How to talk to your model: Instructions, descriptions, and learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

From the earliest years of our lives, humans use language to express our beliefs and desires. Being able to talk to artificial agents about our preferences would thus fulfill a central goal of value alignment. Yet today, we lack computational models explaining such flexible and abstract language use. To address this challenge, we consider social learning in a linear bandit setting and ask how a human might communicate preferences over behaviors (i.e. the reward function). We study two distinct types of language: *instructions*, which ground to concrete actions, and *descriptions*, which provide information about the reward function. To explain how humans use these forms of language, we suggest they reason about both known *present* and unknown *future* states: instructions optimize for the present, while descriptions generalize to the future. We formalize this choice by extending reward design to consider a distribution over states. We then define a pragmatic listener agent that infers the speaker's reward function by reasoning about *how* the speaker expresses themselves. We validate our models with a behavioral experiment, demonstrating that (1) our speaker model predicts spontaneous human behavior, and (2) our pragmatic listener is able to recover their reward functions. Finally, we show that in traditional reinforcement learning settings, pragmatic social learning can integrate with and accelerate individual learning. Our findings suggest that social learning from a wider range of language—in particular, expanding the field's present focus on instructions to include learning from descriptions—is a promising approach for both value alignment and reinforcement learning more broadly.

1 Introduction

As artificial agents proliferate in society, aligning them with human values is increasingly important [1–3]. But how can we build machines that understand what we want? Prior work has highlighted the difficulty of specifying our desires in numerical reward functions [3-5]. Here, we explore language as a means to communicate them. While most previous work on language input to AI systems focuses on *instructions* [6-19]—we study instructions alongside more abstract, *descriptive* language [20-24]. We examine how humans communicate about rewards, and formalize learning from this input.

To consider how humans communicate about reward functions, imagine taking up mushroom foraging. How would you learn the rewards associated with different fungi (i.e. which are delicious and which are deadly)? In such a setting, learning from direct experience [25] is risky; most humans would seek to learn *socially* instead. How might we learn reward functions from others? Prior work in reinforcement learning (RL) has examined a number of social learning strategies, including

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

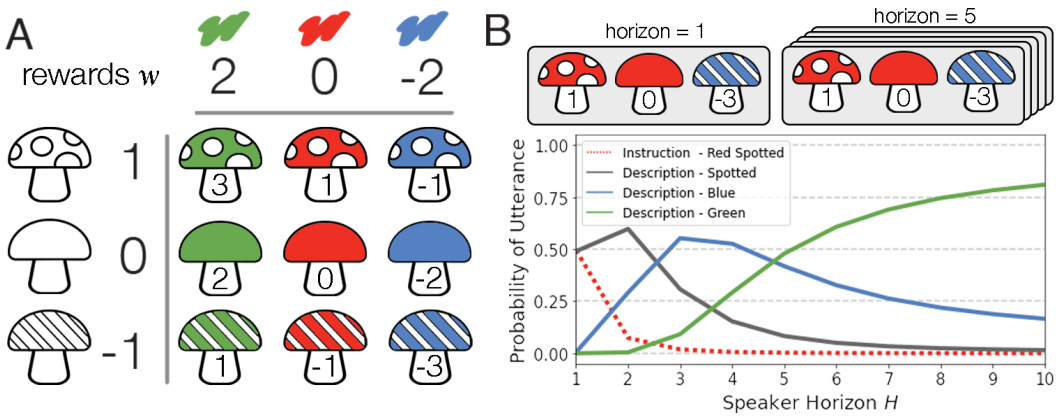

Figure 1: **A**: Our linear bandit setting. Rewards associated with features determine whether actions (mushrooms) are high or low reward. **B**: The speaker's *horizon* affects how they communicate about the reward function. At short horizons, speakers prefer instructions (which precisely select the optimal action). At longer horizons encompassing more unknown states, speakers prefer descriptions (which generalize better).

passive *inverse reinforcement learning* (observe an expert pick mushrooms, then infer their reward function [26, 27]) or active *preference learning* (offer an expert pairs of mushrooms, observe which one they eat, and infer their reward function [28–30]).

However, few humans would rely on such indirect data if they had access to a cooperative teacher [31–34]. For example, an expert guiding a foraging trip might *demonstrate* [35, 36] or verbally *instruct* [7] the learner to pick certain mushrooms, licensing stronger inferences. While such pedagogical actions have been useful for guiding RL agents [10, 17–19], natural language affords richer, under-explored forms of teaching. For example, an expert teaching a seminar might *describe* how to recognize edible or toxic mushrooms based on their features.[1] Descriptive language is particularly powerful if learners can expect experts to prioritize *relevant* and *context-sensitive* information [38, 39].

First, we present a formal model of cooperative social learning in a linear bandit setting (Fig 1). We propose a speaker model that chooses utterances to maximize the listener's expected rewards over some task *horizon*, generalizing reward design [40] to a distribution over states. We formalize and analyze *instructions* (which designate a specific action to take) and *descriptions* (which teach the reward function). We find that short-horizon speakers (focused on a single, known state) prefer instructions, while long-horizon speakers (reasoning about future, unknown states) prefer descriptions.

Second, we consider how a listener might learn from such a speaker. We define a pragmatic listener that performs *inverse* reward design [IRD, 4] to learn about rewards from instructions and descriptions. While prior work suggests pragmatic learning can be vulnerable to model mis-specification [41], we show that jointly inferring the speaker's horizon and reward function can mitigate this risk.

Finally, we conduct a behavioral experiment showing our models support strong reward inference from human-chosen utterances. We integrate this social inference with traditional RL, with social information accelerating learning and reducing regret. Overall, our results suggest that descriptive language and pragmatic inference are powerful mechanisms for value alignment and learning.

2 Related work

Classic RL assumes that the reward function is given to the agent [25]. However, in practice, it is difficult to specify a reward function to obtain desired behavior [3], motivating *learning* the reward

[1] Or write a book on the topic, e.g. [37].

function from social input.[2] Methods for learning reward functions often simply assume the expert is simply acting optimally, but recent *pragmatic* methods instead assume the expert is actively teaching.

Learning reward functions from observed actions. When the desired behavior is known (but the reward function is not), inverse reinforcement learning [IRL, 12, 14, 26, 27, 45–47] can be used to infer an expert's reward function from their actions. However, such approaches face fundamental issues with identifiability: passively observed behavior can often be explained by multiple reward functions. This ambiguity can be addressed, for instance, by allowing the robot to actively query the human [29, 30, 48]. After reviewing work on learning rewards from language, we will turn to another important class of IRL approaches based on *pragmatics* [35, 41, 49, 50].

Learning reward functions from instructions. Instructions use language to communicate specific actions or goals [6–8]. Prior work in reinforcement learning has recognized that instructions can reflect more general latent desires, and used them to shape [15–17] or infer reward functions [9–14]. Our work provides a theoretical basis for these models, situating them within a broader framework of humans' language use: we ask when a human would want to use instructions (vs. other forms of language), and—given an instruction (vs. other forms of language)—how a robot should generalize.

Learning reward functions from descriptions. A relatively smaller body of work seeks to learn reward functions directly from existing bodies of text [20–22, 51] or interactive, free-form language input [23, 24]. While reward-descriptive language directly encodes generalized information, related lines of work use other "descriptive" language, typically instructions or trajectory annotations, to assist in (hierarchical) RL. Rather than learn the reward function, these approaches use externally provided [52–58] or internally-generated language [59–62] to help structure the task representation or guide exploration. While the education literature suggests rich descriptive feedback is crucial for human teaching [63, 64], it poses a severe grounding challenge [65, 66] as the language no longer maps directly to the agent's action or perception space. Here, we assume groundings are known; this allows us to characterize the utility of reward-descriptive language to gain theoretical understanding.

Learning reward functions from pedagogy. The previous algorithms we reviewed all assume that training examples (whether demonstrations or language) are generated by a human that is *indifferent to how they learn*. Recent work has begun to challenge this assumption by considering *pedagogical* settings [33–36]. In particular, the Rational Speech Act framework [67, 68] builds on classic Gricean theories [69] to formulate *optimal communication* in terms of recursive Bayesian inference and decision-making. These ideas have been applied to a variety of language tasks including reference games [70, 71], captioning [72], and instruction following [73].

Developing analogues of (linguistic) pragmatics in reinforcement learning — i.e., algorithms that assume data are *intentionally designed to be informative* — is currently an active area of research [35, 41, 49, 50]. In particular, inverse reward design (IRD [4]) applies pragmatic inference directly to (numerical) reward functions: rather than take the provided reward function literally, the IRD agent quantifies its uncertainty over the function to mitigate alignment risk. In this work, we unify these different strains of pragmatic reasoning, extending IRD to a broader class of natural language input.

3 Social learning as reward design

3.1 Pragmatic reasoning

Our theoretical approach is based on the Rational Speech Act framework [67, 68], which formalizes pragmatics as recursive reasoning. It begins by defining a *literal listener* L_0 who interprets utterances using the literal meanings of the language. A *pragmatic speaker* S_1 then (soft-)optimally selects utterances u to maximize a utility, typically the epistemic goal of increasing the literal listener's beliefs in the true world state w^*: $U(u; w^*) = \ln P_{L_0}(w^* \mid u)$. A *pragmatic listener* L_1 then infers

[2] Another approach bypasses learning the reward function and allows the human to provide a reward signal directly [42, 43]. This is challenging, however, as people do not typically provide classical reinforcement [44].

the intended meaning of a given utterance by assuming it was generated by the speaker:

$$P_{L_0}(w \mid u) \propto \mathcal{L}(w, u) \tag{1}$$

$$P_{S_1}(u \mid w) \propto \exp\{\beta_{S_1} \cdot U(u; w)\} \tag{2}$$

$$P_{L_1}(w \mid u) \propto P_{S_1}(u \mid w) P(w) \tag{3}$$

where $\mathcal{L}$ is a "lexicon" function mapping utterances to states and $\beta_{S_1} > 0$ is an inverse temperature parameter. This recursion can be continued [74], but the behavior of interest emerges at the L_1 level.

3.2 Social learning in linear bandits

We consider social learning in a *linear bandit* setting, a popular testbed for RL [75–79]. Formally, we define a set of A possible actions. Actions are associated with a binary feature vector $\phi : A \to \{0, 1\}^K$ (e.g. a mushroom may be green (or not), or striped (or not)). The rewards of each action are then defined as a linear combination of these features (e.g. green mushrooms tend to be tasty; this is a common assumption in IRL [4, 26, 46]):

$$R(a, w) = w^\top \phi(a) \tag{4}$$

so w is a vector that defines the value of each feature (see Fig. 1A). Each task consists of a sequence of H i.i.d. states. At each time step $t < H$, the agent is presented with a state s_t consisting of a subset of possible actions: $s_t \subseteq A$ (e.g., a particular mushroom patch). They choose an action $a \in s_t$ according to their policy, $\pi_L : S \to \Delta(A)$.

While the bandit problem is typically considered an individual learning problem, we instead ask how an agent should learn from a cooperative, knowledgeable partner. We formalize this *social* learning problem by introducing a second agent: a speaker who knows the true rewards w and the initial state s_0, and produces an utterance u. The listener updates their policy to $\pi_L(a \mid u, s)$ before beginning to choose actions. Intuitively, the horizon H determines how many actions the listener will perform under this updated policy. $H = 1$ represents maximum supervision (i.e. guided foraging), whereas $H \to \infty$ is minimal supervision (teaching the listener to forage independently in future settings). We first assume H is known to both listener and speaker, then relax this assumption. This framework exposes two interrelated problems. First, what should a helpful speaker say? And second, how should the listener update their policy in light of this information?

3.3 Speakers as reward designers

Rather than defining the speaker's utility $U(u; w)$ as Gricean informativeness [69] (i.e. inducing true beliefs, Section 3.1), we suggest that a cooperative speaker should *maximize the listener's rewards*, thus grounding utility in the listener's subsequent actions. When the state is known, the *present* utility of an utterance is the expected reward from using the resulting policy to choose an action:

$$U_{\text{Present}}(u \mid s, w) = \sum_{a \in s} \pi_L(a \mid u, s) R(a, w) \tag{5}$$

This formulation is equivalent to *reward design* [4, 40], where the reward designer chooses a proxy reward for a single, known MDP. However, because only the first state is known, we must also consider how well the policy *generalizes*. Thus, unlike reward design, speakers may reason about future states. The *future* utility of an utterance with respect to a distribution over states $P(s)$ is:

$$U_{\text{Future}}(u \mid w) = \sum_{s \in S} U_{\text{Present}}(u \mid s, w) P(s) \tag{6}$$

Because states are i.i.d. in the bandit setting, a speaker optimizing for a horizon H can be defined as a linear combination of Eqs. 5 and 6:

$$U_{S_1}(u \mid w, s, H) = \frac{U_{\text{Present}} + (H-1) U_{\text{Future}}}{H} = \frac{1}{H} U_{\text{Present}} + (1 - \frac{1}{H}) U_{\text{Future}} \tag{7}$$

where $H = 1$ reduces to Eq. 5 and as $H \to \infty$ reduces to Eq. 6. The following theorem shows that increasing the speaker horizon increases the expected utility of the listener across the distribution of future states. This is because the speaker's incentive to bias the proxy reward to the local context is reduced as the horizon they consider grows.

Theorem 1 (Reward design over longer horizons)**.** *Let S be a set of states that each represent a decision context. Let π_L be an arbitrary listener policy. Consider the behavior of a speaker that chooses utterances u based on Eq. 7. As $H \to \infty$, the future utility generated by π_L increases.*

The proof for this and other theorems can be found in Appendix A.

3.4 Formalizing instructions

We now consider how utterances affect the listener's policy. *Instructions* map to specific actions or trajectories [7, 12]; in our work, "instruction" utterances correspond to the nine actions (Fig 1A). Given an instruction, a literal listener executes the corresponding action. If the action is not available, the listener acts randomly:

$$\pi_{L_0}(a \mid u, s) = \begin{cases} 0 & \text{if } a \notin s \\ \delta_{[\![u]\!](a)} & \text{if } [\![u]\!] \in s \\ \frac{1}{|s|} & \text{otherwise} \end{cases} \tag{8}$$

where $\delta_{[\![u]\!](a)}$ represents the meaning of u, evaluating to one when utterance u grounds to a and zero otherwise.[3] An instruction is a *partial policy*: it designates the correct action in a subset of states.

This formalization of instructions allows us to identify an overlap between pragmatic inference from demonstrations, instructions, and reward design. In learning from demonstrations with, e.g., IRL, we assume that demonstrations are drawn from a distribution of optimal behavior. Thus, inference identifies reward functions such that the demonstrated trajectories have high reward in comparison to alternative trajectories available to the demonstrator. We will show that this is the same as: (1) pragmatic inference over instructions, under the assumption that instructions have a one-to-one mapping to available trajectories; and (2) pragmatic inference over true reward given a proxy, under the assumption that the set of possible proxy rewards has a one-to-one mapping with trajectories. This reduction allows us to connect inference across these modalities and apply the results from, e.g., Theorem 1 to show that with a longer speaker horizon improves expected listener performance.

Theorem 2 (Reward design with instructions is equivalent to demonstrations.)**.** *Let s be a state, represented as a local context with a set of actions $a \in s$ that can be taken. Let $\mathcal{U}_{instruct}$ be a set of instruction utterances that reference each $a \in s$ and let $\widetilde{\mathcal{R}}$ be the following set of (proxy) reward functions $\{R(a) = \mathbb{I}[a = a'] | a' \in s\}$, where $\mathbb{I}$ represents the indicator function. Then, the posterior distribution obtained after an observation of noisily-optimal behavior is the same as that obtained from a speaker maximizing Eq. (5) and that obtained from IRD with the set of proxies $\widetilde{\mathcal{R}}$.*

3.5 Formalizing descriptions

Rather than mapping to a specific action, descriptions provide information about the reward function [23, 24, 47]. Following [80], we model descriptions as providing the reward of a single feature, similar to feature queries [30]. Descriptions are thus a tuple: a one-hot binary feature vector and a scalar value, $\langle \mathbb{1}_K, \mathbb{R} \rangle$. These are messages like ⟨Blue, -2⟩. In this work, we consider the set of 6 features $\times$ 5 values in $[-2, -1, 0, 1, 2]$, yielding 30 descriptive utterances. Formally, L_0 "rules out" inconsistent hypotheses about reward weights w:

$$L_0(w \mid u_{\text{description}}) \propto \delta_{[\![u]\!](w)} P(w) \tag{9}$$

where $\delta_{[\![u]\!](w)}$ represents the meaning of u, evaluating to one when u is true of w and zero otherwise. In this work, we assume $P(w)$ is uniform and there is no correlation between weights. The listener then marginalizes over possible reward functions to choose an action:

$$\pi_{L_0}(a \mid u, s) \propto \exp\{\beta_{L_0} \cdot \sum_w R(a, w) L_0(w \mid u))\} \tag{10}$$

where β_{L_0} is again an inverse temperature parameter.[4]

[3] We assume that lexical groundings $\mathcal{L}$ are known, i.e. the literal listener knows the meaning of utterances.

[4] Due to the recursive structure of speaker and listener, simultaneously varying β_{S_1} and β_{L_0} introduces identifiability issues. We therefore fix $\beta_{L_0} = 3$ throughout this work, and only vary β_{S_1}. This implies that all speakers address the same listener, but do so with different degrees of optimality.

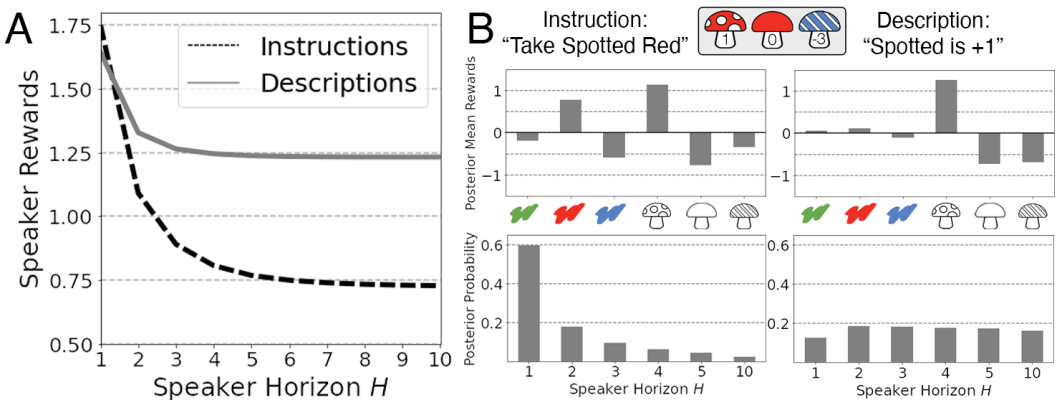

Figure 2: **A**: Speaker rewards (Eq. 7, averaged over all 84 start states) for a speaker with access to only instructions or only descriptions. Instructions outperform descriptions at short horizons, but descriptions offer much stronger generalization. **B**: Latent-horizon pragmatics in this start state (Eq. 12). While both utterances suggest the listener should take the spotted red mushroom, an instruction (left) suggests they are short-horizon while a description (right) suggests they are long-horizon. This leads to different inference about their reward function (see Appendix B for full posteriors and more examples).

3.6 Comparing instructions and descriptions

Prior work suggests that humans use a mix of instructive and descriptive language [23, 24]. What modulates this—when should a rational speaker prefer instructions over descriptions? To explore the effects of speaker horizons on choice of utterance, we simulate a nearly-optimal speaker ($\beta_{S_1} = 10$). Fig. 1A shows our bandit setting. We assume the listener begins with a uniform prior over reward weights throughout, and use states consisting of three unique actions (giving 84 possible states).

Fig 1B shows how the speaker's choice of utterance varies as their horizon lengthens. At short horizons, the speaker optimizes for present rewards (Eq. 5) and chooses instructions and descriptions that target the "Spotted Red" action. At longer horizons, future rewards (Eq. 6) play a larger role, and the speaker blends the two objectives (Eq. 7) by describing the highly negative blue feature. Finally, at sufficiently long horizons, future rewards dominate and it settles on describing the green feature—which is irrelevant to the start state, but the most important feature for generalization. To quantify how rational speakers should use instructions and descriptions, we repeat the task for all 84 start states using horizons ranging 1-10 and different utterance sets. Fig 2A plots rewards for speakers with access to *only* instructions *or* descriptions, illustrating why this shift from instructions to descriptions occurs. Instructions outperform descriptions at short horizons (achieving the theoretical maximum average reward of 1.75); as the horizon lengthens, however, descriptions generalize much better.[5]. Overall, as their horizon lengthens, speakers with access to both instructions and descriptions choose descriptions exclusively, producing highly generalizable information (see Appendix B).

3.7 Learning from Utterances: a Pragmatic Listener

We now ask how the listener should *learn* from the speaker's utterance, using pragmatic inference to recover information about the speaker's reward function.

Known Horizon. Following the standard RSA formulation, a pragmatic listener L_1 can invert the speaker model. If the speaker's horizon H is known, this is equivalent to inverse reward design [4]:

$$L_1(w \mid s, u, H) \propto S_1(u \mid w, s, H)P(w) \tag{11}$$

Given an instruction, L_1 infers the reward weights that would make such an instruction optimal [9, 14]; given a description, L_1 can recover information about features that were not mentioned [23]. The L_1

[5]Because our states consist of only three actions, it is often possible to find a description that uniquely identifies the best action. This allows descriptions to perform nearly as well as instructions at $H = 1$. However, as the number of available increases, instructions become increasingly advantageous; see Appendix B.

listener then chooses actions by substituting this posterior belief into Eq. 10. In practice, however, the speaker's horizon is *not* known, and so this approach is not feasible for real-world applications. This model thus serves as a theoretically-optimal baseline, and we next consider three practical ways of handling this uncertainty.

Assuming the horizon. Perhaps the most straightforward approach is to simply assume a speaker horizon. However, prior work has highlighted the risks of assuming a human is actively trying to teach [41], and suggests that the safest approach is to assume they are not. To test the effects of such mis-specification, in Section 4 we use two pragmatic listener models which assume a short ($H = 1$) or long ($H = 4$) horizon. A pragmatic listener assuming $H = 1$ will *constrain* inference: it will assume the utterance reflects only the speaker's preference between actions in the present state, and thus generalize conservatively. In contrast, a pragmatic listener assuming $H \gg 1$ expects the utterance to generalize broadly, risking *overfitting*.

Inferring the horizon. To mitigate the risk of horizon misspecification, we can instead assume the speaker's horizon is unknown. Given an utterance, the *latent-horizon* pragmatic listener jointly infers both their horizon and rewards, then marginalizes out the horizon:

$$L_1(w \mid s, u) \propto \sum_H S_1(u \mid w, s, H) P(H) P(w) \tag{12}$$

Intuitively, because long-horizon speakers prefer instructions (and short-horizon speakers prefer instructions, Fig. 1B), the latent-horizon pragmatic listener can use the utterance type to infer the speaker's horizon and determine the appropriate scope of generalization. To demonstrate this, we simulate a pragmatic listener with a uniform prior over $H \in [1, 2, 3, 4, 5, 10]$. Fig. 2B shows example utterances and resulting inference about the speaker's rewards and latent horizon. Crucially, while both example utterances indicate a preference for the spotted red mushroom, the description suggests the speaker is $H > 1$ and uniquely identifies the spotted feature as high-value.

4 Behavioral Experiment

To validate our theoretical models, we collected a behavioral dataset. Participants played the role of a mushroom foraging guide and produced utterances for tourists to help them choose good mushrooms. We manipulated the speaker's horizon by varying tourists' itineraries: each tourist was shown visiting a different visible mushroom patch, plus a variable number of unknown future patches (0, 1, or 3, matching horizons 1, 2, and 4 respectively). In the following sections, we analyze the participants' choice of utterances, compare the resulting inference from different listener models, and show how this socially-learned information can accelerate traditional reinforcement learning.

4.1 Experiment Setup

We recruited 119 participants on the Prolific experimental platform (`prolific.co`). Participants were trained and tested on the game dynamics then advised a total of 28 tourists. They were told to consider the tourists' itinerary and choose "the most helpful utterance" from drop-down menus allowing them to specify an instruction or description. Because the space of possible descriptions (30) is larger than possible instructions (9), choosing a description required more effort than choosing an instruction. To equalize this, we reduced the set of descriptions by removing neutral features ("red" and "solid") and the 0 value, yielding $4 \times 4 = 16$ possible descriptive utterances. For the remainder of the paper, all pragmatic inference listeners assume the speaker chooses from this reduced utterance set. Participants received no feedback on how their utterances affected tourists' behaviors, ensuring they chose utterances according to their own sense of how to help. After screening out participants who failed comprehension or attention checks, we were left with 99 participants who produced a total of 2772 utterances. For more details on the data collection, see Appendix C.

We next use this set of utterances to explore (1) exclusively social learning (i.e. whether our models can infer the speakers' reward function), and then (2) integrated social and reinforcement learning (i.e. whether this socially-learned information can reduce regret in traditional RL).

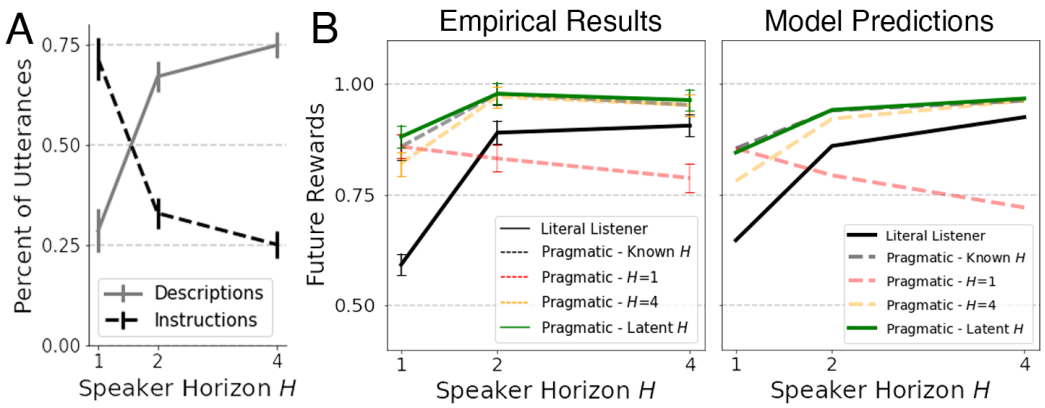

Figure 3: Results from behavioral experiment. **A**: Participants preferred instructions when $H = 1$ and descriptions otherwise. **B**, left: empirical future rewards (Eq. 6) from human utterances. As the horizon lengthens, humans choose more generalizable utterances; our latent-horizon listener successfully recovers their reward function. Right: simulations using our speaker model with $\beta_{S_1} = 3$. All error bars show 95% CI.

4.2 Inferring rewards: social learning from human utterances

Overall, our results validated our theoretical models. First, participants were sensitive to the horizon manipulation: there was a statistically-significant shift in their utterance choices between horizons (for $H = 1$ to 2, $\chi^2(26, 1856) = 336.1, p < .001$; for $H = 2$ to 4, $\chi^2(26, 1833) = 40.3, p = .04$). Almost all participants (96 out of 99) used a mix of instructions and descriptions, favoring instructions at $H = 1$ and descriptions at $H > 1$ (Fig. 3A). This led to lower (literal) future rewards when $H = 1$, and higher future rewards at longer horizons (Fig 3B, "Empirical Results"). To calibrate our pragmatic listeners, we tested $\beta_{S_1} \in [1, 10]$ and found that $\beta_{S_1} = 3$ optimized both Known H and Latent H listeners. Simulating utterances produced by our speaker model and resulting pragmatic inference a match to theoretical predictions (Fig. 3B, "Model Predictions", see Appendix C for details).

Our pragmatic listeners offered statistically-significant improvements over the literal listener (Table 1). These gains were particularly large when the speaker has a short horizon, matching our simulations. Somewhat surprisingly, our Latent H model ($M = 0.94$, $SD = 0.39$) outperformed all other models, including the Known H model ($M = .93$, $SD = .40$), by a significant albeit small margin (mean difference 0.01, paired-samples t-test $t(2771) = 4.18, p < .001$; see Appendix D for other pairwise tests). This is particularly notable because it underscores the inevitability of model misspecification [41]: even when we experimentally controlled the horizon, participants did not perfectly follow our theoretical models, leading our Latent H model to outperform. We also confirmed the risks of *assuming* a horizon: the two fixed-horizon listeners ($H = 1$ and $H = 4$) underperformed the Latent H listener.

Despite these successes, we found a notable discrepancy between theoretical predictions and empirical results, suggesting a possible future refinement. As described in [80], utility-maximizing speakers agnostic to the truth of an utterance may send messages wtih exaggerated values (e.g. preferring the false utterance ⟨Spotted, +2⟩ to the true utterance ⟨Spotted, +1⟩). In our experiment, however, participants regularly chose the lower-reward, true utterance (Appendix C). Intriguingly, breaking out results by utterance type indicates that pragmatic gains come primarily on instructions, with pragmatic models on average faring slightly *worse* than the literal on descriptive utterances (Appendix E). This suggests that our reward-design objective is in fact too weak: participants' tendency to use true descriptions licenses stronger inference, which we return to in the discussion.

4.3 Why learn socially? Reinforcement learning with social priors

We now explore how this form of social learning could augment traditional RL approaches. In particular, we take the inferences from our listeners in the previous section and use them as a prior

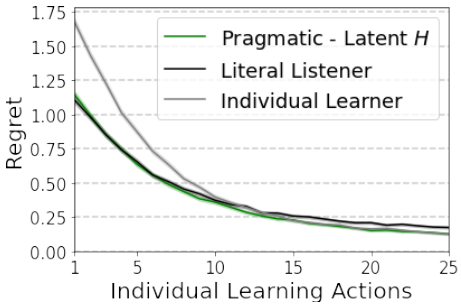

Figure 4: Regret with and without social information from our behavioral experiment.

Listener	Future rewards from utterance	Regret after 25 actions
Individual	-	12.14
Literal	.79	10.23
Prag - Known H	.93	**9.42***
Prag - $H = 1$	.83	9.76
Prag - $H = 4$	.91	9.67
Prag - Latent H	**.94*****	9.55

Table 1: Results from human utterances. Bolded results significantly better than others (see Appendix D for statistical testing).

for reinforcement learning (or in the case of literal instructions, we integrate them into the learner's policy; see Appendix F for details on the experimental setup). We then use Thompson sampling [81–83] to learn the reward function and compare regret over the course of learning for five independent learning trials on each context-utterance pair. We report results for our listener agents, as well as an agent with no social information (the "Individual" agent; Fig. 4, Table 1).

We find that social information—and in particular, pragmatic social learning—can substantially reduce regret. Intriguingly, when comparing regret, our Known H pragmatic listener now achieves the best results; this is likely due to the greater uncertainty inherent in the Latent H inference. To confirm these differences are statistically significant, we use a linear regression to predict regret with a fixed effect of listener and random effects for each of the 2772 utterance-context pairs, again comparing the Latent H listener to all other listeners. This confirms that the Known H listener achieves lower regret than the Latent H listener ($\beta = -.12$, $t(80400) = -2.14, p = .03$), and all other models suffer higher regret (see Appendix D). Finally, while the literal listener achieves lower regret initially than the pure reinforcement learner, it subsequently asymptotes as it rigidly obeys instructions (similar to the issue raised in [14]). This suggests that while instructions provide useful guidance early in learning (as they help the agent choose and learn from good actions), a more effective instruction-following strategy could be to gradually shift to belief-based actions.

5 Discussion

We introduced a unifying model of communication as reward design [40] to explain humans' use of instructions and descriptions, allowing pragmatic inference of their reward functions—a critical capability for value alignment [3, 49, 84]. Analyses show that instructions are optimal at short horizons, but descriptive language affords much stronger generalization. This allows our pragmatic listener to perform *inverse* reward design [4] to jointly infer the speaker's horizon and reward function, reducing risk of model mis-specification in pragmatic inference [41]. Finally, our behavioral experiment shows that humans follow our theoretical predictions, demonstrating the benefits of pragmatic inference for value alignment and accelerating traditional reinforcement learning.

We note several limitations and future directions. First, our behavioral experiment suggests possible improvements to our speaker model: participants displayed a bias towards *truthful* descriptions. This licenses stronger pragmatic inference [85] such as enriching our model with traditional RSA objectives [67, 68], or reformulating it to minimize regret during individual learning [86]. Second, for theoretical clarity and experimental feasibility, we used a simple bandit setting and pre-defined utterances. Future work could address more challenging environments and open-ended language.

More generally, our work provides a theoretical basis for further work at the intersection of social and reinforcement learning. Linear bandits are already a popular testbed for RL [75, 77]. We hope that our formalization of social learning in linear bandits supports more work in this setting [23, 24]. Finally, while most previous work focuses on instructions, our results highlight the power of descriptions for generalization, illustrating the potential of learning from richer forms of language.

References

[1] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437, September 2020.

[2] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, USA, 2019.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems*, 2017.

[5] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *arXiv preprint arXiv:2104.13906*, 2021.

[6] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[7] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*, 2011.

[8] David Chen and Raymond Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI Conference on Artificial Intelligence*, 2011.

[9] 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.

[10] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. In *International Conference on Learning Representations*, 2018.

[11] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Seyed Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. Learning to understand goal specifications by modelling reward. In *International Conference on Learning Representations*, 2019.

[12] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 2020.

[13] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*, 2022.

[14] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should robots be obedient? In *International Joint Conference on Artificial Intelligence*, 2017.

[15] Richard Maclin and Jude W. Shavlik. Incorporating advice into agents that learn from reinforcements. In *AAAI Conference on Artificial Intelligence*, 1994.

[16] Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. *AAAI Workshop - Technical Report*, 06 2004.

[17] Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2019.

[18] J Luketina, N Nardelli, G Farquhar, J Foerster, J Andreas, E Grefenstette, S Whiteson, and T Rocktäschel. A survey of reinforcement learning informed by natural language. In *International Joint Conference on Artificial Intelligence*, 2019.

[19] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, 2020.

[20] Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. Grounding language for transfer in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 63:849–874, 2018.

[21] Victor Zhong, Tim Rocktäschel, and Edward Grefenstette. RTFM: Generalising to novel environment dynamics via reading. *arXiv preprint arXiv:1910.08210*, 2019.

[22] Austin W Hanjie, Victor Y Zhong, and Karthik Narasimhan. Grounding language to entities and dynamics for generalization in reinforcement learning. In *International Conference on Machine Learning*, 2021.

[23] Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. Learning rewards from linguistic feedback. In *AAAI Conference on Artificial Intelligence*, 2021.

[24] Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. Inferring rewards from language in context. *arXiv preprint arXiv:2204.02515*, 2022.

[25] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[26] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

[27] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.

[28] Douglas B Markant and Todd M Gureckis. Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, 143(1):94, 2014.

[29] 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*, 2017.

[30] Chandrayee Basu, Mukesh Singhal, and Anca D Dragan. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2018.

[31] Natalia Vélez and Hyowon Gweon. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, 38: 110–115, 2021.

[32] Hyowon Gweon. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, 25(10):896–910, 2021.

[33] Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 2020.

[34] Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55 – 89, 2014.

[35] Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus Doing: Teaching by Demonstration. In *Advances in Neural Information Processing Systems*, 2016.

[36] Mark K. Ho, Fiery Cushman, Michael L. Littman, and Joseph L. Austerweil. Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, 150(11):2246–2272, November 2021.

[37] Frank Hyman. *How to Forage for Mushrooms Without Dying: An Absolute Beginner's Guide to Identifying 29 Wild, Edible Mushrooms*. Storey Publishing, LLC, 2021.

[38] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

[39] Michael Henry Tessler and Noah D Goodman. The language of generalization. *Psychological Review*, 126(3):395, 2019.

[40] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from? In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.

[41] Smitha Milli and Anca D Dragan. Literal or pedagogic human? Analyzing human model misspecification in objective learning. In *Uncertainty in Artificial Intelligence*, 2020.

[42] W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the Fifth International Conference on Knowledge Capture*, page 9–16, New York, NY, USA, 2009. Association for Computing Machinery.

[43] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, 2017.

[44] Mark K. Ho, Fiery Cushman, Michael L. Littman, and Joseph L. Austerweil. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, 148(3):520–549, March 2019.

[45] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2007.

[46] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

[47] Silvia Tulli, Francisco S. Melo, Ana Paiva, and Mohamed Chetouani. Learning from explanations with maximum likelihood inverse reinforcement learning. March 2022.

[48] Christian Daniel, Oliver Kroemer, Malte Viering, Jan Metz, and Jan Peters. Active reward learning with a novel acquisition function. *Auton. Robots*, 39(3):389–405, October 2015.

[49] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29*. 2016.

[50] 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. In *Robotics Research*, pages 49–57. Springer, 2020.

[51] SRK Branavan, David Silver, and Regina Barzilay. Learning to win by reading manuals in a Monte-Carlo framework. *Journal of Artificial Intelligence Research*, 43:661–704, 2012.

[52] John D Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, John DeNero, Pieter Abbeel, and Sergey Levine. Meta-learning language-guided policy learning. In *International Conference on Learning Representations*, 2019.

[53] Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. Skill induction and planning with latent language. *arXiv preprint arXiv:2110.01517*, 2021.

[54] Catherine Wong, Kevin M Ellis, Joshua Tenenbaum, and Jacob Andreas. Leveraging language to learn program abstractions and search heuristics. In *International Conference on Machine Learning*, 2021.

[55] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In *NAACL*, pages 2166–2179, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[56] Valerie Chen, Abhinav Gupta, and Kenneth Marino. Ask your humans: Using human instructions to improve generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

[57] Khanh X Nguyen, Dipendra Misra, Robert Schapire, Miroslav Dudík, and Patrick Shafto. Interactive learning from activity description. In *International Conference on Machine Learning*, 2021.

[58] Yiding Jiang, Shixiang Shane Gu, Kevin P Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2019.

[59] Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, and Pierre-Yves Oudeyer. Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems*, 2020.

[60] Geoffrey Cideron, Mathieu Seurin, Florian Strub, and Olivier Pietquin. Higher: Improving instruction following with hindsight generation for experience replay. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 225–232. IEEE, 2020.

[61] Jesse Mu, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and Edward Grefenstette. Improving intrinsic exploration with language abstractions. *arXiv preprint arXiv:2202.08938*, 2022.

[62] Allison C Tam, Neil C Rabinowitz, Andrew K Lampinen, Nicholas A Roy, Stephanie CY Chan, DJ Strouse, Jane X Wang, Andrea Banino, and Felix Hill. Semantic exploration from language abstractions and pretrained representations. *arXiv preprint arXiv:2204.05080*, 2022.

[63] Valerie J. Shute. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189, 2008.

[64] A.A. Lipnevich and J.K. Smith. Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, 15(4):319–333, 2009.

[65] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

[66] Raymond J Mooney. Learning to Connect language and perception. In *AAAI Conference on Artificial Intelligence*, 2008.

[67] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[68] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818 – 829, 2016.

[69] H. P. Grice. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[70] Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

[71] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Empirical Methods in Natural Language Processing*, 2016.

[72] Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, 2020.

[73] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *NAACL*, June 2018.

[74] Noga Zaslavsky, Jennifer Hu, and Roger P Levy. A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*, 2020.

[75] Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2020.

[76] Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 2017.

[77] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018.

[78] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, volume 15, 2011.

[79] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[80] Theodore Sumers, Robert Hawkins, Mark Ho, and Thomas Griffiths. Extending rational models of communication from beliefs to actions. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2021.

[81] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[82] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[83] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[84] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition, 2009. ISBN 0136042597.

[85] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, volume 97, 2019.

[86] Sophie Bridgers, Julian Jara-Ettinger, and Hyowon Gweon. Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 4(2): 144–152, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the discussion (Section 5).
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work primarily theoretical and intended to have positive societal impact. We do not foresee any material risks at this time..
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See in particular details on our behavioral experiment consent / IRB (Appendix C)).
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See theorems 1 and 2, and the complete proofs in Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will submit code in the supplemental material, and release on GitHub for the camera-ready.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We do not train any models, but we specify the parameter-setting procedures used for the human data in Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All plots contain error bars where relevant and we conduct statistical testing to quantify uncertainty where appropriate.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All analyses were run on consumer hardware (e.g. a Macbook). We note this in Appendix F.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We do not use any existing assets.
   (b) Did you mention the license of the assets? [N/A] We do not use any existing assets.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the (anonymized) data from our experiment in the supplemental materials and will release via GitHub.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The consent form used is linked as part of the experiment.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We hash participant identifiers and include only information collected during the coruse of our experiment. This is discussed in Appendix C.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We link to the full (anonymized) experiment in Appendix C.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not anticipate any participant risks; this is discussed in the consent form on the experiment.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix C.