Linguistic communication as (inverse) reward design

Anonymous ACL submission

Abstract

Natural language is an intuitive and expressive way to communicate reward information to autonomous agents. It encompasses everything from concrete instructions to abstract descriptions of the world. Despite this, natural language is often challenging to learn from: it is difficult for machine learning methods to make appropriate inferences from such a wide range of input. This paper proposes a generalization of reward design as a unifying principle to ground linguistic communication. We first extend reward design to incorporate reasoning about unknown future states in a linear bandit setting. We then define a speaker model which chooses utterances according to this objective. Simulations show that short-horizon speakers (reasoning primarily about a single, known state) tend to use instructions, while long-horizon speakers (reasoning primarily about unknown, future states) tend to describe the reward function. We then define a pragmatic listener which performs inverse reward design by jointly inferring the speaker's latent horizon and rewards. Our findings suggest that this extension of reward design to linguistic communication, including the notion of a latent speaker horizon, is a promising direction for achieving more robust alignment outcomes from natural language supervision.

1 Introduction

Imagine taking up mushroom foraging as a hobby. How would you learn which fungi are delicious and which are deadly? Learning from direct experience (Sutton and Barto, 2018) seems risky. But how might we best learn from others? Prior work in reinforcement learning (RL) has examined a number of social learning strategies, including passive *inverse reinforcement learning* (observe an expert pick mushrooms, then infer their reward function; Ng and Russell, 2000; Abbeel and Ng, 2004) or active *preference learning* (offer an expert pairs of mushrooms, observe which one they eat, and infer their reward function; Markant and Gureckis, 2014; Christiano et al., 2017; Basu et al., 2018).

We posit that few humans would rely on such indirect observations if they had access to a cooperative teacher (Vélez and Gweon, 2021; Gweon, 2021; Wang et al., 2020). For example, an expert guiding a foraging trip might *demonstrate* or verbally *instruct* the learner to pick certain mushrooms rather than others (Shafto et al., 2014; Ho et al., 2016). While such explicit instruction has been a useful tool for guiding RL agents (Goyal et al., 2019; Luketina et al., 2019; Fu et al., 2019; Tellex et al., 2020), natural language affords much richer forms of expression. For example, an expert teaching a seminar might *describe* how to recognize edible or toxic mushrooms from their features.[1] Descriptive language is particularly powerful if learners can expect experts to prioritize *relevant* and *context-sensitive* information (Sperber and Wilson, 1986; Tessler and Goodman, 2019).

To formalize these expectations, we generalize models of *reward design* (Singh et al., 2009) to linguistic communication in a linear bandit setting. Section 2 begins by defining a speaker that chooses utterances to maximize an (imagined) listener's expected rewards over the likely distribution of future states. Section 3 shows that speakers focused on a single state prefer instructions (designating an action to take), while those reasoning about many states prefer descriptions (providing information about the reward function). Finally, we consider how a listener might learn from such a speaker. Section 4 defines a pragmatic listener which performs *inverse* reward design (IRD, Hadfield-Menell et al., 2017), to learn about rewards from both instructions and descriptions.

Using IRD on natural language input offers two distinct benefits over its non-linguistic formulation. First, language is *expressive* yet tractable (for humans): while reward functions are notoriously dif-

[1] Or write a book on the topic, e.g. Hyman (2021).

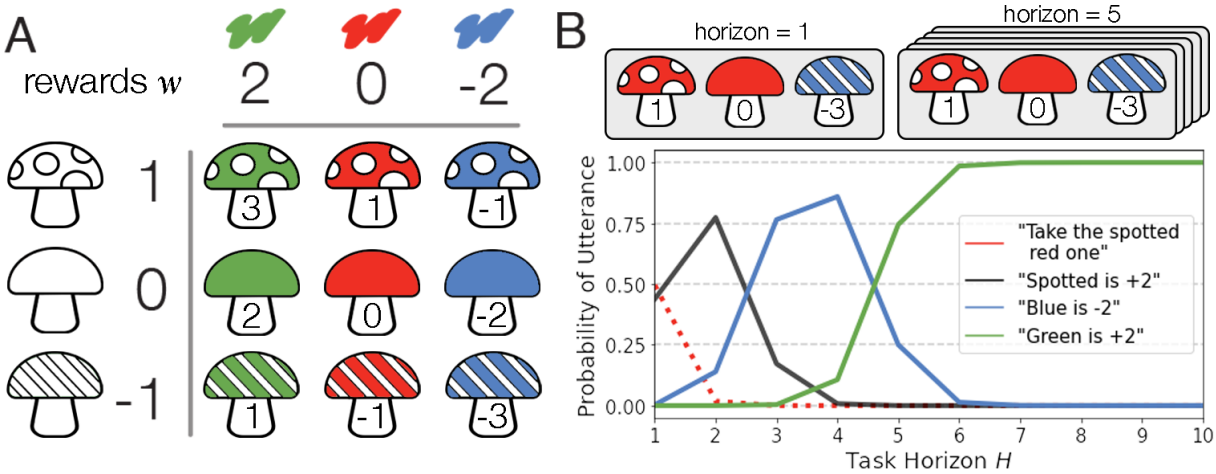

Figure 1: **A**: Rewards associated with features determine whether actions (mushrooms) are high or low reward (tasty or toxic). **B**: Speaker's choice of utterances as a function of horizon H for this start state. At short horizons (maximum supervision), speakers often use instructions or exaggerated descriptions. As the horizon lengthens, there are more unknown states, and speakers prefer truthful descriptions which provide generally useful information. Pragmatic listeners can exploit this pattern to jointly infer a speaker's horizon and rewards.

ficult to specify (Amodei et al., 2016), natural language provides an accessible and expansive space of proxy rewards. Second, language can address *future settings*: speakers can refer to actions or features which are not physically present. Thus, while reward design and IRD assume the reward designer optimizes a known Markov Decision Process (MDP), our formulation relaxes this requirement. We show that pragmatic listeners which jointly infer the speaker's reward function and distribution over states reliably outperform a literal listener.

2 Communication as Reward Design

Linear Bandits We begin by formulating the reward design problem in a *linear bandit* setting (Lattimore and Szepesvari, 2020; Amin et al., 2017). 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; have spots or not). Rewards are defined as a function of these features: $R : \phi(a) \to \mathbb{R}$. We assume they are a linear combination of the features:

$$R(a, w) = w^\top \phi(a) \tag{1}$$

so w is a vector that defines the value of each feature (e.g. green mushrooms are tasty and blue are toxic; 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 as an individual learning problem, we assume that rewards are not directly observable and instead ask how agents should learn *socially*. We formalize the 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 much supervision the speaker exerts. $H = 1$ is maximum supervision (i.e. guided foraging), whereas $H \to \infty$ is minimal supervision (teaching the listener to forage in future settings). We first assume H is known to both listener and speaker, but later relax this assumption.

This social learning framework exposes two interrelated problems. First, what should the speaker agent say to be most helpful? And second, how should the listener update their policy in light of this information?

Speakers as Reward Designers Drawing on the Rational Speech Act framework (RSA, Goodman and Frank, 2016), we define a speaker S_1 that chooses utterances u according to a utility function $U_{S_1}(\cdot)$:

$$S_1(u) \propto \exp\left(\beta_{S_1} \cdot U_{S_1}(u)\right) \tag{2}$$

where β_{S_1} is the speaker's soft-max temperature.

But what utility is appropriate? Rather than defining utility simply as Gricean informativeness (Grice, 1975), i.e. inducing true beliefs, we

suggest that a cooperative speaker should *maximize the listener's rewards*, thus grounding utility in terms of the listener's subsequent actions.[2]

When the state is known, the *present* utility of an utterance is the expected reward from using the resulting policy to choose an action in that state:

$$U_{\text{Present}}(u \mid s, w) = \sum_{a \in s} \pi_L(a \mid u, s) R(a, w) \quad (3)$$

This formulation is equivalent to the *reward design* objective (Singh et al., 2009; Hadfield-Menell et al., 2017), 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* to other mushroom patches. Thus, unlike the reward design objective, speakers may reason about future states. We represent the *future* utility of an utterance with respect to some distribution over states $P(s)$:

$$U_{\text{Future}}(u \mid w) = \sum_{s \in S} U_{\text{Present}}(u \mid s, w) P(s) \quad (4)$$

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. 3 and 4:

$$U_{S_1}(u \mid w, s, H) = U_{\text{Present}} + (H-1) U_{\text{Future}} \quad (5)$$

where $H = 1$ reduces to Eq. 3. We next define how utterances may affect the listener's policy.

3 Choosing Optimal Utterances

We formally define two classes of utterances, *instructions* and *descriptions*, by specifying how they affect the policy of a "literal" listener. We then show how varying the horizon H systematically affects the speaker's choice of utterance.

Instructions Instructions map to specific actions or trajectories (Tellex et al., 2011; Jeon et al., 2020). Given an instruction, a literal listener executes the corresponding action. If the action is not available, the listener chooses an action randomly:

$$\pi_{L_0}(a \mid u_{\text{instruct}}, s) = \begin{cases} \delta_{[\![u]\!](a)} & \text{if } a \in s \\ \frac{1}{|s|} & \text{if } a \notin s \end{cases} \quad (6)$$

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.

[2]For other recent action-oriented RSA formulations, see (Jiang et al., 2021; Stacy et al., 2021; Sumers et al., 2021a).

[3]We assume that groundings are known, i.e. the literal listener understands the meaning of utterances.

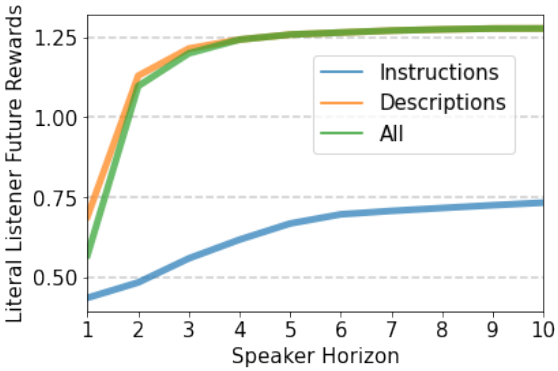

Figure 2: "Future" rewards (Eq. 4, averaged over all 84 start states) for a literal listener as a function of horizon and available utterances. At longer horizons, speakers with access to descriptions produce utterances that generalize well.

Descriptions Rather than mapping to a specific action, descriptions provide information about the world (Ling and Fidler, 2017; Narasimhan et al., 2018; Sumers et al., 2021b). Following Sumers et al. (2021a), we assume that descriptions provide the reward of a single feature, similar to feature queries (Basu et al., 2018).

Formally, we define descriptions as 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⟩. Given a description, a literal listener "rules out" inconsistent hypotheses about reward weights w:

$$L_0(w \mid u_{\text{description}}) \propto \delta_{[\![u]\!](w)} P(w) \quad (7)$$

where $\delta_{[\![u]\!](w)}$ represents the meaning of u, evaluating to one when u is true of w and zero otherwise. Intuitively, descriptions set L_0's beliefs about the reward of a single feature without affecting others. Descriptions need not be accurate; for example, ⟨Spotted, +2⟩ is a false but valid utterance.

The listener then marginalizes over possible reward functions to choose an action:

$$\pi_{L_0}(a \mid u, s) \propto \exp\{\beta_{L_0} \cdot \sum_{w} R(a, w) L_0(w \mid u))\} \quad (8)$$

where β_{L_0} is again a softmax optimality.

Horizons and Utterance Preferences We use simulations to explore the effects of speaker horizons and utterance sets. Fig. 1A shows our bandit setting. "Instruction" utterances correspond to the nine actions. "Description" utterances are the 6 features $\times$ 5 values in $[-2, -1, 0, 1, 2]$, yielding 30 feature-value tuples. We assume the listener begins with a uniform prior over reward weights and

set $\beta_{L_0} = 3, \beta_{S_1} = 10$.[4] We use states consisting of three unique actions, giving 84 possible states.

To quantify how the horizon H affects the generalization of the listener's policy, we repeat the task for all 84 start states using horizons ranging 1-10 and different utterance sets. Fig 1B shows one example, and Fig 2 plots a literal listener's average future rewards. When the horizon is short (small H), speakers focus on the visible state, producing utterances which generalize poorly (low future rewards). As H increases, they provide more generally useful information. Finally, instructions are most useful at short horizons; speakers with access to descriptions use them exclusively when $H > 2$.

4 Learning from Utterances

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

Known Horizon Following the standard RSA formulation, a pragmatic listener L_1 can invert the speaker model. When the speaker's horizon H is known, this is equivalent to inverse reward design (Hadfield-Menell et al., 2017):

$$L_1(w \mid s, u, H) \propto S_1(u \mid w, s, H)P(w) \quad (9)$$

Given an instruction, L_1 can recover information about the reward weights; given a description, L_1 can recover information about features that were not mentioned. The L_1 listener then chooses actions with respect to this posterior by substituting it into Eq. 8. Fig. 3 shows the gain in "future" rewards for a pragmatic listener (L_1 - L_0) when the speaker has access to both instructions and descriptions, and their horizon is known. Pragmatics are particularly helpful when the speaker has a short horizon and is *not* attempting to provide general information.

Misaligned Horizons However, unlike IRD, in linguistic communication the speaker's horizon H is not explicitly known. Prior work has highlighted the risks of assuming a human is behaving pedagogically when they are not (Milli and Dragan, 2020), so we test one form of misalignment: when the speaker $H = 1$ but the listener H ranges from 1-10. Fig. 3 shows that when the pragmatic listener assumes a longer horizon, it overgeneralizes and performs worse than L_0.

[4]Because our action space is small (each state has only 3 actions), descriptions are often equivalent to instructions. A lower β_{L_0} helps compensate for this.

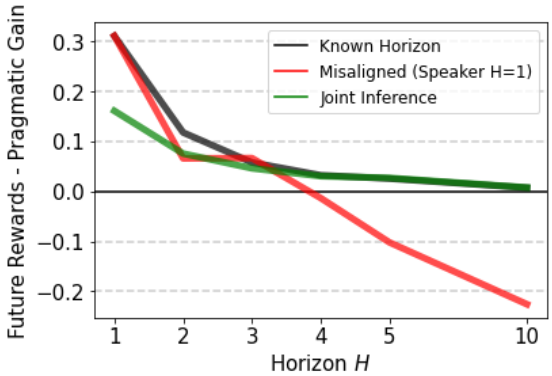

Figure 3: "Future" reward gain from pragmatic inference (Eq. 4, $L_1 - L_0$ averaged over all 84 start states). Reward inference works best when the listener knows the speaker's horizon, but can reduce performance if this assumption is incorrect. Jointly inferring the rewards and horizon (Eq. 10) mitigates this risk.

Inference over Speaker Horizons To mitigate the risk of horizon misalignment, we can instead assume the speaker's horizon is unknown. Given an utterance, the listener jointly infers both their horizon and rewards, then marginalizes out the horizon:

$$L_1(w \mid s, u) \propto \sum_H S_1(u \mid w, s, H)P(H)P(w) \quad (10)$$

We test a pragmatic listener with a uniform prior over $H \in [1, 2, 3, 4, 5, 10]$. This results in more conservative reward inference, but avoids the misalignment risk posed by assuming the speaker's horizon. Fig. 3 shows the results.

5 Discussion

In this work, we formalized communication as reward design, allowing us to to unify instructions and descriptions under a single objective. Simulations show that instructions are optimal when the state is known, but descriptions are optimal when considering a distribution over states. Finally, a pragmatic listener can jointly infer the speaker's horizon and reward function.

One important limitation of this work is our reliance on simulations. Future work should validate the speaker model proposed here with behavioral data. Finally, developmental studies indicate that even young children reason about exploration costs when teaching (Bridgers et al., 2020), suggesting that the reward design objective could be extended further to incorporate reasoning about individual learning.

References

Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

Kareem Amin, Nan Jiang, and Satinder Singh. 2017. Repeated inverse reinforcement learning. *Advances in neural information processing systems*, 30.

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

Chandrayee Basu, Mukesh Singhal, and Anca D Dragan. 2018. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 132–140. IEEE.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307.

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

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

Prasoon Goyal, Scott Niekum, and Raymond J Mooney. 2019. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2385–2391. AAAI Press.

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

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

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

Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. 2016. Showing versus Doing: Teaching by Demonstration. In *Advances in Neural Information Processing Systems*, pages 3027–3035.

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

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

Kaiwen Jiang, Stephanie Stacy, Adelpha Chan, Chuyu Wei, Federico Rossano, Yixin Zhu, and Tao Gao. 2021. Individual vs. joint perception: a pragmatic model of pointing as smithian helping. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Tor Lattimore and Csaba Szepesvari. 2020. *Bandit Algorithms*. Cambridge University Press.

Huan Ling and S. Fidler. 2017. Teaching machines to describe images with natural language feedback. In *NIPS*.

J Luketina, N Nardelli, G Farquhar, J Foerster, J Andreas, E Grefenstette, S Whiteson, and T Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *IJCAI 2019*, volume 57. AAAI Press.

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

Smitha Milli and Anca D Dragan. 2020. Literal or pedagogic human? analyzing human model misspecification in objective learning. In *Uncertainty in artificial intelligence*, pages 925–934. PMLR.

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

Andrew Y Ng and Stuart J Russell. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*.

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

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

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

Stephanie Stacy, Chenfei Li, Minglu Zhao, Yiling Yun, Qingyi Zhao, Max Kleiman-Weiner, and Tao Gao. 2021. Modeling communication to coordinate perspectives in cooperation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

T Sumers, RD Hawkins, M Ho, and TL Griffiths. 2021a. Extending rational models of communication from beliefs to actions. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. 2021b. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010.

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

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

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

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

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

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. 2020. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33:17582–17593.