# A Proofs

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

*Proof.* (sketch) The future utility for the listener in Eq. 6 is exactly the expected utility generated by $\pi_L$, averaged across all possible decision contexts. Thus, the limit of Eq. 7 as $H \to \infty$ is the expected utility generated by $\pi_L$. Then, consider two utterances $u_1, u_2$, with $U_{Future}(u_1|w) > U_{Future}(u_2|w)$. The odds ratio $\frac{P_{S_1}(u_1)}{P_{S_1}(u_2)}$ between can be shown to increase with $H$ by taking the derivative. Thus, utterances become increasingly preferred by the speaker, hence the expected utility of the updated listener policy $\pi_L(a|s, u)$ under those utterances increases. □

**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 $\widetilde{\mathcal{R}}$.*

*Proof.* The likelihood function for noisily-optimal behavior gives $P(a|s, w) \propto \exp\left(\beta R(a, w)\right)$. Thus, the posterior distribution over reward, given that action $a$ was taken in state $s$ can be written

$$P(w|s, a) = \frac{\exp\left(\beta R(a, w)\right)}{\sum_{a' \in s} \exp\left(\beta R(a', w)\right)}. \tag{13}$$

To show that this is equivalent to the posterior from observing an instruction from a short-sighted speaker that can select utterances from $\mathcal{U}_{instruct}$ according to Eq. (2) with utility function Eq. 5, we observe that the utterance likelihood function and the action likelihood function are in one-to-one correspondence. The utterance likelihood is

$$P(u|s, w) \propto \exp\left(\beta_{S_1} \sum_{a \in s} \pi_{L_0}(a|u, s) R(a, w)\right). \tag{14}$$

We can see that this is equivalent to Eq. 13 by substituting $\pi_L$ from Eq. 8 and then using the one-to-one mapping from $\mathcal{U}_{instruct}$ to $\{a \in s\}$ to rename variables.

Next, we show that this is equivalent to (locally-optimal) reward design for $s$. In state $s$, the optimal policy, given the proxy, is to take the only action that gets reward according to the proxy. Thus, this is equivalent to Eq. 8. As a result, a reward designer optimizing over this set of proxies will behave as if they are selecting utterances from $\mathcal{U}_{instruct}$. A similar line of reasoning shows the result. □

# B Speaker simulations and pragmatic inference

## B.1 Instructions vs Descriptions

In the main text, we used a fixed number of available actions (a context $S$ with $|S| = 3$ objects). Here, we further explore the effect of horizon on the choice of instructions vs. descriptions under different numbers of available actions. As in the main text, we assume that the speaker uses a near-optimal softmax temperature for clarity by setting $\beta_{S_1} = 10$.

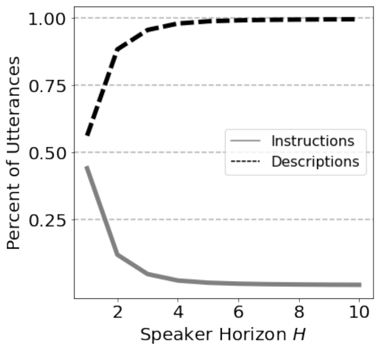


**Figure S1:** As noted in Section 3.6, speakers exhibit a strong preference for descriptions as their horizon lengthens.

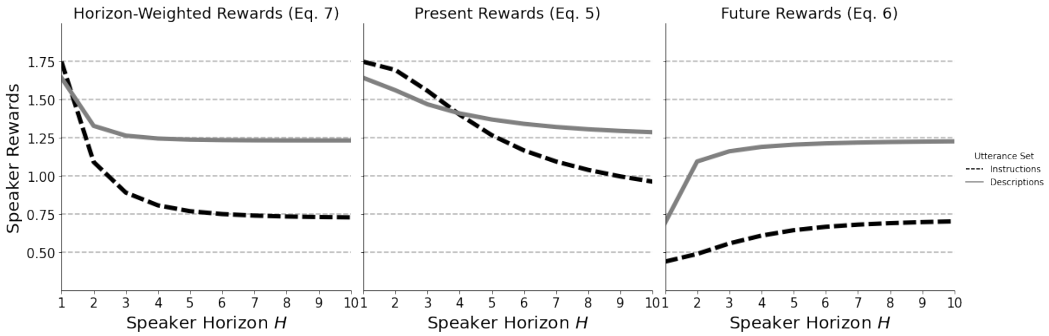


**Figure S2:** Breaking out speaker rewards (Fig. 2A) by reward type and utterance. "Horizon-Weighted Rewards" (left) is the same as Fig. 2A. Instructions afford high "Present Rewards" (center) but generalize poorly (low "Future Rewards", right). As a result, rational speakers with access to instructions only remain biased towards the present context even as their horizon lengthens. This can be seen by comparing "Present" and "Future" rewards at long horizons (e.g. $H = 10$). Description-only speakers exhibit little bias towards the present context ("Present" and "Future" rewards are nearly equal), while instruction-only speakers remain biased towards the present context ("Present" > "Future" rewards).

We find that instructions become more useful as the number of available actions increases. They can always uniquely select the best action in a given state (even when all nine possible objects are present), whereas it is not always possible to use a description to identify the best action.

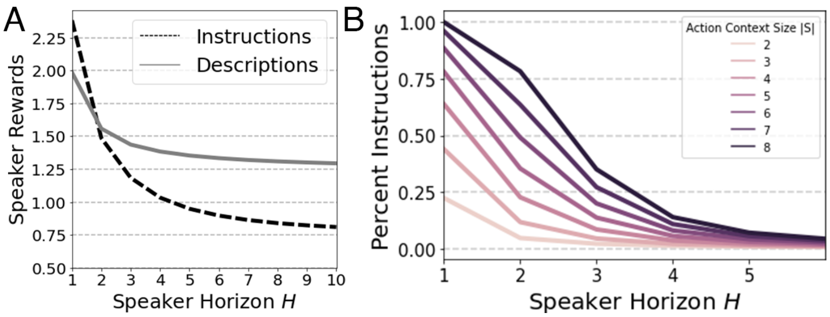


**Figure S3:** **A**: Same as Fig. 2A, but with a state-size $|S|$ of 5 instead of 3. At short horizons, the relative utility of instructions increases with state size (e.g. as the action space grows, instructions are more useful). **B**: Speaker's probability of using an instruction as a function of number of available actions $|S|$ and horizon $H$ (note that Fig. S1 shows the curve for $|S| = 3$). As the number of actions increases, speakers prefer instructions.

Next, building on Fig. 2B in the main text, we include full posteriors over reward functions and additional utterances.

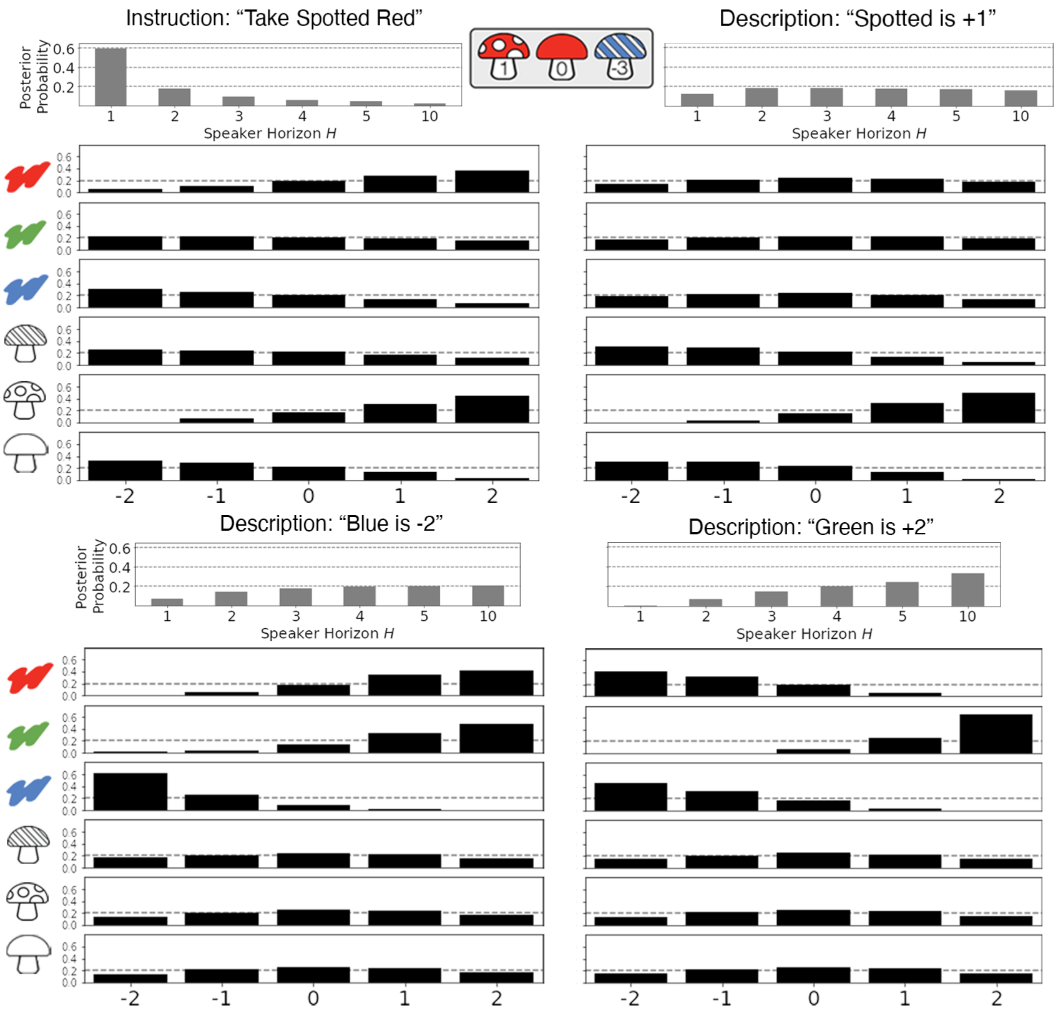


**Figure S4:** Same as Fig. 2B, but with full utterance posteriors over possible reward functions and additional utterances. For utterance posteriors, the gray dashed line indicates the prior (e.g. uniform over all possible values). Note that descriptions suggest that unmentioned features are lower-magnitude (e.g. for the bottom-right “Green is +2” utterance, the listener infers that all textures—Striped, Spotted, and Solid—are unlikely to be -2 or +2). Finally, note that with all descriptive utterances, the listener assigns non-negligible probability mass to values *other* than the specified one (e.g. in the top-right, the listener infers a substantial probability that the Spotted feature is actually +2). This suggests that integrating a “truthfulness bias” could improve our models (see Section 4.2 for a discussion).

# C Behavioral Experiment

The full (anonymized) experiment can be viewed at `pragmatic-bandits.herokuapp.com`. Note that the app is running on free dynos so you may need to wait 5-10 seconds for it to load.

## C.1 Experiment Details

**Participant compensation.** Participants earned an average hourly wage of \$12.10 and the total amount spent on participant compensation was \$444. The mean time spent in the experiment was 18.5 minutes.

**IRB approvals.** This study was approved by **Anonymized** University’s IRB. All participants gave informed consent; the consent form can be seen at the experiment URL above. As described in the Checklist, no significant participant risks were anticipated.

**Anonymized data.** Anonymized data, including participant responses and free-form exit survey responses, is available in the supplementary zip file and will be released along with the code for this paper. Note that the worker IDs provided have been hashed to prevent re-identification of participants on the platform.

**Trials.** We split our 84 states into 3 sets of 28; each participant saw one of these sets. Each participant additionally saw 8 "attention check" trials (constant across all participants). These "attention checks" forced the participant to use a description with a pre-selected feature (4 "Spotted" and 4 "Striped"). The participant then chose a value from $[-1, +1]$. The trials were selected to ensure that the true value would lead the learner to choose a good mushroom. Participants who failed to select the true value on at least 7/8 trials (e.g. >75% of the time) were still paid the full bonus, but their data was excluded from the analysis. We further exclude the attention check trials from the analysis.

**Feature Randomization.** To avoid saliency biases (e.g. color may be more salient than texture), mushroom feature values were randomized across participants. Fig. S5 shows one example of an alternative featurization scheme. Note that all responses were converted back the the "canonical" feature map shown in Fig. 1 for analysis.

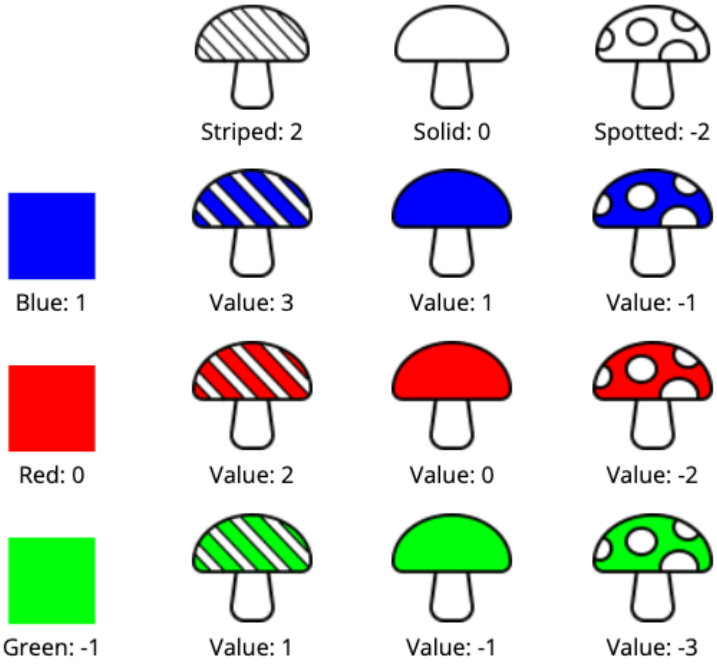
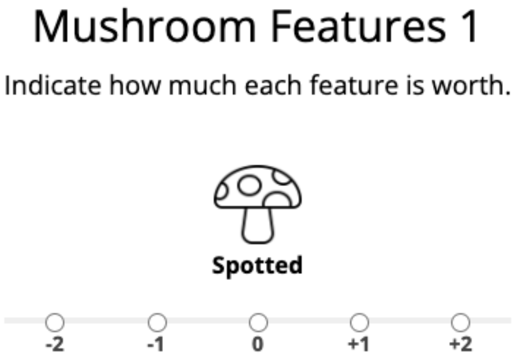


**Figure S5:** Throughout the experiment, participants could trigger a pop-up giving them the value of all features (and all actions). Note that features were randomized to avoid saliency biases, and so this set of feature values does not match Fig. 1.

**Instructions** We include several screenshots of key instruction pages, but recommend viewing the full experiment at `pragmatic-bandits.herokuapp.com` for details.

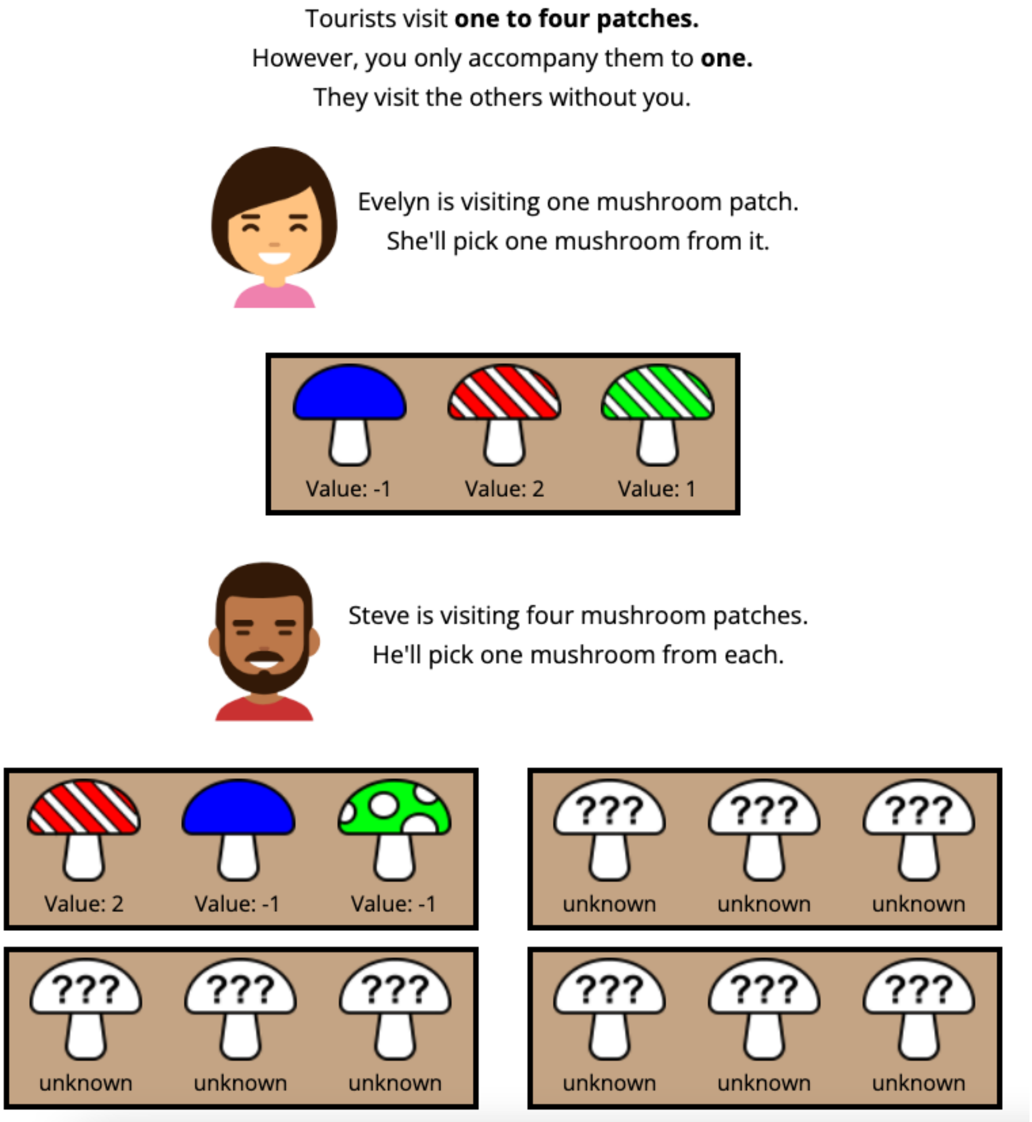


Mushroom Features 1

Indicate how much each feature is worth.

○ -2 ○ -1 ○ 0 ○ +1 ○ +2

**Figure S6:** One example quiz question. To ensure comprehension of the linear bandit setup, participants were tested on their knowledge of all features. If they failed the quiz, the experiment terminated early and they earned $2; if they passed, they completed the full experiment and earned $4.



Tourists visit **one to four patches.**

However, you only accompany them to **one.**

They visit the others without you.

Evelyn is visiting one mushroom patch.

She'll pick one mushroom from it.

Value: -1 Value: 2 Value: 1

Steve is visiting four mushroom patches.

He'll pick one mushroom from each.

Value: 2 Value: -1 Value: -1

??? ??? ???

unknown unknown unknown

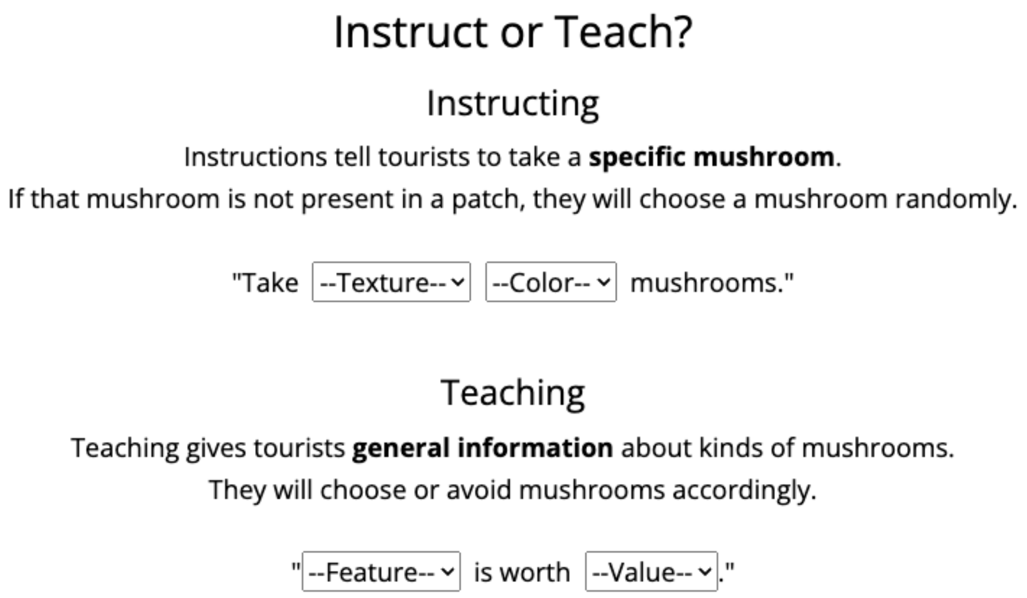
??? ??? ???

unknown unknown unknown

??? ??? ???

unknown unknown unknown

**Figure S7:** Experiment instructions: introducing the notion of the *horizon*. Participants were told they could only accompany the tourist to one patch, but that depending on their itinerary, tourists could go on to visit other (unknown) patches afterwards.



Instruct or Teach?

Instructing

Instructions tell tourists to take a **specific mushroom**.
If that mushroom is not present in a patch, they will choose a mushroom randomly.

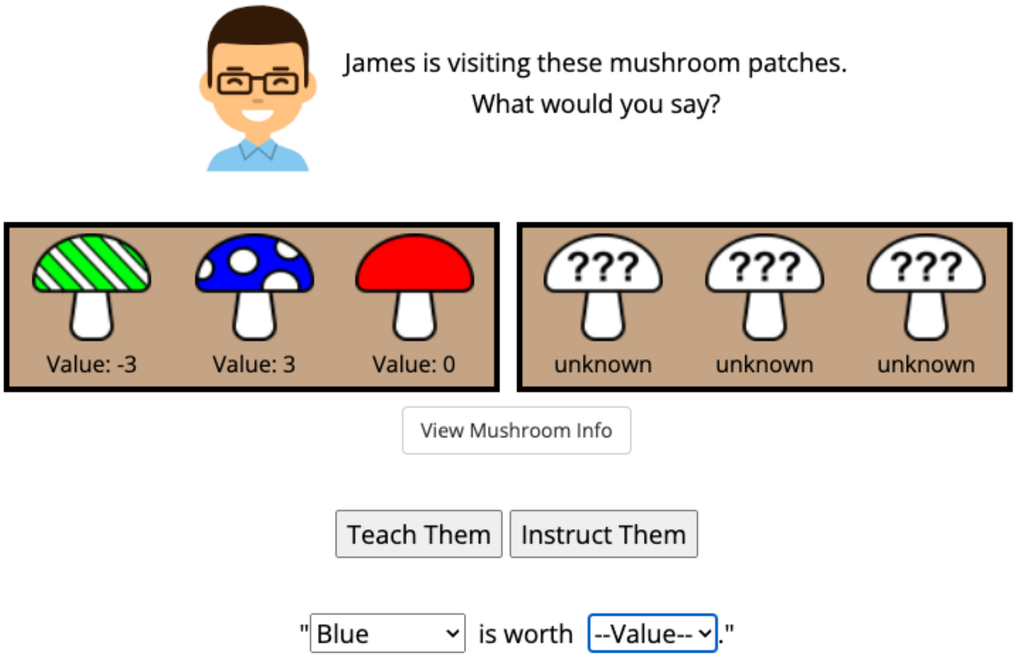
"Take --Texture-- --Color-- mushrooms."

Teaching

Teaching gives tourists **general information** about kinds of mushrooms.
They will choose or avoid mushrooms accordingly.

"--Feature-- is worth --Value--."

**Figure S8:** Experiment instructions: introducing the notion of instructions and descriptions. Participants were told to consider the tourist's itinerary (Fig. S7) and help the tourist pick good mushrooms throughout their visit.



James is visiting these mushroom patches.
What would you say?

"Blue is worth --Value--."

**Figure S9:** One example trial from the experiment. Clicking the "Teach Them" button revealed drop-down menus to select a "Description" utterance, while clicking "Instruct Them" yielded menus to select an "Instruction."

## C.2 Participant Utterance Choices

The supplementary materials contain all participant responses (see the Experiment Analysis Jupyter notebook for analysis code). Here, we summarize some of the key patterns in the data.

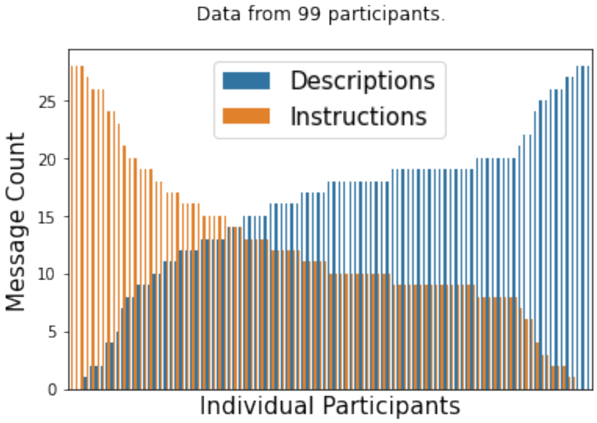


**Figure S10:** Breakdown of utterance types (instruction vs descriptions) for all participants. Most participants used a mix of both: 3 used only instructions, 3 used only descriptions, and 93 used at least one message of each type.

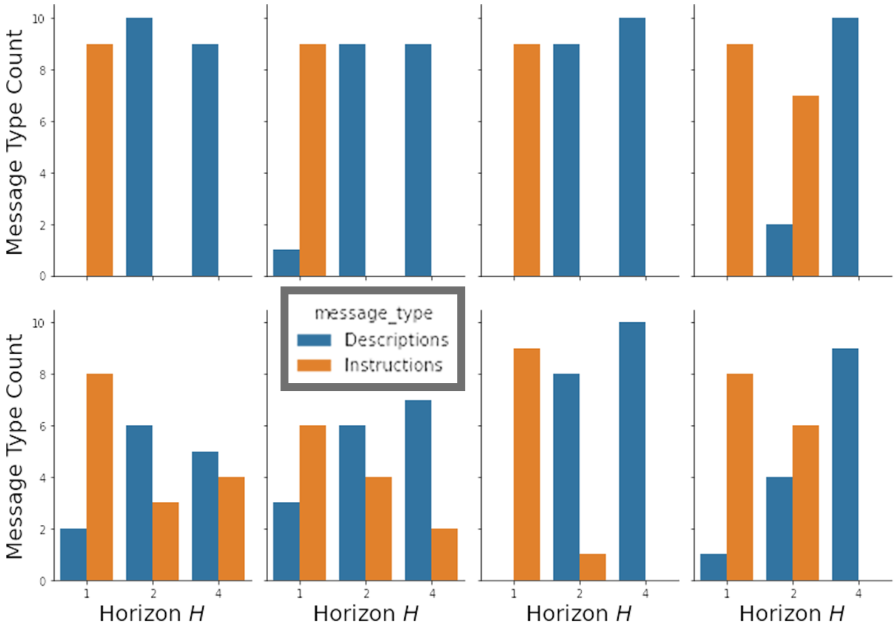


**Figure S11:** Instruction / description breakdown by horizon for 8 random participants. While individual preferences varied substantially, virtually all participants displayed an increasing preference for descriptions as the horizon increased.

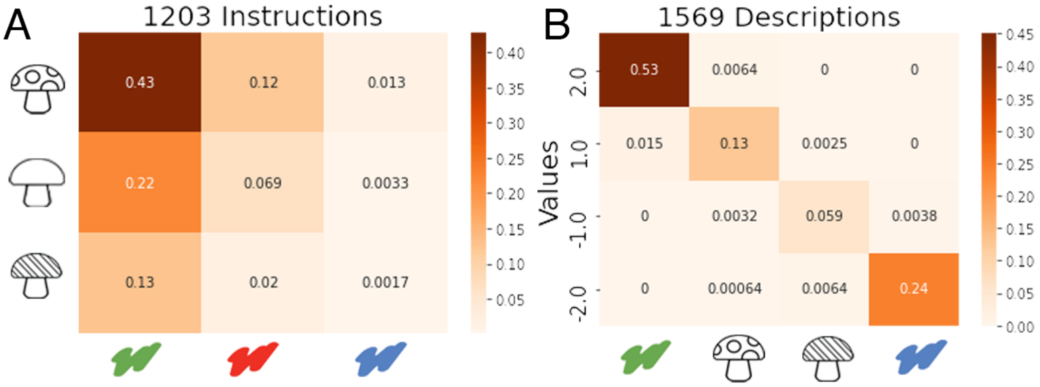


**Figure S12:** Within-type distribution of utterances chosen by participants. **A**: When giving instructions participants (unsurprisingly) almost always chose positive-reward actions, e.g. "Take the spotted green mushroom," in the top-left quadrant. **B**: When giving descriptions, participants almost always chose *true* utterances (e.g. "Spotted is +1"), even though our reward-maximizing model predicts exaggeration (e.g. "Spotted is +2"). See Section 4.2 for discussion.

### C.3 Choosing $\beta_{S_1}$

To choose a $\beta_{S_1}$ for our behavioral experiment, we used a grid search over integers $\beta_{S_1} \in [1, 10]$ and evaluated our primary models (Pragmatic - Known $H$ and Pragmatic - Latent $H$). We chose $\beta_{S_1} = 3$, which optimized future reward from the human data for both speakers. Note that while "Pragmatic - Known $H$ was numerically optimal at $\beta_{S_1} = 2$ (expected reward $= 0.9308, SD = 0.39$), there was not find a significant difference between this and $\beta_{S_1} = 3$ (expected reward $= 0.9295, SD = 0.40$); paired-samples $t(2771) = -1.70, p = 0.09$.

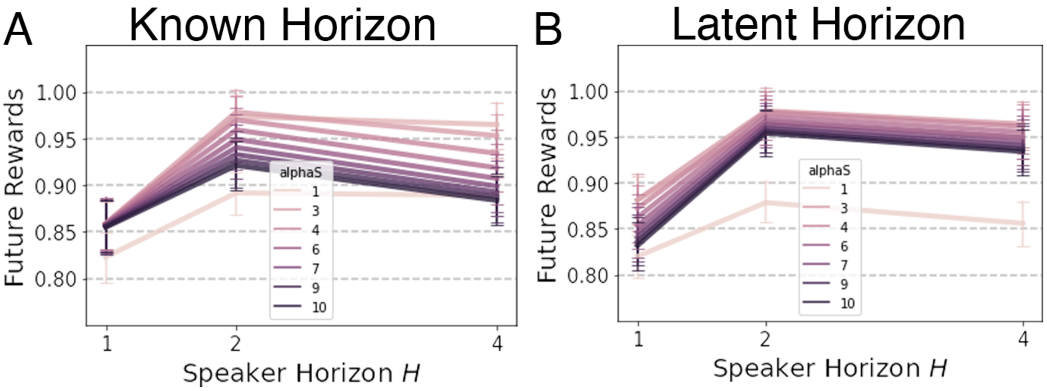


**Figure S13:** Performance of different pragmatic listener models as a function of horizon $H$ and speaker-optimality $\beta_{S_1}$. Qualitatively, the Latent-$H$ model (right) was less sensitive than the Known-$H$ model (left).

| $\beta_{S_1}$ | Known $H$ | Latent $H$ |
|---|---|---|
| 1 | 0.87 | 0.85 |
| 2 | 0.93 | 0.93 |
| **3** | **0.93** | **0.94** |
| 4 | 0.92 | 0.94 |
| 5 | 0.91 | 0.93 |
| 6 | 0.90 | 0.92 |
| 7 | 0.90 | 0.92 |
| 8 | 0.89 | 0.91 |
| 9 | 0.89 | 0.91 |
| 10 | 0.89 | 0.91 |

**Table S1:** Mean "Future Rewards" for our two primary models of interest as a function of the $\beta_{S_1}$ parameter. Note that while "Pragmatic - Known $H$ was slightly higher at $\beta_{S_1} = 2$ there was not find a significant difference between this value and $\beta_{S_1} = 3$, hence we use the latter to be consistent across both models.

### C.4 Simulating model behavior

To compare the empirical pattern of utterances observed from humans in our experiment against the predictions of our theoretical speaker model, we use simulations to generate a distribution over utterances and directly compare the results (Fig. 3, "Model Predictions"). We set $\beta_{S_1} = 3$ as described above. The utterance set is composed of the 9 instructions and 16 descriptions defined in Section 4.1, for a total of 25 possible utterances.

First, for each $H \in [1, 2, 4]$ and each of the 84 states $s \in S$, we run the speaker model to produce a distribution over the 25 possible utterances (Eq. 7). We then calculate the literal future rewards resulting from each utterance (Eq. 8 for instructions and Eqs.9, 10 for descriptions). We then calculate the *expected future reward* obtained by the literal listener by weighting the rewards for each utterance by the probability of the speaker producing that utterance, and averaging over all 84 start states.

Similarly, to evaluate the pragmatic listener, we perform pragmatic inference over each utterance (Eq. 11) to recover the speaker's reward function, evaluate the future rewards (Eq. 6) from the resulting beliefs, and again weight by the speaker's distribution over utterances.

## D Statistical testing

See the R notebook for statistical testing code.

### D.1 Paired T-Tests (§ 4.2)

| Comparison | Mean Difference | 95% CI | | t | df | p-val |
|---|---|---|---|---|---|---|
| vs. Pragmatic (Known $H$) | 0.011 | 0.006 | 0.017 | 4.1798 | 2771 | <.001 |
| vs. Pragmatic ($H = 4$) | 0.026 | 0.022 | 0.031 | 10.81 | 2771 | <.001 |
| vs. Pragmatic ($H = 1$ ) | 0.11 | 0.10 | 0.13 | 20.545 | 2771 | <.001 |
| vs. Literal | 0.15 | 0.13 | 0.16 | 23.364 | 2771 | <.001 |

**Table S2:** Pairwise $t$-tests comparing the "Future Rewards" obtained by the Latent-$H$ listener to other models for the 2772 utterances from our behavioral experiment. These results indicate that the Latent $H$ model outperforms all other models.

| Comparison | Mean Difference | 95% CI | | t | df | p-val |
|---|---|---|---|---|---|---|
| vs. Pragmatic (Known $H$) | -0.13 | -0.15 | -0.12 | -19.854 | 2771 | <.001 |
| vs. Pragmatic ($H = 4$) | -0.12 | -0.13 | -0.11 | -23.324 | 2771 | <.001 |
| vs. Pragmatic ($H = 1$) | -0.03 | -0.05 | -0.01 | -3.202 | 2771 | <.01 |

**Table S3:** Pairwise $t$-tests comparing the "Future Rewards" obtained by the Literal listener to the remaining other models for the 2772 utterances from our behavioral experiment. These results indicate that all pragmatic models outperform the Literal model.

### D.2 Mixed-effects regression model (§ 4.3)

The following analysis tests for a significant difference in regret when using the model's social-learning posterior as a prior for individual learning (see Section F for details). Note that lower regret is better, so negative coefficients indicate better performance.

We dummy-coded our different models as a categorical variable with the Latent $H$ listener as the reference level. We included random intercepts for each unique utterance from our experiment (e.g. for each of the 2772 utterances chosen by participants) to account for some utterances being systematically easier or harder than others. The resulting coefficients indicate that the Latent $H$ listener outperformed all models except for the Known $H$ model, which achieved slightly lower regret.

| | Effect | Term | Estimate | Std Error | Statistic | df | $p$ value |
|---|---|---|---|---|---|---|---|
| 1 | fixed | (Intercept) | 9.55 | 0.05 | 195.66 | 14282.57 | $< 0.001$ |
| 2 | Fixed | Individual | 2.60 | 0.06 | 45.41 | 80383.00 | $< 0.001$ |
| 3 | Fixed | Literal | 0.68 | 0.06 | 11.93 | 80383.00 | $< 0.001$ |
| 4 | Fixed | Prag (Known $H$) | -0.12 | 0.06 | -2.14 | 80383.00 | 0.03 |
| 5 | Fixed | Prag ($H = 1$) | 0.22 | 0.06 | 3.77 | 80383.00 | $< 0.001$ |
| 6 | Fixed | Prag ($H = 4$) | 0.13 | 0.06 | 2.24 | 80383.00 | 0.03 |
| 7 | Random Effect | sd__(Intercept) | 1.44 | | | | |
| 8 | Residual | sd__Observation | 4.76 | | | | |

## E When does pragmatic reasoning help?

In this section, we examine the utterances produced in the human experiment (Section 4 and Appendix C) to explore when, exactly, pragmatic reasoning is most useful. We analyze the performance of the Latent $H$ pragmatic model in comparison to the Literal listener. Concretely, we take the 2772 utterances produced in our behavioral experiment and evaluate the "future rewards" (Eq. 6, the expected rewards over all possible states, ) resulting from a literal interpretation of the utterance against those resulting from a pragmatic interpretation.

| Utterance Type | Count | Mean Pragmatic Gain |
|---|---|---|
| Instruction | 1203 | .42 ± .23 |
| Description | 1569 | -.07 ± .22 |

**Table S4:** Average pragmatic gain for different utterance types (+/- standard deviations). Pragmatics on instructions helps substantially by converting from partial policies to rewards, but pragmatics on descriptions marginally *reduces* the average reward obtained.

We find that under these conditions, pragmatic inference primarily helps with *instructions*, rather than descriptions (Table S4): converting a partial policy into inference over the reward function allows much stronger generalization. Across the 1203 instruction utterances in our experiment, the pragmatic listener achieved a large and statistically-significant gain ($M = .423, SD = .232$), $t(1202) = 63.34, p < .001$. In contrast, on the 1569 descriptive utterances, the pragmatic listener suffered a small but statistically-significant loss ($M = -.067, SD = .215$), $t(1568) = -12.29, p < .001$.

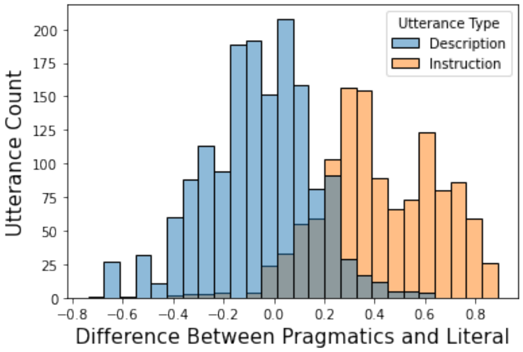


**Figure S14:** Distribution of pragmatic gain (Pragmatic listener with Latent $H$ vs. Literal listener) for the 2772 utterances in our behavioral experiment. Pragmatic inference substantially improves rewards for instructions, but marginally reduces rewards for descriptions on expectation.

Analysis of utterance posteriors (Fig. S4) shows one notable disconnect with empirical human behavior regarding descriptive utterances. The pragmatic listener *does not* preserve the literal truth conditions of descriptive utterances: for the three descriptions shown in Fig. S4, the listener places substantial probability mass on values *other* than the specified one (e.g. believing that "Spotted is +1" suggests "Spotted is +2" is plausible). Yet in our experiment, participants almost always choose true utterances (see Fig. S12B). This suggests future work integrating truthfulness and reward objectives, effectively combining our current objective with classic Gricean notions.

# F Social vs. individual reinforcement learning (§ 4.3)

To study the potential benefits of integrating social and reinforcement learning, we integrated the reward information learned from our behavioral experiment into a classic Thompson sampling individual learner in § 4.3. Here, we provide details on this integration. Code for these simulations can be found in the Supplemental Materials.

## F.1 Individual learning: Thompson sampling in linear bandits

We first define a simple individual learner in our linear bandit setting using Thompson sampling [81–83]. The agent begins with a prior distribution over possible reward functions. At each timestep, they (1) observe a new state $s_t$ consisting of three possible actions; (2) sample a reward weight vector $w_t$ from their belief distribution over reward weights, and (3) act optimally according to that reward vector. They observe the reward of that action, and use this observation to update their beliefs for the next timestep.

We implement this algorithm using a Gaussian prior and likelihood function, assuming observation noise from a unit Gaussian. Thus, after taking action $a$, the agent receives rewards according to:

$$R(a) \sim \mathcal{N}(\phi(a)^\top w, 1) \tag{15}$$

We use a wide multivariate Gaussian prior: $w^0 \sim \mathcal{N}(0, \Sigma_0)$ where $\Sigma_0 = 5I$. After each action, we perform conjugate Bayesian updates to obtain a posterior (i.e. use Bayesian linear regression), which we use for for the next timestep.

We perform rejection sampling to ensure the sampled belief is compatible with the (discrete, bounded) reward function. We first sample a (continuous) weight vector from our multivariate Gaussian beliefs, then round the weights to integer values and reject the sample if any of the resulting values fall outside the range of possible reward values, $[-2, 2]$.

We note that these simulation parameters are arbitrary. Our aim is to demonstrate the general utility of social information to reduce regret *even when individual learning is entirely possible*. We thus defined a relatively straightforward, low-noise individual learning setting. However, we could easily make individual learning arbitrarily more difficult (e.g. by increasing the observation noise), which would in turn increase the relative value of social information.

### F.2 Integrating pragmatic inference: Importance sampling

In order to integrate social information about the reward function, we incorporate an additional importance sampling step. Given a particular pragmatic model and a utterance-context-horizon tuple, we first use the pragmatic model to generate a *social* posterior over reward functions (Eq. 11 or 12). This defines a probability for every possible reward function (e.g. every reward weight vector $w$). We then initialize our individual learner as described above.

When the individual learner performs Thompson sampling, it now performs an additional importance sampling step. Rather than sample a *single* reward vector from its Gaussian prior, it samples a minimum of 100 possible reward vectors. As described above, it first discretizes these vectors then re-weights them according to the probability of each vector from its pragmatic social inference. Finally, it samples a single reward vector from this re-weighted set and uses this to choose an action.

### F.3 Integrating literal information

We use a similar procedure to test individual learning with our literal listener. For descriptive utterances, we use the listener's posterior over reward functions (Eq. 9). However, because there are a handful of false utterances in the experimental data (e.g. "Spotted is -1"), using a hard constraint breaks the importance-sampling procedure described above. We therefore instead use soft-conditioning by setting a very low likelihood on inconsistent worlds ($\epsilon = 1^{-10}$) instead of ruling them out entirely. We use this posterior for importance sampling as described above.

For instructions, we modify the action selection step. We set the listener's policy to take the instructed action if available. If the action is not available, then they follow the Thompson sampling procedure described above. This is the simplest and most "obedient" interpretation of instructions [14]. We find that it yields rapid learning early on, as the instruction guides exploration. However, 57% of instructions designate sub-optimal actions (see Fig. S12A). A literal listener instructed to take one of these (e.g. "Take solid green mushrooms") is forced to continue taking them *even after* inferring spotted green mushrooms are likely worth more. This constraint on their policy eventually leads their regret to asymptote below the more flexible pragmatic learner (Fig. 4). As discussed in the main text and noted in prior work [14], more flexible approaches to instruction-following could avoid this pitfall.

### F.4 Simulation details

All simulations were run on consumer hardware (a MacBook Pro). For each of the 2772 utterances in our behavioral experiment, we ran 5 independent Thompson sampling simulations, each spanning 25 timesteps. We repeated this process for each of the pragmatic listener models (Known $H$, Latent $H$, $H = 1$, and $H = 4$) in our experiment, giving us 13860 Thompson sampling simulations each model. We then ran the same number of independent simulations for the "Individual" learner (which used only the Gaussian prior described in Appendix F.1).