Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning

Smitha Milli, Anca D. Dragan
University of California, Berkeley
{smilli, anca}@berkeley.edu

Abstract

It is incredibly easy for a system designer to misspecify the objective for an autonomous system ("robot"), thus motivating the desire to have the robot *learn* the objective from human behavior instead. Recent work has suggested that people have an interest in the robot performing well, and will thus behave *pedagogically*, choosing actions that are informative to the robot. In turn, robots benefit from interpreting the behavior by accounting for this pedagogy. In this work, we focus on misspecification: we argue that robots might not know whether people are being pedagogic or literal and that it is important to ask which assumption is *safer* to make. We cast objective learning into the more general form of a common-payoff game between the robot and human, and prove that in any such game literal interpretation is *more robust* to misspecification. Experiments with human data support our theoretical results and point to the sensitivity of the pedagogic assumption.

1 INTRODUCTION

It is notoriously difficult for system designers to directly specify the correct objective for a system (Krakovna, 2018; Lehman et al., 2018; Clark & Amodei, 2017). This difficulty has sparked a line of work that instead aims to infer the correct objective from other forms of human input, such as demonstrations (Ng & Russell, 2000; Abbeel & Ng, 2004; Ziebart et al., 2008), comparisons (Wirth et al., 2017; Sadigh et al., 2017; Christiano et al., 2017), or corrections (Bajcsy et al., 2017; Jain et al., 2015).

These methods operate under the assumption that human behavior is near-optimal with respect to the objective to be inferred. This assumption makes sense in many domains, like an autonomous car learning an objective function for driving through observing human drivers (Levine & Koltun, 2012). However, recent work shows that this assumption may not hold in collaborative settings (Ho et al., 2016), in which the human is *aware* that the robot needs to learn. In such settings, the person might optimize for *teaching* the robot about the objective, which is not the same as directly optimizing the objective itself (Dragan et al., 2013b; Hadfield-Menell et al., 2016; Ho et al., 2016). Figure 1 shows an example of the difference between the two. When the human optimizes for the objective, she takes any optimal path to the goal. When the human optimizes for teaching the objective, she takes the path that best signals the objective.

We refer to these two types of human behavior as *literal* and *pedagogic*, and note that the robot can interpret human behavior using either model:

1. The **literal human** directly optimizes for the objective.
2. The **literal robot** infers the objective while assuming the human is literal.
3. The **pedagogic human** optimizes for teaching the *literal* robot the objective.
4. The **pedagogic robot** (sometimes called "pragmatic" (Fisac et al., 2018)) infers the objective while assuming the human is *pedagogic*.

⋮

In general, this recursion could go on further, and in theory, it is always better for the human and robot to be at a deeper level of recursion. The potential for increased performance suggests that we should try to make our robots more pedagogic. Indeed, this is the direction suggested and pursued by Fisac et al. (2018); Malik et al. (2018); Hadfield-Menell et al. (2016)

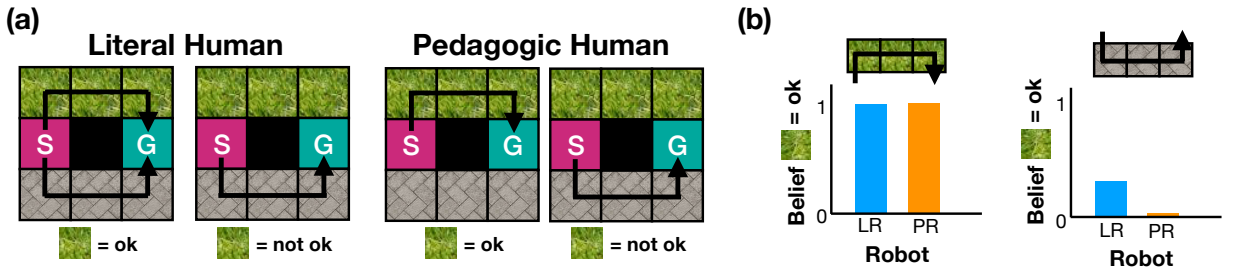

Figure 1: In this scenario, the robot infers whether it is ok to walk on the grass or not from a demonstration provided by the human (it already knows pavement is always ok). (a) The literal and pedagogic human's demonstrations. When grass is ok, the literal human is equally likely to walk on the grass or pavement. On the other hand, when grass is ok, the pedagogic human always walks on the grass in order to signal that grass is ok. (b) The beliefs of the literal robot **LR** and pedagogic robot **PR** after observing a demonstration. **LR** and **PR** assume human is literal and pedagogic, respectively. **LR** and **PR** differ in how strong their beliefs are after witnessing the human walk on pavement, leading to different problems when the human is misspecified. If the human is literal, but the robot is pedagogic, then it makes too strong of an inference. On the other hand, if the human is pedagogic, but the robot is literal, then it makes too weak of an inference.

However, the increased performance is contingent on the human being pedagogic. In practice, it may be difficult to know whether the human is acting literally or pedagogically, and furthermore, different humans may behave differently. We argue that the robot should be robust to misspecification of the *human*, and thus it is important to ask which assumption – literal or pedagogic – is *safer* to make.

Which is worse—interpreting a pedagogic human literally (literal robot + pedagogic human) or interpreting a literal human pedagogically (pedagogic robot + literal human)? In both cases, the model of the human is incorrect, and we might expect that which is better depends on the context and the task; surprisingly, we are able to prove that regardless of the task, the literal robot is more robust, suggesting that it may be safer to simply use a *literal* robot. Our contributions are the following:

- Section 2: *Theoretical Analysis.* We cast objective learning into the more general form of a common-payoff game between the human and robot. We then prove that a pedagogic robot and a literal human always do worse than a literal robot and a pedagogic human, showing that misspecification is worse in one direction than the other.

- Section 3: *Empirical Analysis.* We test the effects of misspecification on data from human teaching. The data confirms that assuming pedagogic behavior when people are literal is worse than assuming literal behavior when people are pedagogic. Surprisingly, we find that the literal robot does better than the pedagogic robot *even when people are trying to be pedagogic*, because of the discrepancy between the pedagogic *model* of humans and real human behavior.

- Section 4: *Theoretical vs Empirical.* Our empirical results are surprising, in that, the pedagogic model is a state of the art cognitive science model that is fit to the human data and has relatively high predictive accuracy, yet the pedagogic robot does worse than the literal robot *with humans who are trying to be pedagogic*. We use our theory to derive a hypothesis for why this could be. The hypothesis is that in practice different humans vary in how literal or pedagogic they are, which our theory implies will degrade the performance of the pedagogic robot more than the literal robot. We find positive evidence for this hypothesis, indicating that robustness to a *population* of humans is an important consideration in choosing a pedagogic versus literal robot.

- Section 5: *Can we "fix" the pedagogic robot?* An intuitive idea for improving the pedagogic robot is to give it a model of the pedagogic human that has higher predictive accuracy. For example, what if instead of assuming all people are pedagogic, the robot estimated how pedagogic each person is? Unfortunately, we find that this makes no difference. And in fact, we show that better models can actually *worsen* performance due to a subtle, yet remarkable fact: a human model with higher *predictive accuracy* does not necessarily imply higher *inferential accuracy* for the robot.

In conclusion, we found that not only are pedagogic robots less robust to the human's recursion level, they can also perform worse when people are actually being

pedagogic – even with a state of the art pedagogic model tuned to human data. This points to a surprising brittleness of the pedagogic assumption. Further, we point out that a more predictive human model will not necessarily solve the problem because improving the model's predictive accuracy does not necessarily lead to better robot inference.

The difficulty of *objective specification* was an important motivation for pursuing objective learning in the first place. But in our pursuit of objective learning, we ought to be careful to not simply trade the problem of objective specification for the equally, if not more, difficult problem of *human specification*. In practice, humans will deviate from our models of them, and thus, it is important to understand which assumptions are robust and which are not.

Rather than trying to make the robot more pedagogic, which requires brittle assumptions on the human, an alternative direction may be to help the *human* become more pedagogic. How can we make robots that are easier for humans to teach? Perhaps simpler robots are actually better, if it means that humans can more easily teach them.

2 THEORETICAL ANALYSIS

In this section, we cast objective learning into the more general form of a common-payoff[1] game between the robot and human. We then formalize the literal/pedagogic robot/human, and prove that in *any* common-payoff game a literal robot and pedagogic human perform better than a pedagogic robot and literal human.

2.1 GENERALIZING OBJECTIVE LEARNING TO CI(RL)

Before proceeding to our proof, we give background on how common-payoff games generalize standard objective learning. In particular, objective learning can be modeled as a cooperative inverse reinforcement learning (CIRL) game (Hadfield-Menell et al., 2016), a common-payoff game in which only the human knows an objective/reward function r.

Formally, a CIRL game is defined as a tuple $\langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T, \{\mathcal{R}, r\}, P_0, \gamma\}$ where $\mathcal{S}$ is the set of states, $\mathcal{A}^{\mathbf{H}}$ and $\mathcal{A}^{\mathbf{R}}$ are the set of actions available to the human and robot, $T(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{R}})$ is the transition distribution specifying the probability of transitioning to a new state s' given the previous state s and the actions $a^{\mathbf{H}}$ and $a^{\mathbf{R}}$ of both agents, R is the space of reward functions, $r : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \to \mathbb{R}$ is the shared reward function (known only to $\mathbf{H}$), P_0 is the initial distribution over states and reward functions, and γ is the discount factor.

The joint payoff in a CIRL game is traditionally *value*, expected sum of rewards. Since only the human in CIRL knows the shared reward function, this indirectly incentivizes the human to act in ways that signal the reward function and incentivizes the robot to learn about the reward function from the human's behavior[2]. However, we can also consider a version that directly incentivizes reward inference by making the payoff the accuracy of the robot's inference.

To illustrate how objective learning settings are special cases of CIRL games, we give an example for the learning from demonstrations setting.

Example 2.1 (Demonstration-CI(RL))**.** Learning from demonstrations can be modeled as a CIRL game with two phases. In the first phase, the human provides demonstrations. In the second phase...

(a) (CIRL) the robot acts in the environment. The game's joint payoff for the robot and human is the value (expected sum of rewards) attained by the robot.
(b) (CI) the robot outputs an estimate of the reward function. The game's joint payoff for the robot and human is a measure of the accuracy of the robot's inference.

The second formulation (b) is a *cooperative inference* (CI) problem (Yang et al., 2018; Wang et al., 2019). In CI, there is a teacher (e.g human) who is teaching a learner (e.g. robot) a hypothesis r (e.g the reward) via data d (e.g. demonstrations). The human has a distribution over demonstrations $p^{\mathbf{H}}(d \mid r)$ and the robot has a distribution over rewards $p^{\mathbf{R}}(r \mid d)$. Given a starting distribution of human demonstrations, $p_0^{\mathbf{H}}(d \mid r)$, the optimal solution for these two distributions can be found via fixed-point iteration of the following recursive equations[3]:

$$p_k^{\mathbf{R}}(r \mid d) \propto p_k^{\mathbf{H}}(d \mid r)\,, \tag{1}$$

$$p_{k+1}^{\mathbf{H}}(d \mid r) \propto p_k^{\mathbf{R}}(r \mid d)\,. \tag{2}$$

[1] A game in which all agents have the same payoff.

[2] In fact, in Demonstration-CIRL (Example 2.1), the best that the robot can do is infer a posterior distribution over rewards from the human's demonstrations, and then act optimally with respect to the posterior mean.

[3] Wang et al. (2019) show that fixed-point iteration converges for all discrete distributions. For simplicity, we have written equations (1) and (2) assuming that the human and robot's prior over rewards and demonstrations is uniform, but in general, the equations can incorporate any prior (Yang et al., 2018).

2.2 PROOF: LITERAL ROBOTS ARE MORE ROBUST

We now proceed to formalize what we mean by literal and pedagogic and to prove that the literal robot is more robust to misspecification of whether the human is literal or pedagogic. Suppose that the human and robot are acting in a cooperative game. Let $\mathcal{H}$ and $\mathcal{R}$ be the space of policies for the human and robot, respectively. The joint payoff for the human and robot is denoted by $U : \mathcal{H} \times \mathcal{R} \to \mathbb{R}$. The joint payoff function could be value, as assumed by CIRL, or accuracy of inference, as assumed by CI.

To define literal and pedagogic, we define a set of recursive policies for the human and robot. Let H_0 be a starting human policy. For $k \geq 0$, define the recursive policies

$$R_k = \mathbf{BR}(H_k)\,, \tag{3}$$
$$H_{k+1} = \mathbf{IR}(R_k)\,. \tag{4}$$

We assume that at each level of recursion the robot does a best response $\mathbf{BR} : \mathcal{H} \to \mathcal{R}$. However, for the human, we only assume that at each step she improves over her previous policy. This allows for arbitrary irrationalities, so long as the next policy is at least as good as the previous one. We call this an "improving" response $\mathbf{IR} : \mathcal{R} \to \mathcal{H}$, and define both types of responses below.

Definition 2.1. A *best response* for the robot is a function $\mathbf{BR} : \mathcal{H} \to \mathcal{R}$ such that for any human policy $H \in \mathcal{H}$ and robot policy $R \in \mathcal{R}$,

$$U(H, \mathbf{BR}(H)) \geq U(H, R)\,.$$

Definition 2.2. An *improving response* for the human is a function $\mathbf{IR} : \mathcal{R} \to \mathcal{H}$ such that $\forall\, k \geq 0$,

$$U(\mathbf{IR}(R_k), R_k) \geq U(H_k, R_k)\,.$$

We give special emphasis to what we call the *literal* human and robot, H_0 and R_0, and the *pedagogic* human and robot, H_1 and R_1.[4] We now provide an example of the literal and pedagogic policies for the Demonstration-CI setting (Example 2.1b). In this case, the human and robot policies can be modeled by the recursive CI equations (1) and (2).

1. The **literal human** H_0 is noisily-optimal with respect to the reward function r. The probability $p_0^{\mathbf{H}}(d \mid r)$ that she gives a demonstration d is exponentially proportional to the reward of the demonstration, denoted by $r(d)$:

$$p_0^{\mathbf{H}}(d \mid r) \propto \exp(r(d))\,.$$

2. The **literal robot** R_0 does a $\mathbf{BR}$ to the literal human, i.e. does Bayesian inference assuming the human is literal,

$$p_0^{\mathbf{R}}(r \mid d) \propto p_0^{\mathbf{H}}(d \mid r)\,,$$

and then uses the posterior mode as its estimate for the reward r.

3. The **pedagogic human** H_1 picks demonstrations that are informative to the literal robot[5]:

$$p_1^{\mathbf{H}}(d \mid r) \propto p_0^{\mathbf{R}}(r \mid d)\,.$$

Note this is not a best response, which would unrealistically require the human to choose $\arg\max_d p_0^{\mathbf{R}}(r \mid d)$.

4. The **pedagogic robot** R_1 does a $\mathbf{BR}$ to the pedagogic human, i.e. does Bayesian inference assuming the human is literal,

$$p_1^{\mathbf{R}}(r \mid d) \propto p_1^{\mathbf{H}}(d \mid r)\,,$$

and then uses the posterior mode as its estimate for the reward r.

We show that for *any* common-payoff game the payoffs for the literal/pedagogic human/robot pairs have the following ranking:

$$\begin{aligned}&\text{Pedagogic } R_1, H_1 \geq \text{Literal } R_0, \text{Pedagogic } H_1\\ &\geq \text{Literal } R_0, H_0 \geq \text{Pedagogic } R_1, \text{Literal } H_0\,.\end{aligned} \tag{5}$$

In particular, a pedagogic robot R_1 and a literal human H_0 always do worse than a literal robot R_0 and a pedagogic human H_1, showing misspecification is worse one way than the other. The ranking has the following straight-forward proof.

Claim 2.1. *In any common-payoff game, the ranking of payoffs between a literal/pedagogic human/robot is*

$$U(H_1, R_1) \geq U(H_1, R_0) \geq U(H_0, R_0) \geq U(H_0, R_1)\,.$$

Proof. Since $R_1 = \mathbf{BR}(H_1)$, we have $U(H_1, R_1) \geq U(H_1, R_0)$. Since $H_1 = \mathbf{IR}(R_0)$, we have $U(H_1, R_0) \geq U(H_0, R_0)$. Since $R_0 = \mathbf{BR}(H_0)$, we have $U(H_0, R_0) \geq U(H_0, R_1)$. □

[4] There is nothing that requires the literal level to be at recursion level 0 and the pedagogic level to be at recursion level 1. Our proof of Claim 2.1 holds when the literal level is any $k \geq 0$ and the pedagogic level is $k + 1$.

[5] Typically, the human is modeled as being exponentially more informative: $p_1^{\mathbf{H}}(d \mid r) \propto \exp(p_0^{\mathbf{R}}(r \mid d))$. This is the form we will use in our experiments.

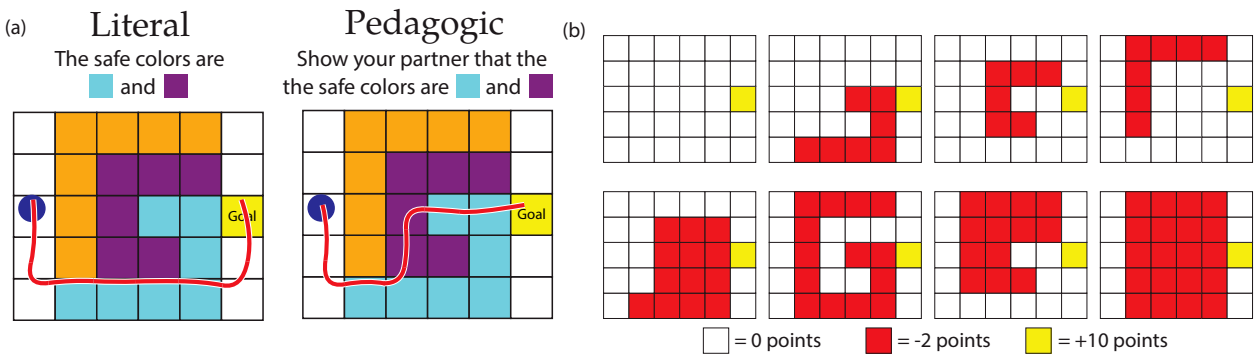

Figure 2: (a) The instructions given to participants in the literal and pedagogic condition, and a sample demonstration from both conditions. In the pedagogic case, participants were more likely to visit multiple safe colors, as well as to loop over safe tiles multiple times. (b) All possible reward functions. Each tile color can be either safe (0 points) or dangerous (-2 points). Figure modified from Ho et al. (2018).

3 EMPIRICAL ANALYSIS

In this section, we test our results in practice using experimental data from actual humans, who provide demonstrations that the robot uses to infer the objective from. These experiments are a way of stress-testing our theoretical results, which assumed that at each level of recursion, the robot computes a best response and that the human never does worse than her previous level. In practice, there will be a difference between what humans actually do, and the model of the human, which could cause both the human and robot to break our theoretical assumptions.

3.1 EXPERIMENTAL DESIGN

We test the performance of literal/pedagogic robot/human pairs on data from Experiment 2 in (Ho et al., 2016), which was subsequently followed up on in (Ho et al., 2018). In the experiment, humans are asked to act in different types of gridworlds. The gridworld has one goal state that is worth 10 points and three other types of tiles (orange, purple, cyan). Each type of tile can each be either "safe" (0 points) or "dangerous" (-2 points). Thus, there are $2^3 = 8$ possible reward functions, which are depicted in Figure 2b.

Sixty particpants were recruited from Mechanical Turk. The participants are told that they will get two cents of bonus for each point they get. Participants were split into two conditions, a literal and pedagogic condition, depicted in Figure 2a. In the literal condition, the participant only gets points for their own actions in the gridworld. In the pedagogic condition, the participant is told that their demonstration will be shown to another person, a learner, who will then apply what they learn from the demonstration to act in a separate gridworld. The participant still gets points based on their own actions, but is also told that the number of points the learner receives will be added as a bonus.

3.2 HUMAN AND ROBOT MODELS

We model the robot and human following Ho et al. (2018). We use the same model parameters that Ho et al. (2018) found to be the best qualitative match to the human demonstrations.

Notation. Let $\mathcal{S}$ be the set of states and $\mathcal{A}$ be the set of actions. The gridworld has a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. The optimal Q-value function for a reward function r is denoted by $Q_r^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The robot has a uniform prior belief over the reward function, i.e, it puts uniform probability on all $2^3 = 8$ reward functions depicted in Figure 2b. The human and robot models are as follows.

1. **Literal H.** At each time step t, the probability $H_L(a_t|s_t, r)$ that the literal human takes action a_t given state s_t and the reward r is exponentially proportional to the optimal Q-value,

$$H_L(a_t \mid s_t, r) \propto \exp(Q_r^*(s_t, a_t)/\tau_L)\,. \quad (6)$$

   The temperature parameter τ_L controls how noisy the human is.

2. **Literal R.** The robot does Bayesian inference, while assuming that the human is literal. The robot's posterior belief at time $t+1$ is

$$R_L^{t+1}(r) \propto H_L(a_t \mid s_t, r) \cdot T(s_{t+1} \mid s_t, a_t) \cdot R_L^t(r)\,. \quad (7)$$

3. **Pedagogic H.** The pedagogic human optimizes a reward function r' that trades-off between optimizing

the reward in the gridworld and teaching the literal robot the reward. At time step t, the reward is $r'(s_t, a_t, s_{t+1}) = r(s_t, a_t, s_{t+1}) + \kappa(R_L^{t+1}(r) - R_L^t(r))$. The parameter $\kappa \geq 0$ controls how "pedagogic" the human is. The probability $H_P(a_t \mid s_t, r)$ that the pedagogic human takes action a_t given state s_t and the reward r is exponentially proportional to the optimal Q-value associated with the modified reward r',

$$H_P(a_t \mid s_t, r) \propto \exp(Q^*_{r'}(s_t, a_t)/\tau_P)\,. \quad (8)$$

The temperature parameter τ_P controls how noisy the human is.

4. **Pedagogic R.** The robot does Bayesian inference, while assuming that the human is literal. The robot's posterior belief at time $t+1$ is

$$R_P^{t+1}(r) \propto H_P(a_t \mid s_t, r) \cdot T(s_{t+1} \mid s_t, a_t) \cdot R_P^t(r)\,. \quad (9)$$

3.3 RESULTS

We evaluate each literal/pedagogic robot/human pair on the accuracy of the robot's inference, i.e, $\mathbb{P}(\hat{r} = r)$, where r is the true reward and $\hat{r}$ is the robot's guess. We take the robot's guess $\hat{r}$ to be the mode of its belief, as given by the robot models (7) and (9). We test each pair with both the demonstrations generated by actual humans and demonstrations generated by simulating humans according to the human models (6) and (7).

Figure 3 depicts our experimental results[6]. We refer to the actual human as **AH**, the human model as **H**, and the robot as **R**. Consistent with the theory, the performance of pedagogic **R** and literal **AH** is (significantly) worse than that of literal **R** and pedagogic **AH**, validating that misspecification is worse one way than the other. However, the overall ranking of robot/human pairs does not match the theoretical ranking (Equation 5) we expected. Surprisingly, even when **AH** is pedagogic, pedagogic **R** performs (insignificantly) worse than literal **R**. The empirical ranking is

$$\text{Literal } \mathbf{R}, \text{Pedagogic } \mathbf{AH} \geq \text{Pedagogic } \mathbf{R}, \mathbf{AH} \geq \text{Literal } \mathbf{R}, \mathbf{AH} \geq \text{Pedagogic } \mathbf{R}, \text{Literal } \mathbf{AH}\,.$$

The empirical ranking carries a stronger implication than our theoretical ranking—it implies that regardless of whether the human is literal or pedagogic, the robot should be literal.

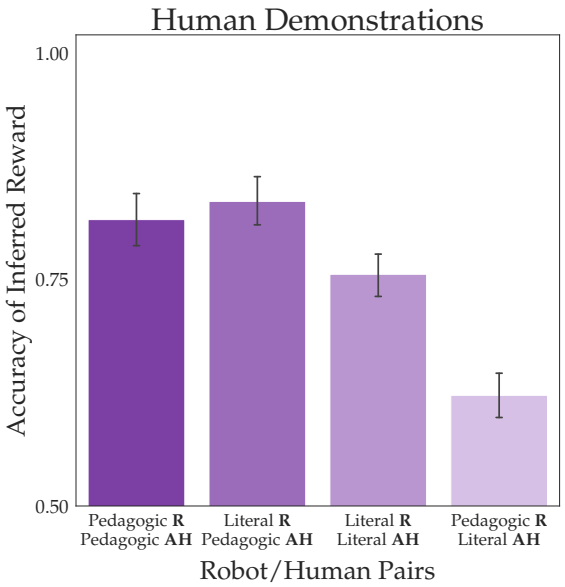

Figure 3: The accuracy of the robot's inferred reward for different pairs of human/robot pairs under the demonstrations provided by actual humans.

4 THEORETICAL VS EMPIRICAL

In our empirical analysis, we found a perplexing result: the literal robot does better than the pedagogic robot *even when people are trying to be pedagogic*. How could this be? Figure 4 shows the accuracy of robot/human pairs when the human demonstrations are simulated from the literal **H** and pedagogic **H** models described in equations (6) and (8). In the simulations, we find, as expected, that pedagogic **R** and pedagogic **H** do better than literal **R** and pedagogic **H**. So clearly, the reason pedagogic **R** does worse in the human experiments is that pedagogic **AH** (actual humans) is not the same as the pedagogic **H** (the human model).

But, in *what way* does pedagogic **AH** deviate from pedagogic **H**? Why is literal **R** more robust to the deviation than pedagogic **R**? We derive a hypothesis from our theory. In the theoretical ranking (Equation 5), the performance of literal **R** and literal/pedagogic **H** is sandwiched between the performance of pedagogic **R** + pedagogic **H** and pedagogic **R** + literal **H**. Thus, literal **R** is more robust to whether **H** is literal or pedagogic than pedagogic **R**. This suggests an explanation for why literal **R** does better, namely, that pedagogic **AH** is actually sometimes literal!

Hypothesis 4.1. *Pedagogic **AH** is actually "in between"*

[6] All error bars and confidence bands in the paper depict bootstrapped 95% confidence intervals.

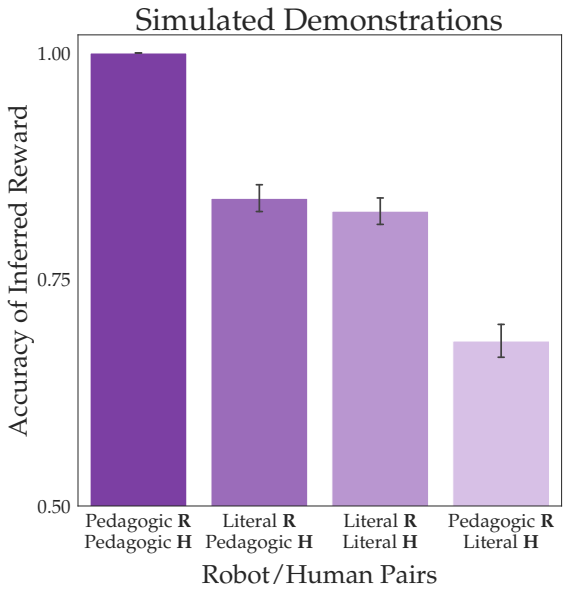

Figure 4: The accuracy of the robot's inferred reward for different pairs of human/robot pairs under the demonstrations provided by humans simulated according to the human models in Section 3.2.

*literal **H** and pedagogic **H**. If this is true, then our theory implies that literal **R** will be affected less than pedagogic **R**, potentially explaining why pedagogic **R** does worse with pedagogic **AH** than literal **R**.*

To test Hypothesis 4.1, we create two models that are mixtures of the literal and pedagogic model. We then measure how much better the mixture models are at predicting pedagogic **AH**, and how much the literal and pedagogic **R** are affected by demonstrations generated from the mixture models.

4.1 DEMONSTRATION MIXTURE MODEL

First, we test what we call the "demonstration mixture model". It is possible that the humans in the pedagogic condition from Ho et al. (2016) followed different strategies; some may have attempted to be pedagogic, but others may have simply been literal. If we look at which model, literal or pedagogic, is a better fit on a per-individual basis, we find that 89.7% of literal **AH** are better described by literal **H**, but in comparison, only 70.0% of pedagogic **AH** are better described by pedagogic **H**. If we simulate pedagogic humans as only following the model 70.0% of the time, then the accuracy of the pedagogic robot drops to 90% (Figure 5a). The literal robot is hardly impacted.

4.2 ACTION MIXTURE MODEL

We now test a more continuous version of the previous setting. We now model humans as acting according to a mixture policy H_m between the literal H_L and pedagogic H_P policies. In particular, at each time t the probability the human picks action a_t from state s in an environment with reward r is

$$H_M(a_t \mid s_t, r) = \\ \alpha H_P(a_t \mid s_t, r) + (1-\alpha) H_L(a_t \mid s_t, r)\,. \qquad (10)$$

The parameter α is the probability of picking an action according to the pedagogic model. We plot the likelihood of the actual pedagogic human demonstrations as a function of α, (10). The best value of α over the whole population was $\alpha_P = 0.5$ (Figure 6a). But surprisingly, even a mixture with $\alpha = 0.01$ or $\alpha = 0.99$ is far better than either the literal ($\alpha = 0$) or pedagogic ($\alpha = 1$) model. In addition, the best estimates for α on an individual basis are somewhat bimodal (Figure 6b), indicating that there is high individual variation.

Figure 5b shows the performamce of the literal and pedagogic robot as α varies. At the best population level $\alpha_P = 0.5$, we find that the pedagogic robot gets 97% accuracy. However, if we simulate the pedagogic humans using the values of α estimated at an individual level, then the pedagogic robot's accuracy drops to 90%, highlighting the importance of individual variation. The literal robot again remains unaffected.

4.3 DISCUSSION

In both the demonstration and action mixture models, we found that a mixture between pedagogic **H** and literal **H** was a much better fit to **AH**. In both cases, when pedagogic **H** is simulated according to the more accurate mixture model pedagogic **R**'s accuracy drops ten percentage points, but literal **R**'s accuracy hardly changes. Thus our results provide positive evidence for Hypothesis 4.1.

However, Hypothesis 4.1 does not explain the full story. The hypothesis can only account for a ten percentage drop in accuracy, but even with this drop, pedagogic **R** would still be better than literal **R**. Furthermore, with a cursory glance at Figure 5, one might be tempted to consider pedagogic **R** quite robust, as it remains high-performing for large ranges of α. However, as usual, the real problem is the *unknown unknowns*. Our empirical results imply that there are other ways that humans deviate from the model and that literal **R** is more robust than pedagogic **R** to these unknown deviations. Rather than robustness to α, the more compelling reason for choosing to use literal **R** is robustness to these unknown deviations.

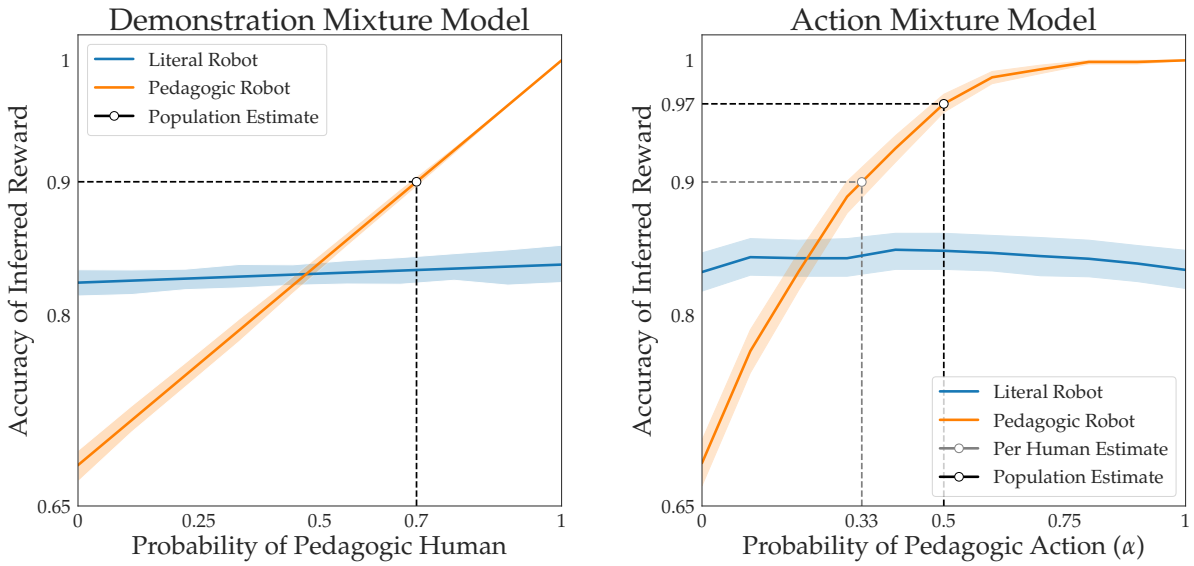

Figure 5: The performance of the literal and pedagogic robot when demonstrations are simulated according to the demonstration mixture model (left/a) or the action mixture model (right/b).

5 CAN WE "FIX" THE PEDAGOGIC ROBOT?

An intuitive idea for improving the performance of the pedagogic robot **R** is to give it a model of the pedagogic human **H** that is more predictive. Unfortunately, it is not so simple. For example, in Section 4.2, we showed that a mixture between the literal and pedagogic human model was a much better fit to actual pedagogic humans than either the literal or pedagogic human model. What if we had pedagogic **R** use the more accurate mixture model as its model of the pedagogic human? Unfortunately, as shown in Figure 5a (purple line), even if pedagogic **R** uses a better predictive model, i.e. the action mixture model, it does not perform any better.

In fact, it is possible for pedagogic **R** to do *worse* when given a *more* predictive model of the human. The reason is that a model that is better for *predicting* behavior (e.g. human demonstrations) is not necessarily better for *inferring* underlying, latent variables (e.g. the reward function), which is what is important for the robot. This may help explain why pedagogic **R** does worse than literal **R** with pedagogic **AH**, even though the pedagogic **H** model assumed by pedagogic **R** is more predictive of pedagogic **AH** than the literal **H** model assumed by literal **R**.

To illustrate, suppose there is a latent variable $\theta \in \Theta$ with prior distribution $p(\theta)$ and observed data $x \in \mathcal{X}$ generated by some distribution $p(x \mid \theta)$. In our setting, θ corresponds to the objective and x corresponds to the human input. For simplicity, we assume Θ and $\mathcal{X}$ are finite. We have access to a training dataset $\mathcal{D} = \{(\theta_i, x_i)\}_{i=1}^n$ of size n. A *predictive model* $m(x \mid \theta)$ models the conditional probability of the data x given latent variable θ for all $x \in \mathcal{X}, \theta \in \Theta$. In our case, the predictive model is the model of the human. The *predictive likelihood* $\mathcal{L}_{\mathcal{X}}$ of a predictive model m is simply the likelihood of the data under the model:

$$\mathcal{L}_{\mathcal{X}}(m) = \prod_{i=1}^{n} m(x_i \mid \theta_i)\,. \tag{11}$$

The *inferential likelihood* is the likelihood of the latent variables after applying Bayes' rule:

$$\mathcal{L}_{\Theta}(m) = \prod_{i=1}^{n} \frac{m(x_i \mid \theta_i)p(\theta_i)}{\sum_{\theta} m(x_i \mid \theta)p(\theta)}\,. \tag{12}$$

Next, we show higher predictive likelihood does not necessarily imply higher inferential likelihood.

Claim 5.1 (Predictive vs inferential likelihood)**.** *There exist settings in which there are two predictive models m_1, m_2 such that $\mathcal{L}_{\mathcal{X}}(m_1) > \mathcal{L}_{\mathcal{X}}(m_2)$, but $\mathcal{L}_{\Theta}(m_1) < \mathcal{L}_{\Theta}(m_2)$.*

Proof. Suppose that $\Theta = \{\theta_1, \theta_2\}$ and $\mathcal{X} = \{x_1, x_2, x_3\}$, the prior $p(\theta)$ is uniform over Θ, and the

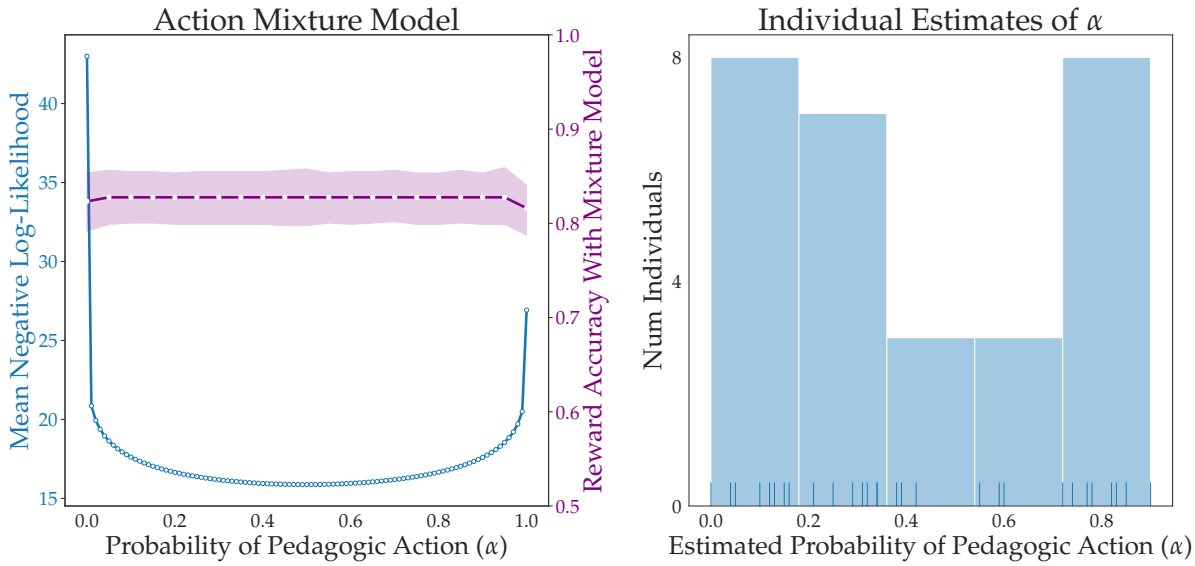

Figure 6: (left/a) In blue is the mean negative log-likelihood of the pedagogic human demonstrations under the action mixture model with probability α. All individuals are assumed to have the same value of α. Note that a mixture between the literal and pedagogic model is far better than either the literal model ($\alpha = 0$) or the pedagogic model ($\alpha = 1$). In purple is the accuracy of the robot's inferred reward, when given demonstrations from pedagogic **AH**, if it assumes the human acts according to the action mixture model. (right/b) The mixture probability α that maximized log-likelihood for each individual pedagogic **AH**.

dataset $\mathcal{D}$ contains the following $n = 9$ items:

$$\mathcal{D} = \{(\theta_1, x_1), (\theta_1, x_1), (\theta_1, x_2), (\theta_2, x_2), (\theta_2, x_2), (\theta_2, x_3), (\theta_2, x_3), (\theta_2, x_3)\}.$$

Define the models $m_1(x \mid \theta)$ and $m_2(x \mid \theta)$ by the following conditional probabilities tables.

$m_1(x \mid \theta)$

	x_1	x_2	x_3
θ_1	2/3	1/3	0
θ_2	0	1/3	2/3

$m_2(x \mid \theta)$

	x_1	x_2	x_3
θ_1	2/3	1/3	0
θ_2	0	2/3	1/3

The model m_1 has predictive likelihood $\mathcal{L}_{\mathcal{X}}(m_1) = (2/3)^6(1/3)^3$ and inferential likelihood $\mathcal{L}_{\Theta}(m_1) = (1/2)^3$. The model m_2 has predictive likelihood $\mathcal{L}_{\mathcal{X}}(m_2) = (2/3)^4(1/3)^5$ and inferential likelihood $\mathcal{L}_{\Theta}(m_2) = (1/3)(2/3)^3$. Thus, $\mathcal{L}_{\mathcal{X}}(m_1) > \mathcal{L}_{\mathcal{X}}(m_2)$, but $\mathcal{L}_{\Theta}(m_1) < \mathcal{L}_{\Theta}(m_2)$. □

In our context, Claim 5.1 means that a human model that is better in terms of *prediction* may actually be worse for the robot to use for *inference*. An alternative approach could be to directly fit models that optimize for inferential likelihood. Unfortunately, this optimization becomes much trickier because of the normalization over the latent variable space Θ in the denominator of (12), see e.g. (Dragan et al., 2013a).

Acknowledgements

We thank John Miller, Rohin Shah, and Mark Ho for providing feedback on a draft of the paper. In addition, we thank Mark Ho for his helpfulness in answering any questions about his work, for sharing his code, and for points of discussion along the way.

The work in this paper was supported by the National Science Foundation National Robotics Initiative and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement

learning from human preferences. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Jack Clark and Dario Amodei. Faulty reward functions in the wild, Mar 2017. URL https://blog.openai.com/faulty-reward-functions/.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *Conference on Human-Robot Interaction (HRI)*, 2013a.

Anca D. Dragan, Siddhartha S. Srinivasa, and Kenton C. T. Lee. Teleoperation with intelligent and customizable interfaces. *Journal of Human-Robot Interaction*, 2013b.

Jaime Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Mennell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. Pragmatic-pedagogic value alignment. In *International Symposium on Robotics Research (ISRR)*, 2018.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, pp. 3909–3917, 2016.

Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. In *Neural Information Processing Systems (NeurIPS)*, 2016.

Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. Effectively learning from pedagogical demonstrations. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2018.

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

Victoria Krakovna. Specification gaming examples in AI, Jun 2018. URL https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Belson, David M Bryson, Nick Cheney, 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.

Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. *International Conference on Machine Learning (ICML)*, 2012.

Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. An efficient, generalized Bellman update for cooperative inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

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.

Pei Wang, Pushpi Paranamana, and Patrick Shafto. Generalizing the theory of cooperative inference. *AISTATS*, 2019.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research (JMLR)*, 2017.

Scott Cheng-Hsin Yang, Yue Yu, Arash Givchi, Pei Wang, Wai Keen Vong, and Patrick Shafto. Optimal cooperative inference. *AISTATS*, 2018.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.