect of modeling human rationality level on learning rewards from multiple feedback types. *arXiv preprint arXiv:2208.10687*, 2022.

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., and Levine, S. Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6244–6251. IEEE, 2018.

Hong, J., Bhatia, K., and Dragan, A. On the sensitivity of reward inference to misspecified human models. *arXiv preprint arXiv:2212.04717*, 2022.

Icarte, R. T., Klassen, T., Valenzano, R., and McIlraith, S. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pp. 2107–2116. PMLR, 2018.

Jeon, H. J., Milli, S., and Dragan, A. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33: 4415–4426, 2020.

Kim, K., Garg, S., Shiragur, K., and Ermon, S. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 5496–5505. PMLR, 2021.

Knox, W. B., Hatgis-Kessell, S., Booth, S., Niekum, S., Stone, P., and Allievi, A. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*, 2022.

Kreps, D. M. and Porteus, E. L. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica:*

journal of the Econometric Society, pp. 185–200, 1978.

Laidlaw, C. and Dragan, A. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. *arXiv preprint arXiv:2204.10759*, 2022.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

McKinney, L. E., Duan, Y., Krueger, D., and Gleave, A. On the fragility of learned reward functions. In *Deep Reinforcement Learning Workshop at NeurIPS 2022*, 2022.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *The Seventeenth International Conference on Machine Learning*, pp. 663–670, 2000.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

Paster, K., Pitis, S., McIlraith, S. A., and Ba, J. Return augmentation gives supervised rl temporal compositionality. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

Pitis, S. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7949–7956, 2019.

Pitis, S. Multi-objective agency requires non-markovian rewards. *Under Review*, 2023.

Pitis, S. and Zhang, M. R. Objective social choice: Using auxiliary information to improve voting outcomes. *arXiv preprint arXiv:2001.10092*, 2020.

Pitis, S., Creager, E., and Garg, A. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 2020.

Pitis, S., Creager, E., Mandlekar, A., and Garg, A. Mocoda: Model-based counterfactual data augmentation. In *Advances in Neural Information Processing Systems*, 2022.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *IJCAI*, 2007.

Schultheis, M., Rothkopf, C. A., and Koeppl, H. Reinforcement learning with non-exponential discounting. In *Advances in neural information processing systems*, 2022.

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, pp. 5670–5679. PMLR, 2019.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Simon, H. A. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

Skalse, J. and Abate, A. Misspecification in inverse reinforcement learning. *arXiv preprint arXiv:2212.03201*, 2022.

Skalse, J. M. V., Howe, N. H., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, 2022.

Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., and Gleave, A. Invariance in policy optimisation and partial identifiability in reward learning. 2023.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Tien, J., He, J. Z.-Y., Erickson, Z., Dragan, A., and Brown, D. S. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Tversky, A. and Kahneman, D. Rational choice and the framing of decisions. *Journal of business*, pp. S251–S278, 1986.

White, M. Unifying task specification in reinforcement learning. In *The Thirty-fourth International Conference on Machine Learning*, 2017.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.