IRRATIONALITY CAN HELP REWARD INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Specifying reward functions is difficult, which motivates the area of reward inference: learning rewards from human behavior. The starting assumption in the area is that human behavior is optimal given the desired reward function, but in reality people have many different forms of irrationality, from noise to myopia to risk aversion and beyond. This fact seems like it will be strictly harmful to reward inference: it is already hard to infer the reward from rational behavior, and noise and systematic biases make actions have less direct of a relationship with the reward. Our insight in this work is that, contrary to expectations, irrationality can actually help rather than hinder reward inference. For some types and amounts of irrationality, the expert now produces more varied policies compared to rational behavior, which help disambiguate among different reward parameters – those that otherwise correspond to the same rational behavior. We put this to the test in a systematic analysis of the effect of irrationality on reward inference. We start by covering the space of irrationalities as deviations from the Bellman update, simulate expert behavior, and measure the accuracy of inference to contrast the different types and study the gains and losses. We provide a mutual information-based analysis of our findings, and wrap up by discussing the need to accurately model irrationality, as well as to what extent we might expect (or be able to train) real people to exhibit helpful irrationalities when teaching rewards to learners.

1 INTRODUCTION

The application of reinforcement learning (RL) in increasingly complex environments has been most successful for problems that are already represented by a specified reward function (Lillicrap et al., 2015; Mnih et al., 2015; 2016; Silver et al., 2016). Unfortunately, not only do real-world tasks usually lack an explicit exogenously-specified reward function, but attempting to specify one tends to lead to unexpected side-effects as the agent is faced with new situations (Lehman et al., 2018).

This has motivated the area of *reward inference*: the process of estimating a reward function from human inputs. The inputs are traditionally demonstrations, leading to inverse reinforcement learning (IRL) (Ng et al., 2000; Abbeel & Ng, 2004) or inverse optimal control (IOC) (Kalman, 1964; Jameson & Kreindler, 1973; Mombaur et al., 2010; Finn et al., 2016). Recent work has expanded the range of inputs significantly,to comparisons (Wirth et al., 2017; Sadigh et al., 2017; Christiano et al., 2017), natural language instructions (MacGlashan et al., 2015; Fu et al., 2019), physical corrections (Jain et al., 2015; Bajcsy et al., 2017), proxy rewards (Hadfield-Menell et al., 2017; Ratner et al., 2018), or scalar reward values (Griffith et al., 2013; Loftin et al., 2014).

The central assumption behind these methods is that human behavior is rational, i.e. optimal with respect to the desired reward (cumulative, in expectation). Unfortunately, decades of research in behavioral economics and cognitive science Chipman (2014) has unearthed a deluge of *irrationalities*, i.e. of ways in which people deviate from optimal decision making: hyperbolic discounting, scope insensitivity, optimism bias, decision noise, certainty effects, loss aversion, status quo bias, etc.

Work on reward inference has predominantly used one model of irrationality: decision-making noise, where the probability of an action relates to the value that action has. The most widely used model by far is a Bolzmann distribution stemming from the Luce-Sherpard rule (Luce, 1959; Shepard, 1957; Lucas et al., 2009) and the principle of maximum (causal) entropy in (Ziebart et al., 2008; 2010), which we will refer to as *Bolzmann-rationality* (Fisac et al., 2017). Recent work has started to incorporate systematic biases though, like risk-aversion (Singh et al., 2017), having the wrong dynamics belief (Reddy et al., 2018), and myopia and hyperbolic discounting (Evans & Goodman, 2015; Evans et al., 2016).

Learning from irrational experts feels like daunting task: reward inference is already hard with rational behavior, but now a learner needs to make sense of behavior that is noisy or systematically biased. Our goal in this work is to characterize just how muddied the waters are – how (and how much) do different irrationalities affect reward inference?

> Our insight is that, contrary to expectations, irrationality can actually *help*, rather than hinder, reward inference.

Our explanation is that how good reward inference is depends on the mutual information between the policies produced by the expert and the reward parameters to be inferred. While it is often possible for two reward parameters to produce the same *rational* behavior, irrationalities can sometimes produce different behaviors that *disambiguate* between those same two reward parameters. For instance, noise can help when it is related to the value function, as Boltzmann noise is, because it distinguishes the difference in values even when the optimal action stays the same. Optimism can be helpful because the expert takes fewer risk-avoiding actions and acts more directly on their goal.

Overall, we contribute 1) an analysis and comparison of the effects of different biases on reward inference testing our insight, 2) a way to systematically formalize and cover the space of irrationalities in order to conduct such an analysis, and 3) evidence for the importance of assuming the right type of irrationality during inference.

Our good news is that irrationalities can indeed be an ally for inference. Of course, this is not always true – the details of which irrationality type and how much of it also matter. We see these results as opening the door to a better understanding of reward inference, as well as to practical ways of making inference easier by asking for the right kind of expert demonstrations – after all, in some cases it might be easier for people to act optimistically or myopically than to act rationally. Our results reinforce that optimal teaching is different from optimal doing, but point out that some forms of teaching might actually be easier than doing.

2 Method

2.1 Exploring irrationality through simulation

Our goal is to explore the effect irrationalities have on reward inference *if the learner knows about them* – we explore the need for the learner to accurately model irrationalities in section 4.2. While ideally we would recruit human subjects with different irrationalities and measure how well we can learn rewards, this is prohibitive because we do not get to dictate someone's irrationality type: people exhibit a mix of them, some yet to be discovered. Further, measuring accuracy of inference is complicated by the fact that we do not have ground truth access to the desired reward: the learner can measure agreement with some test set, but the test set itself is produced subject to the same irrationalities that produced the training data. As experimenters, we would remain deluded about the human's true intentions and preferences.

To address this issue, we *simulate* expert behavior subject to different irrationalities based on ground truth reward functions, run reward inference, and measure the performance against the ground truth, i.e. the accuracy of a Bayesian posterior on the reward function given the (simulated) expert's inputs.

2.2 Types and degrees of irrationality

There are many possible irrationalities that people exhibit (Chipman, 2014), far more than what we could study in one paper. They come with varying degrees of mathematical formalization and replication across human studies. To provide good coverage of this space, we start from the Bellman update, and systematically manipulate its terms and operators to produce a variety of different irrationalities that deviate from the optimal MDP policy in complementary ways. For instance, operating on the discount factor can model more myopic behavior, while operating on the transition function can model optimism or the illusion of control. Figure 1 summarizes our approach, which we detail below.

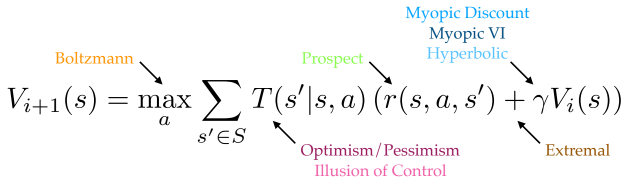


Figure 1: We modify the components of the Bellman update to cover different types of irrationalities: changing the max into a softmax to capture noise, changing the transition function to capture optimism/pessimism or the illusion of control, changing the reward values to capture the nonlinear perception of gains and losses (prospect theory), changing the average reward over time into a maximum (extremal), and changing the discounting to capture more myopic decision-making.

2.2.1 Rational Expert

The **rational** expert does value iteration using the Bellman update from figure 1. Our models change this update to produce different types of non-rational behavior.

2.2.2 Modifying the Max Operator: Bolzmann

Boltzmann-rationality modifies the maximum over actions $\max_a$ with a Boltzmann operator with parameter β:

$$V_{i+1}(s) = \text{Boltz}_a^\beta \sum_{s' \in S} T(s'|s,a)\,(r(s,a,s') + \gamma V_i(s))$$

Where $\text{Boltz}^\beta(\mathbf{x}) = \sum_i x_i e^{\beta x_i} / \sum_i e^{\beta x_i}$ (Ziebart et al., 2010; Asadi & Littman, 2017) This models that people will not be perfect, but rather noisily pick actions in a way that is related to the Q-value of those actions. The constant β is called the *rationality constant*, because as $\beta \to \infty$, the human choices approach perfect rationality (optimality), whereas $\beta = 0$ produces uniformly random choices. This is the standard assumption for reward inference that does not assume perfect rationality, because it easily transforms the rationality assumption into a probability distribution over actions, enabling learners to make sense of imperfect demonstrations that otherwise do not match up with any reward parameters.

2.2.3 Modifying the Transition Function

Our next set of irrationalities manipulate the transition function away from reality.

Illusion of Control. Humans often overestimate their ability to control random events. To model this, we consider experts that use the Bellman update:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} T^n(s'|s,a)\,(r(s,a,s') + \gamma V_i(s))$$

where $T^n(s'|s,a) \propto (T(s'|s,a))^n$. As $n \to \infty$, the demonstrator acts as if it exists in a deterministic environment. As $n \to 0$, the expert acts as if it had an equal chance of transitioning to every possible successor state. At $n = 1$, the expert is the rational expert.

Optimism/Pessimism. Humans tend to systematically overestimate their chance experiencing of positive over negative events. We model this using experts that modify the probability they get outcomes based on the value of those outcomes:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} T^{1/\tau}(s'|s,a)\,(r(s,a,s') + \gamma V_i(s))$$

where $T^{1/\tau}(s'|s,a) \propto T(s'|s,a) e^{\left(r(s,a,s') + \gamma V_i(s)\right)/\tau}$. $1/\tau$ controls how pessimistic or optimistic the expert is. As $1/\tau \to +\infty$, the expert becomes increasingly certain that good transitions will happen. As $1/\tau \to -\infty$, the expert becomes increasingly certain that bad transitions will happen. As $1/\tau \to 0$, the expert approaches the rational expert.

2.2.4 Modifying the Reward: Prospect Theory

Next, we consider experts that use the modified Bellman update:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} T(s'|s,a) \left(f(r(s,a,s')) + \gamma V_i(s) \right)$$

where $f : \mathbb{R} \to \mathbb{R}$ is some scalar function. This is equivalent to solving the MDP with reward $f \circ r$. This allows us to model human behavior such as loss aversion and scope insensitivity.

Prospect Theory Kahneman & Tversky (2013) inspires us to consider a particular family of reward transforms:

$$f_c(r) = \begin{cases} \log(1 + |r|) & r > 0 \\ 0 & r = 0 \\ -c\log(1 + |r|) & r < 0 \end{cases}$$

c controls how loss averse the expert is. As $c \to \infty$, the expert primarily focuses on avoiding negative rewards. As $c \to 0$, the expert focuses on maximizing positive rewards and

2.2.5 Modifying the Sum between Reward and Future Value: Extremal

Extremal. Humans seem to exhibit duration neglect, sometimes only caring about the maximum intensity of an experiennce (Do et al., 2008). We model this using experts that use the Bellman step:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} T(s'|s,a) \left(\max\left[r(s,a,s'), (1-\alpha) r(s,a,s') + \alpha V_i(s) \right] \right)$$

These experts maximize the expected maximum reward along a trajectory, instead of the expected sum of rewards. As $\alpha \to 1$, the expert maximizes the expected maximum reward they achieve along their full trajectory. As $\alpha \to 0$, the expert becomes greedy, and only cares about the reward they achieve in the next timestep.

2.2.6 Modifying the Discounting

Myopic Discount. In practice, humans are often myopic, only considering immediate rewards. One way to model this is to decrease gamma in the Bellman update. At $\gamma = 1$, this is the rational expert. As $\gamma \to 0$, the expert becomes greedy and only acts to maximize immediate reward.

Myopic VI. As another way to model human myopia, we consider a expert that performs only h steps of Bellman updates. That is, this expert cares equally about rewards for horizon h, and discount to 0 reward after that. As $h \to \infty$, this expert becomes rational. If $h = 1$, this expert only cares about the immediate reward.

Hyperbolic Discounting. Human also exhibit hyperbolic discounting, with a high discount rate for the immediate future and a low discount rate for the far future. Alexander & Brown (2010) formulate this as the following Bellman update:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} T(s'|s,a) \left(r(s,a,s') + V_i(s) \right) / (1 + kV_i(s))$$

k modulates how much the expert prefers rewards now versus the future. As $k \to 0$, this expert becomes the rational expert.

3 Impact of Irrationalities on Reward Inference

3.1 Experimental Design

Simulation Environment. To reduce possible confounding from our choice of environment, we used a small 5x5 gridworld where the irrationalities nonetheless cause experts to exhibit different behavior. Our gridworld consists of three types of cells: ice, holes, and rewards. The expert can start in any ice cell. At each ice cell, the expert can move in one of the four cardinal directions. With

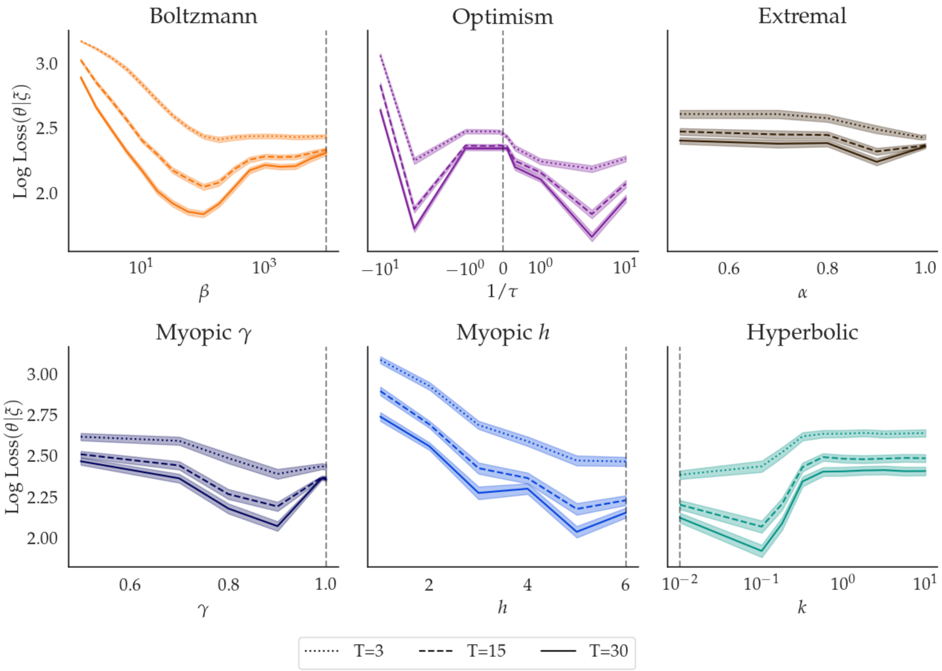


Figure 2: The log loss (lower = better) of the posterior as a function of the parameter we vary for each irrationality type. These six irrationalities all have parameter settings that outperform rational experts. For the models that interpolate to rational expert, we denote the value that is closest to rational using a dashed vertical line.

probability 0.8, they will go in that direction. With probability 0.2, they will instead go in one of the two adjacent directions. Holes and rewards are terminal states, and return the expert back to their start state. They receive a penalty of -10 for falling into a hole and $\theta_i \in [0, 4]$ for entering into the ith reward cell.

Dependent Measures. To separate the inference difficulty caused by suboptimal inference from the difficulty caused by expert irrationality, we perform the exact Bayesian update on the trajectory θ (Ramachandran & Amir, 2007), which gives us the posterior on θ given ξ:

$$P(\theta|\xi) = \frac{P(\xi|\theta)P(\theta)}{\int_{\theta'} P(\xi|\theta')P(\theta')}$$

We use two metrics to measure the difficulty of inference The first is the expected **log loss** of this posterior, or negative log-likelihood:

$$\text{Log Loss}(\theta|\xi) = E_{\theta,\xi\sim\pi_\theta}\left[-\log P(\theta|\xi)\right].$$

A low log loss implies that we are assigning a high likelihood to the true θ. As we are performing exact Bayesian inference with the true model $P(\xi|\theta)$ and prior $P(\theta)$, the log loss is equal to the entropy of the posterior $H(\theta|\xi)$.

The second metric is the $\mathbf{L^2}$**-distance** between the mean posterior θ and the actual theta:

$$L^2(\theta|\xi) = E_{\theta^*,\xi\sim\pi_{\theta^*}}\left[||E[\theta|\xi] - \theta^*||^2\right]$$

The closer the inferred posterior mean of θ is to the actual value θ^*, the lower the loss.

For each irrationality type, we calculate the performance of reward inference on trajectories of a fixed length T, with respect to the two metrics above. To sample a trajectory of length T from a expert, we fix θ^* and start state s. Then, we perform the expert's (possibly modified) Bellman updates until convergence to recover the policy π_{θ^*}. Finally, we generate rollouts starting from state s until T state, action pairs have been sampled from π_{θ^*}.

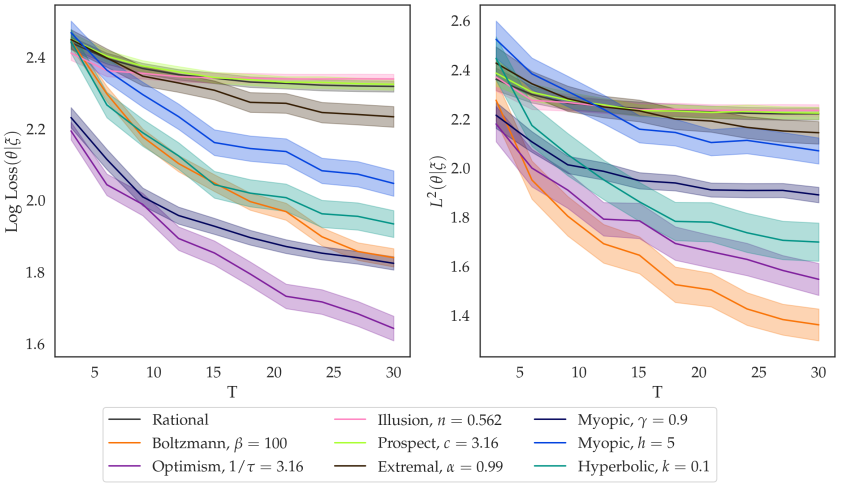


Figure 3: A best case analysis for each irrationality type: the log loss/L^2 distance from mean (lower=better) for experts, as a function of the length of trajectory observed. Each irrationality uses the parameter value that is most informative. As discussed in section 3.2, different irrationality types have different slopes and converge to different values. In addition, the best performing irrationality type according to log loss *is not* the best performing type according to L^2 loss.

3.2 Analysis

Impact of Each Irrationality. We found that of the 8 irrationalities we studied, 6 had parameter settings that lead to lower log loss than the rational expert. We report how the parameter influences the log loss for each of these experts in figure 2.[1] For $T = 30$, Optimism with $1/\tau = 3.16$ performed the best, followed by Boltzmann with $\beta = 100$ and Hyperbolic with $k = 0.1$. Both forms of Myopia also outperformed the rational expert, with best performance occurring at $\gamma = 0.9$ and $h = 5$. Finally, the Extremal expert also slightly outperformed the rational expert, with best performance at $\alpha = 0.9$. Notably, in every case, neither the most irrational expert nor the perfectly rational expert was the most informative.

Impact of Data for Different Irrationalities. Next, we investigate how the quality of inference varies as we increase the length of the observed trajectory T. We report our results for the best performing parameter for each irrationality type in figure 3. Interestingly, while both metrics decrease monotonically regardless of irrationality type, the rate at which they decrease differs by the irrationality type, and the best performing irrationality type according to log loss (Optimism) is not the best performing type according to L^2 distance (Boltzmann).

What is behind these differences? To explain these results, we use the notion of mutual information $\mathbf{I}(X;Y)$ between two variables, defined as:

$$\mathbf{I}(X;Y) = E_{X,Y}\left[\log\left(\frac{P(X,Y)}{P(X)P(Y)}\right)\right] = H(X) - H(X|Y)$$

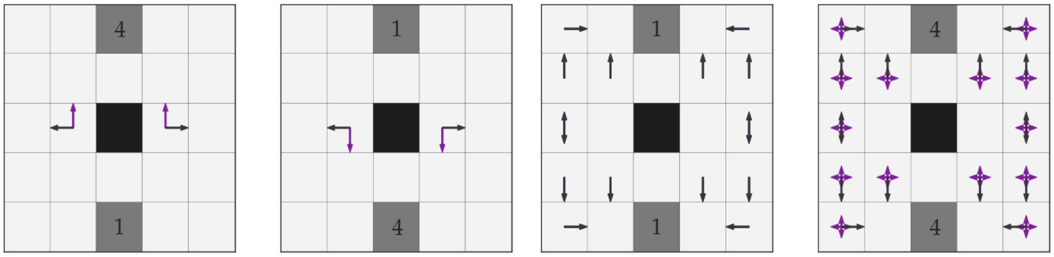
The mutual information measures how much our uncertainty about X decreases by observing Y.

For reward inference, the term we care about is the mutual information between the expert's trajectory and the reward parameters

$$\mathbf{I}(\theta;\xi) = E_{\theta,\xi\sim\theta}\left[\log\left(\frac{P(\theta,\xi)}{P(\theta)P(\xi)}\right)\right] = H(\theta) - H(\theta|\xi)$$

The mutual information $\mathbf{I}(\theta;\xi)$ is equal to a constant minus the posterior log loss under the true model. A expert with mutual information will cause the learner to have a lower posterior log loss.

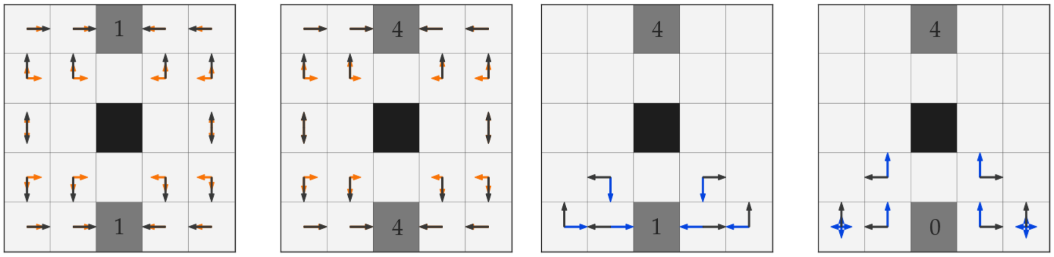
[1] The plots for the other two irrationalities are included in the appendix.



(a) Optimism Bias($\beta = 3.16$)

(b) Pessimism Bias($\beta = -3.16$

Figure 4: (a) Optimism bias produces different actions for $\theta^* = (4, 1)$ vs. $\theta^* = (1, 4)$ in the states shown: the rational policy is to go away from the hole regardless of θ, but an optimistic expert takes the chance and goes for the larger reward – up in the first case, down in the second. (b) Pessimism bias produces different actions for $\theta^* = (1, 1)$ vs. $\theta^* = (4, 4)$: when the reward is sufficiently large, the expert becomes convinced that no action it takes will lead to the reward, leading it to perform random actions.



(a) Boltzmann ($\beta = 100$)

(b) Myopia($h = 5$)

Figure 5: (a) Boltzmann-rationality produces different policies for $\theta^* = (1, 1)$ vs. $\theta^* = (4, 4)$: when $||\theta||$ is larger, the policy becomes closer to that of the rational expert. (b) A Myopic expert produces different policies for $\theta^* = (4, 1)$ vs. $\theta^* = (4, 0)$: while the rational expert always detours around the hole and attempts to reach the larger reward, myopia causes the myopic expert to go for the smaller source of reward when it is non-zero.

By the information processing inequality, we have the bound $\mathbf{I}(\theta; \xi) \le \mathbf{I}(\theta; \pi)$.

To have higher mutual information, different θs should be mapped to different policies πs. Indeed, we found that the experts that were able to outperform the rational expert were able to disambiguate between θs that the rational expert could not. To visualize this, we show examples of how the policy of several irrational experts differ when the rational expert's policies are identical in figures 4 and 5.

We plot the correlation between $\mathbf{I}(\theta; \xi)$ and $\mathbf{I}(\theta; \pi)$ in figure 6. Experts that have more informative policies tend to have more informative trajectories, but the correlation is not perfect. Notably, the Optimism expert has the most informative trajectories of length 30, but has less informative policies than the Boltzmann expert.

In the limit of infinite data from every state, we would have $\mathbf{I}(\theta; \xi) \to \mathbf{I}(\theta; \pi)$. However, as each trajectory begins from the same start state, and not every state is reachable with every policy, the bound is not achievable in general, even if we observe an arbitrarily large number of trajectories. This highlights the need for off-policy data in reward inference tasks.

4 Discussion

4.1 Summary

We show that, contrary to what we might expect, suboptimal experts can actually help an agent learn the reward function. Optimism bias, myopia (via heavier discounting or hyperbolic discounting),

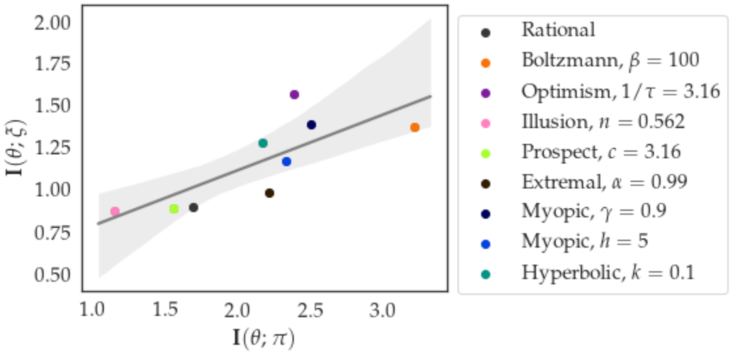


Figure 6: The informativeness of policies correlates with the informativeness of trajectories of length 30, as discussed in section 3.2

and noise via Boltzmann rationality were the most informative irrationalities in our environments, far surpassing the performance of the rational expert for their ideal settings. Our contribution overall was to identify a systematic set of irrationalities by looking at deviations in the terms of the Bellman update, and show that being irrational is not automatically harmful to inference by quantifying and comparing the inference performance for these different types.

4.2 Limitations and Future Work.

Estimating expert irrationality. One major limitation of our work is that our findings hold for when the learner knows the type and parameter value of the irrationality. In practice, reward inference will require solving the difficult task of estimating the irrationality type and degree (Armstrong & Mindermann, 2018; Shah et al., 2019). We still need to quantify to what extent these results still hold given uncertainty about the irrationality model. It does, however, seem crucial to reward inference that learners do reason explicitly about irrationality – not only is the learner unable to take advantage of the irrationality to make better inference if it does not model it, but actually reward inference in general suffers tremendously if the learner assumes the wrong type.

In figure 10 in the Appendix, we compare inference with the true model vs. with assuming a Boltzmann model as default. The results are quite striking: not knowing the irrationality harms inference tremendously. Whether irrationalities help, this means that it is really important to model them.

Generalization to other environments. A second limitation of our work is that we only tested these models in a limited range of environments. Further work is needed to test generalization of our findings across different MDPs of interest. Our analysis of mutual information lends credence to the Boltzmann rationality result generalizing well: these policies are much more varied with the reward parameters. In contrast, how useful the optimism bias is depends on the task: if we know about what to avoid already, as was the case for our learner, the bias is useful; if, on the other hand, we would know the goal but do not know what to avoid, the bias can hinder inference. Overall, this paper merely points out that there is a lot of richness to the ways in which these biases affect inference, and provides a quantitative comparison for a starting domain – much more is needed to gain a deeper understanding of this phenomenon.

Applications to real humans. A third limitation is that we do not know where real humans lie. Do they have the helpful irrationality types? Do they fall in the range of parameters for these types that help inference? And what happens when types combine? While these questions are daunting, there is also a hidden opportunity here: what if we could influence humans to exhibit helpful types of irrationality? It might be much easier for them, for instance, to act myopically than to act rationally. In the end, reward inference is the confluence of two factors: how well the robot learns, and how well the teacher teaches. Our results point out that it might be easier than previously thought to be a good teacher – even easier than being a rational expert.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004. URL: `http://people.eecs.berkeley.edu/~russell/classes/cs294/s11/readings/Abbeel+Ng:2004.pdf`.

William H Alexander and Joshua W Brown. Hyperbolically discounted temporal difference learning. *Neural computation*, 22(6):1511–1527, 2010.

Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, pp. 5598–5609, 2018. URL: `https://papers.nips.cc/paper/7803-occams-razor-is-insufficient-to-infer-the-preferences-of-irrational-agents.pdf`.

Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 243–252. JMLR. org, 2017.

Andrea Bajcsy, Dylan P Losey, Marcia K OMalley, and Anca D Dragan. Learning robot objectives from physical human interaction. *Conference on Robot Learning (CoRL)*, 2017.

Susan E. F. Chipman. *The Oxford Handbook of Cognitive Science*. Oxford University Press, 11 2014. ISBN 9780199842193.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017. URL: `https://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.pdf`.

Amy M Do, Alexander V Rupert, and George Wolford. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review*, 15(1):96–98, 2008.

Owain Evans and Noah D Goodman. Learning the preferences of bounded agents. In *NIPS Workshop on Bounded Optimality*, volume 6, 2015. URL: `https://pdfs.semanticscholar.org/d55b/3f05ad612ecd0ae160850e9f07b1926e51bc.pdf`.

Owain Evans, Andreas Stuhlmüller, and Noah Goodman. Learning the preferences of ignorant, inconsistent agents. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. URL: `https://pdfs.semanticscholar.org/31bf/e42e77a572bd83c0529e0f03bc3dc8af52c2.pdf`.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58, 2016. URL: `http://proceedings.mlr.press/v48/finn16.pdf`.

Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. Pragmatic-pedagogic value alignment. *arXiv preprint arXiv:1707.06354*, 2017. URL: `https://arxiv.org/pdf/1707.06354.pdf`.

Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, pp. 2625–2633, 2013.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in neural information processing systems*, pp. 6765–6774, 2017.

Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

Antony Jameson and Eliezer Kreindler. Inverse problem of linear optimal control. *SIAM Journal on Control*, 11(1):1–19, 1973. URL: https://epubs.siam.org/doi/pdf/10.1137/0311001.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Rudolf Emil Kalman. When is a linear control system optimal? *Journal of Basic Engineering*, 86(1):51–60, 1964. URL: https://asmedigitalcollection.asme.org/fluidsengineering/article-abstract/86/1/51/392203.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453*, 2018.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. URL: https://arxiv.org/pdf/1509.02971.pdf.

Robert Tyler Loftin, James MacGlashan, Bei Peng, Matthew E Taylor, Michael L Littman, Jeff Huang, and David L Roberts. A strategy-aware technique for learning behaviors from discrete human feedback. In *AAAI Conference on Artificial Intelligence*, 2014.

Christopher G Lucas, Thomas L Griffiths, Fei Xu, and Christine Fawcett. A rational model of preference learning and choice prediction by children. In *Advances in neural information processing systems*, pp. 985–992, 2009. URL: http://papers.nips.cc/paper/3579-a-rational-model-of-preference-learning-and-choice-prediction-by-children.pdf.

R Duncan Luce. Individual choice behavior. 1959.

James MacGlashan, Monica Babes-Vroman, Marie desJardins, Michael L Littman, Smaranda Muresan, Shawn Squire, Stefanie Tellex, Dilip Arumugam, and Lei Yang. Grounding english commands to reward functions. In *Robotics: Science and Systems*, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015. URL: https://daiwk.github.io/assets/dqn.pdf.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937, 2016. URL: http://proceedings.mlr.press/v48/mniha16.pdf.

Katja Mombaur, Anh Truong, and Jean-Paul Laumond. From human to humanoid locomotionan inverse optimal control approach. *Autonomous robots*, 28(3):369–383, 2010. URL: https://link.springer.com/content/pdf/10.1007/s10514-009-9170-7.pdf.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000. URL: http://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

Ellis Ratner, Dylan Hadfield-Menell, and Anca D Dragan. Simplifying reward design through divide-and-conquer. *arXiv preprint arXiv:1806.02501*, 2018.

Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *Advances in Neural Information Processing Systems*, pp. 1454–1465, 2018.

Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca D Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. *arXiv preprint arXiv:1906.09624*, 2019. URL: `https://arxiv.org/pdf/1906.09624.pdf`.

Roger N Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345, 1957. URL: `https://link.springer.com/content/pdf/10.1007/BF02288967.pdf`.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016. URL: `http://web.iitd.ac.in/~sumeet/Silver16.pdf`.

Sumeet Singh, Jonathan Lacotte, Anirudha Majumdar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via semi-and non-parametric methods. *arXiv preprint arXiv:1711.10055*, 2017.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. 2010.

Policy	Parameter	Values
Rational	γ	[0.99]
Boltzmann	β	[1, 1.78, 3.16, 5.62, 10, 17.8, 31.6, 56.2, 100, 178, 316, 562, 1000,1780, 3160, 5620, 10000]
Optimism	$1/\tau$	[−10, −3.16, −1, −0.316, −0.1, 0.1, 0.316, 1, 3.16, 10]
Illusion of Control	n	[0.1, 0.178, 0.316, 0.562, 1, 1.78, 3.16, 5.62, 10]
Prospect Theory	c	[0.1, 0.178, 0.316, 0.562, 1., 1.78, 3.16, 5.62, 10]
Extremal	α	[0.5, 0.7, 0.8, 0.9, 0.99, 0.999]
Myopic γ	γ	[0.5, 0.7, 0.8, 0.9, 0.99, 0.999]
Myopic h	h	[1, 2, 3, 4, 5, 6]
Hyperbolic	k	[0.01, 0.1, 0.178, 0.316, 0.562, 1, 1.78, 3.16, 5.62, 10]

Table 1: The parameter values we search over for each policy.

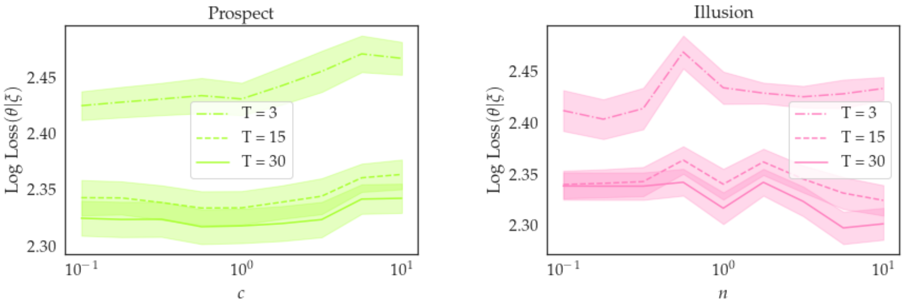


Figure 7: Log loss for the posterior on θ, given trajectories from the Prospect Theory expert and the Illusion of Control expert.

A More experimental details

To enable exact inference, we discretized θ, using 5 evenly spaced points for each θ_i. Our specific grid is included in figures 4 and 5 As there are two reward cells, this gives us 25 possible distinct reward parameters. We assumed a uniform prior on the reward parameter.

We list the parameter values we search over for each policy in table 1. Except for myopic γ and myopic h, we use $\gamma = 0.99$. For myopic h, we use $\gamma = 1$.

From each start state, we sample 10 trajectories of each length for each reward parameter, policy combination.

B Additional results

We include the plots for the log loss of trajectories from the Prospect Theory and Illusion of Control experts in 7

In addition, we include the plots for the L^2 loss for all 8 irrationalities in figures 8 and figure 9.

C Model misspecification greatly impairs inference

Given that several types of irrationality can help inference when the form of irrationality is known, a natural question to ask is how important is it to known the irrationality exactly. To investigate this, we plot the log loss of the posterior of a learner who falsely assumes that the expert is Boltzmann-

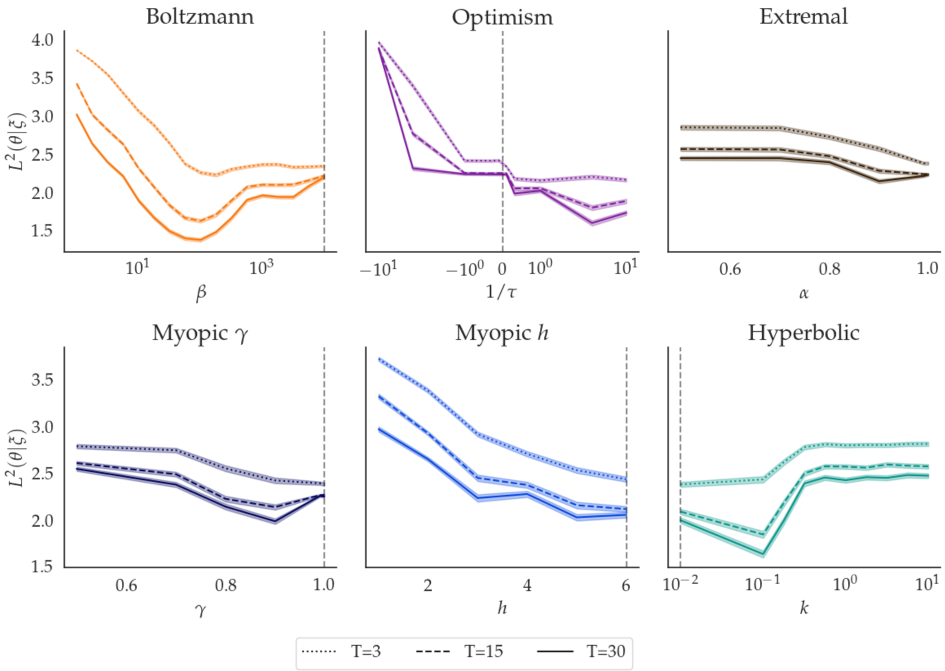


Figure 8: The L^2 distance (lower = better) of posterior mean of θ to the true θ^*,s as a function of the parameter we vary for each irrationality type. These six irrationalities all have parameter settings that outperform rational experts. For the models that interpolate to rational expert, we denote the value that is closest to rational using a dashed vertical line.

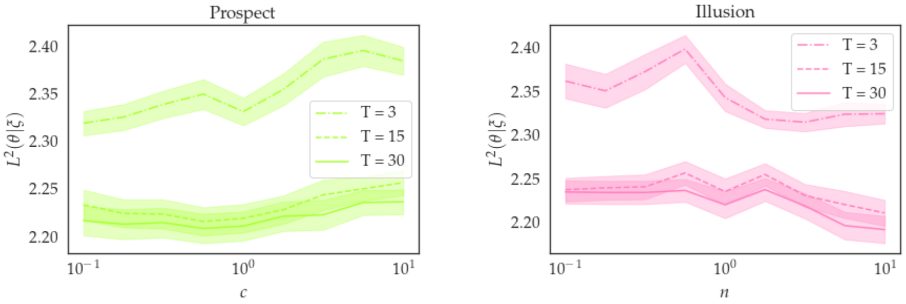


Figure 9: The L^2 distance (lower=better) of the posterior mean θ to th true θ^*, given trajectories from the Prospect Theory expert and the Illusion of Control expert.

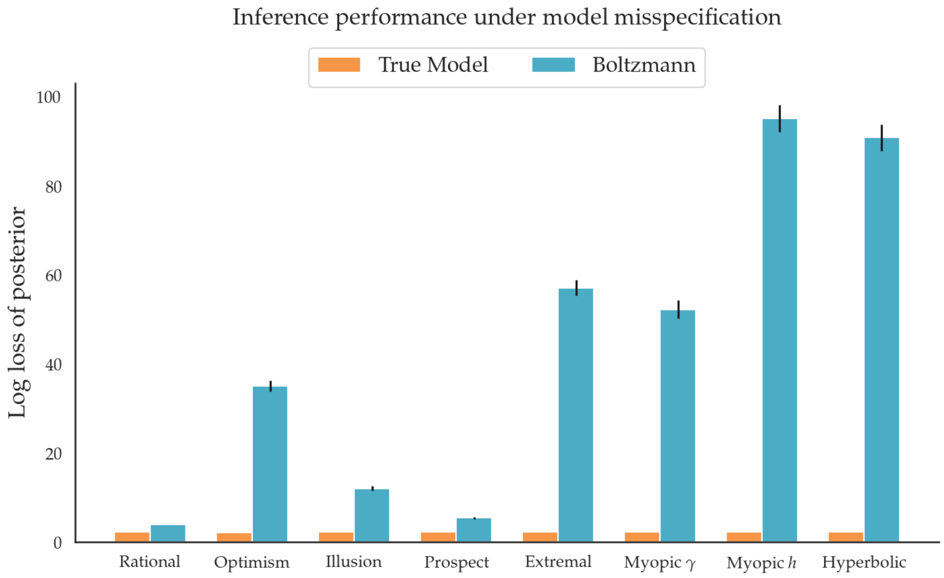


Figure 10: A comparison of reward inference using a correct model of the irrationality type, versus always using a Boltzman model. (Lower log loss = better.) The inference impairment from using the misspecified irrationality model (Boltzmann) greatly outweighs the variation in inference performance caused by the various irrationality types themselves. Hence, compared to using a misspecified model of irrationality, expert irrationality is not in itself a major impairment to reward inference, and sometimes expert irrationality can even helps when a model of the irrationality is known.

rational with $\beta = 100$. Where applicable, the log loss is averaged over possible hyperparameter settings for the expert.

We report the results in figure 10. The log loss of the posterior if we wrongly imagine the expert is Boltzmann-rational far outweighs differences between particular irrationality types.

C.1 Why is using a misspecified irrationality type for inference so bad?

Fundamentally, misspecification is bad for inference because different experts might exhibit the same action only under different reward parameters. For example, consider figure the case where the actual expert is myopic, with small n. Then the myopic agent might go toward a closer reward even if it is much smaller, as shown in figure 11. This would cause the learner to falsely infer that the closer reward is quite large, leading to a posterior with extremely high log loss when the reward is actually smaller.

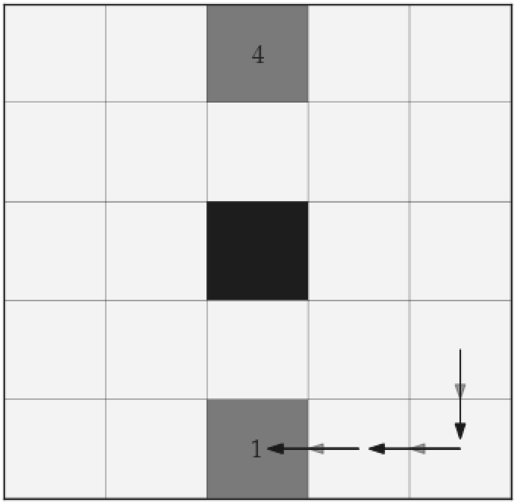


Figure 11: An example of why assuming Boltzmann is bad for a myopic agent - the Boltzmann rational agent would take this trajectory only if the reward at the bottom was not much less than the reward at the top. The myopic agent with $n \leq 4$, however, only ”sees” the reward at the bottom. Consequently, inferring the preferences of the myopic agent as if it were Boltzmann leads to poor performance in this case.