Inferring the Goals of Communicating Agents from Actions and Instructions

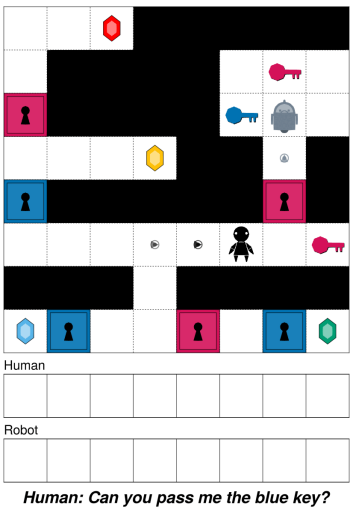
Lance Ying [*1,2] **Tan Zhi-Xuan** [*2] **Vikash K. Mansinghka** [2] **Joshua B. Tenenbaum** [2]

Abstract

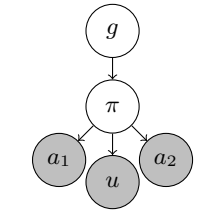
When humans cooperate, they frequently coordinate their activity through both verbal communication and non-verbal actions, using this information to infer a shared goal and plan. How can we model this inferential ability? In this paper, we introduce a model of a cooperative team where one agent, the principal, may communicate natural language instructions about their shared plan to another agent, the assistant, using GPT-3 as a likelihood function for instruction utterances. We then show how a third person observer can infer the team's goal via multi-modal Bayesian inverse planning from actions and instructions, computing the posterior distribution over goals under the assumption that agents will act and communicate rationally to achieve them. We evaluate this approach by comparing it with human goal inferences in a multi-agent gridworld, finding that our model's inferences closely correlate with human judgments ($R = 0.96$). When compared to inference from actions alone, we also find that instructions lead to more rapid and less uncertain goal inference, highlighting the importance of verbal communication for cooperative agents.

1. Introduction

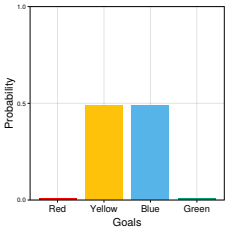
Human cooperation is a flexible, interactive process that involves the mutual observation and interchange of a great variety of signals and cues, providing information about the goals, intentions, beliefs, and other mental states of the people involved. Some of these signals are implicit, such as goal-directed actions, whereas others are explicit, such as verbal communication. In order to navigate cooperative life, social agents like ourselves must integrate this multiplicity of information into coherent theories of others' minds, drawing inferences about shared or individual goals and plans that can serve as guides to cooperative action.



(a) Example cooperative scenario.



(b) Graphical model.



(c) Goal posterior.

Figure 1: An overview of our framework: (a) A human-robot team cooperates to achieve a shared goal g (one of the 4 colored gems), with the human uttering an instruction $u =$ *"Can you pass me the blue key?"* to the robot assistant. (b) We model the team as forming a *joint* plan π to achieve their goal g, which dictates their actions a_1 (human), a_2 (robot). The human also communicates part of this plan as an instruction u. (c) Bayesian inverse planning produces a posterior distribution over goals $p(g|u, a_1, a_2)$.

What is it that explains this inferential ability in humans, and how can it inform the design of cooperative AI systems? We take steps toward an answer by building upon a long tradition in cognitive science that models human linguistic and action understanding as processes of Bayesian interpretation: On one hand, Bayesian theory-of-mind (BToM) posits that humans understand other's actions by inferring the goals and beliefs that explain those actions as rational (Baker et al., 2009; 2017). On the other hand, rational speech act (RSA) theory suggests that humans interpret other's utterances not just in terms of bare semantics, but also the pragmatic intentions they imply (Goodman & Stuhlmüller, 2013; Goodman

[*]Equal contribution [1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Lance Ying <lanceying@seas.harvard.edu>, Tan Zhi-Xuan <xuan@mit.edu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

& Frank, 2016). Since each of these frameworks are formulated in terms of Bayesian inference over the mental states that might explain observed actions or instructions respectively, it is natural to combine them, achieving *joint* inference from actions *and* uttered instructions.

In this paper, we develop a Bayesian model of communicating team agents that incorporates aspects of both these frameworks. The team consists of two agents, a principal (played by a human) who may communicate instructions to an assistant (played by a robot), both of whom act in order to achieve a shared goal (illustrated in Figure 1(a)). Unlike related work that explores how the assistant should infer the principal's goal (Hadfield-Menell et al., 2016; Jeon et al., 2020; Squire et al., 2015), our task is to infer the *team's* goal given their actions and communicated instructions, producing a distribution over goals (Figure 1(c)).

To do so, we follow recent work in cooperative agency by modeling the team as a **group agent**, bypassing the challenge of recursive mental reasoning (Shum et al., 2019; Tang et al., 2020; 2022; Wu et al., 2021). We implement this model as a **probabilistic program** that comprises a goal prior, joint planner, and utterance model (schematically depicted in Figure 1(b)), extending a line of research that uses the flexibility of probabilistic programming to modularly specify agent models in terms of deterministic, probabilistic, and black-box components (Evans et al., 2017; Cusumano-Towner et al., 2017; Seaman et al., 2018; Zhi-Xuan et al., 2020; Berke et al., 2020). This in turn allows us to easily integrate **neural language models as flexible utterance likelihoods** given hypothesized goals and plans, building upon the insight made by Lew et al. (2020) and subsequent papers (Dohan et al., 2022; Li et al., 2023) that (large) language models (LLMs) can be used as modular components in larger probabilistic models.

To evaluate this model, we conduct a series of computational and human experiments that tested the model's ability to accurately infer the goal of a team in a multi-agent grid-world environment, and also how well it explains the goal inferences of third-person human observers when they are provided with the same actions and instructions. For comparison, we also perform experiments in a baseline setting where instructions are omitted, allowing us to isolate the role that language information plays in goal inference. We find that human goal inferences are highly and robustly correlated with the inferences produced by our model, and that language instructions greatly accelerate the convergence of inferences to the true goal, with remaining ambiguity resolved by action information. Collectively, these findings suggest that our model is a viable explanation for how humans infer goals from actions and instructions, as well as a promising route towards building communicative AI assistants that act on the basis of well-calibrated goal inferences.

2. Modeling Communicating Agents

In accordance with the principle of rational action (Gergely & Csibra, 2003; Baker et al., 2009) and rational speech act theory (Goodman & Frank, 2016), we model communicative cooperators as rational agents who act and communicate efficiently to achieve shared goals. However, a complete model of rational Bayesian communication and action requires a great deal of sophistication: Since each agent in the cooperating team may not initially know the team's goal (as is the case for our robot assistant), a third person observer would have to model not only the team's shared goal, but also each agent's *beliefs* about their shared goal, including how those beliefs are formed through goal inference. In addition, agents who know the goal (such as our human principal) would have to be modeled as *pedagogically* selecting utterances in order to best reduce listeners' uncertainty about the shared goal and plan (Shafto et al., 2014). As a further level of sophistication, an observer might model an assistive agent as a *pragmatic listener* who reasons about what a pedagogical speaker might utter (Fisac et al., 2017).

2.1. Communicating Teams as Group Agents

We sidestep these multiple levels of recursive reasoning by opting to model a cooperating team as *a single group agent*: Instead of separately representing the mental states of both the human principal and the robot assistant, we model them as a singular mind with a shared goal g, and a joint plan π. Given this joint plan π, the principal agent utters an instruction u to communicate the plan, and each agent i takes actions $a_{i,t}$ at timestep t according to the plan (a simplified graphical model is shown in Figure 1b):

$$\textit{Goal Prior:} \quad g \sim P(g) \tag{1}$$

$$\textit{Joint Planning:} \quad \pi \sim P(\pi|g) \tag{2}$$

$$\textit{Utterance Model:} \quad u \sim P(u|\pi) \tag{3}$$

$$\textit{Action Selection:} \quad a_{1,t}, a_{2,t} \sim P(a_{1,t}, a_{2,t}|\pi) \tag{4}$$

A peculiar aspect of this model is that the principal is assumed to communicate the plan π through an utterance u, despite both agents supposedly having shared mental states: Why communicate the plan, if everybody knows what it is?

Nonetheless, there are good reasons for using this as a model for the purposes of goal inference. First, since our group agent model is much simpler than the complete model described earlier, it can serve as a *resource-rational approximation* (Lieder & Griffiths, 2020) of the true dynamics of multi-agent communication, allowing third-person observers to infer the goals of cooperating teams while avoiding the need to represent and infer the mental states of individual agents (as in Shum et al. (2019)). Second, our model is highly plausible as an *Imagined We* model from the perspective of an assistive agent who is part of team. In the Imagined We

model UTTERANCE-MODEL(π)
  parameters: $p_{\text{communicate}}, \mathcal{E}$
  $a^*_{1:t} \leftarrow$ ROLLOUT-POLICY(π)
  $\alpha_{1:k} \leftarrow$ EXTRACT-SALIENT-ACTIONS($a^*_{1:t}$)
  $p \leftarrow p_{\text{communicate}}$ **if** $(k > 0)$ **else** $(1 - p_{\text{communicate}})$
  $c \sim$ BERNOULLI(p)
  if $c =$ TRUE **then**
    $u \sim$ LANGUAGE-MODEL($\alpha_{1:k}, \mathcal{E}$)
  end if
end model

(a) Utterance model $P(u, c|\pi)$ as a probabilistic program

Input:	`(handover robot human key2) where (iscolor key2 blue)`
Output:	Hand me the blue key.
Input:	`(unlockr robot key1 door1) where (iscolor door1 red)`
Output:	Can you unlock the red door for me?
Input:	`(handover robot human key1) (handover robot human key2) where (iscolor key1 green) (iscolor key2 red)`
Output:	Can you pass me the green and the red key?

(b) Paired examples $\mathcal{E}$ of salient actions $\alpha_{1:k}$ and utterances u

Figure 2: Our utterance model is a probabilistic program (a) that extracts salient actions $\alpha_{1:k}$ from a joint plan π, then samples an utterance u using a language model (in our case, GPT-3 CURIE) given $\alpha_{1:k}$ and few-shot examples $\mathcal{E}$ in its prompt. Several examples are shown in (b).

(IW) framework, cooperative agents avoid excess recursive reasoning by imagining themselves to be part of a group agent with a shared goal, albeit a goal that may be unknown to individual members (Tang et al., 2020; 2022). To act, group members infer their shared goal by asking, "What is it that *we* want, that best explains our actions so far?" They then direct their actions towards the inferred goal, resulting in decentralized goal convergence. Building upon recent work that applies the IW framework to pragmatic communication (Stacy et al., 2021), our model effectively extends the listener component of the IW framework to account for joint communication and action: To infer a shared goal, assistive agents ask, "What is that *we* want, that best explains our actions *and* instructions?"

2.2. Model Components

Having defined the high level structure of our model, we now describe its individual components. For the goal prior, $P(g)$, we use a uniform distribution over a fixed set of possible goals $g \in G$. In the context of our environment (Figure 1a), a goal g corresponds to the human picking up one of the four colored gems.

To model joint planning in an efficient manner, we make the assumption that agent's actions are *ordered* — i.e., the agents take turns, with the principal (human) acting at each step t while the assistant waits, before the assistant (robot) acts at $t + 1$ while the principal waits. This limits the branching factor of planning, while preserving the optimal solution (Boutilier, 1996). Under this assumption, we model joint planning as the process of computing a joint Boltzmann policy π for a goal g:

$$\pi(a_{i,t}|s_t, g) = \frac{\exp \frac{1}{T} Q^*_g(s_t, a_{i,t})}{\sum_{a'_{i,t}} \exp \frac{1}{T} Q^*_g(s_t, a'_{i,t})} \tag{5}$$

where s_t is the current state, $a_{i,t}$ is the action taken by agent i at s_t, T is a temperature parameter, and $Q^*_g(s_t, a_{i,t})$ is the (negated) cost of the optimal plan from s_t to goal g with $a_{i,t}$ as its first action. This models a team that is *noisily optimal* in how it acts, with the amount of noise controlled by T. Importantly, $Q^*_g(s_t, a_{i,t})$ need not be computed in advance, but can instead be computed online for each action $a_{i,t}$ and state s_t observed during the inference. We do this using real-time adaptive A* search as an incremental shortest-path planner (Koenig & Likhachev, 2006), avoiding the prohibitive cost of computing a Q-value for every state and action via value iteration (used by related inverse reinforcement learning algorithms, e.g. Ramachandran & Amir (2007); Ziebart et al. (2008)), while using the Q-values computed by previous A* searches to inform future searches.

With the policy π computed, we model action selection by sampling actions according to the policy. In addition, we can use π to model the instruction u that the principal agent utters according to the following process:

1. Rollout the policy π with temperature $T = 0$ to get an optimal sequence of actions $a^*_{1:t}$ to the goal.
2. Extract *salient* actions $\alpha_{1:k}$ from $a^*_{1:t}$ to be communicated to the assistive agent.
3. Generate a natural language instruction or request that communicates the salient actions $\alpha_{1:k}$ (or avoid communicating if there are none).

We implement the above process as a probabilistic program that combines deterministic, stochastic and neural components, shown in Figure 2(a). Steps 1 and 2 are deterministic, with step 2 implemented by filtering out non-salient actions like directional movement, and keeping only important actions for the assistant to perform, such as handing over keys or unlocking doors. This approximates a pragmatic speaker in the RSA framework (Goodman et al., 2008), communicating instructions that trade-off informativeness and utterance cost by mentioning only the most relevant actions to achieving the team's shared goal. Step 3 has two parts: (i) if there are $k > 0$ salient actions to communicate, the program decides with high probability to communicate an utterance, with this choice denoted by c; (ii) if this occurs (i.e. $c =$ TRUE), then the utterance u is generated using a neural language model, conditioned on both the salient actions $\alpha_{1:k}$ and a series of few-shot examples $\mathcal{E}$ that are included in the prompt (Figure 2(b)). In our implementation, we use the CURIE variant of GPT-3 (Brown et al., 2020) to serve

as the utterance likelihood $P(u|\alpha_{1:k}, \mathcal{E})$, since we found it reasonably calibrated when evaluating the probability of an utterance u, and did not require the more realistic forward-generation abilities of larger language models. However, any language model that defines a probability distribution over string tokens can in principle be used.

2.3. Goal Inference from Actions and Instructions

Using the model described above, our aim is to compute the posterior distribution over goals g given an instruction u, whether an instruction was communicated c[1], and a series of actions $a_{i,1:t}$ for each agent i:

$$P(g|u, c, a_{1,1:t}, a_{2,1:t}) \propto P(g, u, c, a_{1,1:t}, a_{2,1:t}) = P(g)P(u, c|\pi_g)\prod_{\tau=1}^{t} P(a_{1,\tau}|\pi_g)P(a_{2,\tau}|\pi_g) \quad (6)$$

Since all the terms in the joint distribution can be computed exactly[2], we can perform exact Bayesian inference by updating the unnormalized weights w_t^g for each goal g as new evidence arrives, then normalizing the weights to get the probability P_t^g for goal g at timestep t:

$$w_0^g \leftarrow P(g)P(u, c|g)$$
$$w_t^g \leftarrow w_{t-1}^g P(a_{1,\tau}|\pi_g)P(a_{2,\tau}|\pi_g)$$
$$P_t^g \leftarrow w_t^g / \textstyle\sum_{g'}(w_t^{g'})$$

We implement this inference algorithm as an exact variant of Sequential Inverse Plan Search (Zhi-Xuan et al., 2020) using the particle filtering extension of the Gen probabilistic programming system (Cusumano-Towner et al., 2019; Zhi-Xuan, 2020), which can be configured to support exact inference by disabling random sampling, while still automating all the necessary weight updates.

3. Experiments

To evaluate both the scientific validity and performance of our model, we conducted a human and computational experiments, comparing our model's goal inferences against goal inferences elicited from humans. As a baseline, we used an "Actions Only" model that does not model or condition upon uttered instructions. Below we describe the environment we used to conduct our experiments, the dataset of action and instruction stimuli we generated, followed by the human experiment and model fitting procedures.

3.1. Environment Description

In order to study goal inference in a multi-agent setting, we adapt the Doors, Keys, & Gems gridworld from Zhi-Xuan et al. (2020) into a multi-agent environment, where a human principal and a robot assistant collaborate to retrieve a target gem (Figure 1(a)). In this environment, agents may need to pick up keys and unlock doors so as to reach desired items. Keys and doors are colored, and agents can only unlock a door using a key of the same color, after which the key is exhausted. To allow for cooperative behavior, agents may pass held items to each other if they are on adjacent grid cells. In addition, the robot is not allowed to pick up gems, reflecting its role as an assistive agent. If they have nothing useful to do, agents may also wait at their current location, which carries 60% the cost of other actions.

3.2. Dataset Generation

We constructed 6 instances of the multi-agent Doors, Keys, & Gems environment with varying maze designs and item locations. In each of these environment instances, we created 2–4 action sequences from the initial state to a goal gem, generated through a combination of automated planning and manual modification in order to increase the diversity and goal ambiguity associated with each action sequence. For each action sequence, we wrote a natural language instruction that the human might communicate to the robot in a variety of styles (e.g. requests like *"Can you unlock the red and blue door for me?"*, or commands like *"Pass me the blue key."*). Instructions are all communicated at the beginning of each action sequence. In total, we constructed 20 stimuli of action trajectories paired with instruction utterances, reflecting a range of cooperative scenarios.

3.3. Human Experiment Design

Our human experiment involved two experimental conditions in order the isolate the effect of language information on goal inference: (i) with-instructions and (ii) without-instructions condition. In the with-instructions condition, participants were shown animated trajectories of the human-robot team as stimuli, along with the instructions given by the human to the robot at the initial state (see Figure 3 for selected frames). In the without-instructions condition, the participants were shown the same animated stimuli, but without the instruction. Animated trajectories were segmented at certain judgment points, and participants provided their goal inferences at each judgment point by selecting all gems they thought were likely to be the team's goal (see Appendix). We inserted 4–5 judgment points for each trajectory, depending on the length of the trajectory.

3.4. Participants

We recruited 120 US participants fluent in English via Prolific (age 19–62 with mean 34; 49 women, 67 men, 4 non-binary), 60 of whom were assigned to each of the experimental conditions. Participants were paid at a rate of 15

[1] Note that c must be true if u is observed.

[2] Since π is deterministic given g in our model, we omit $P(\pi|g)$ from the expression and replace π with π_g.

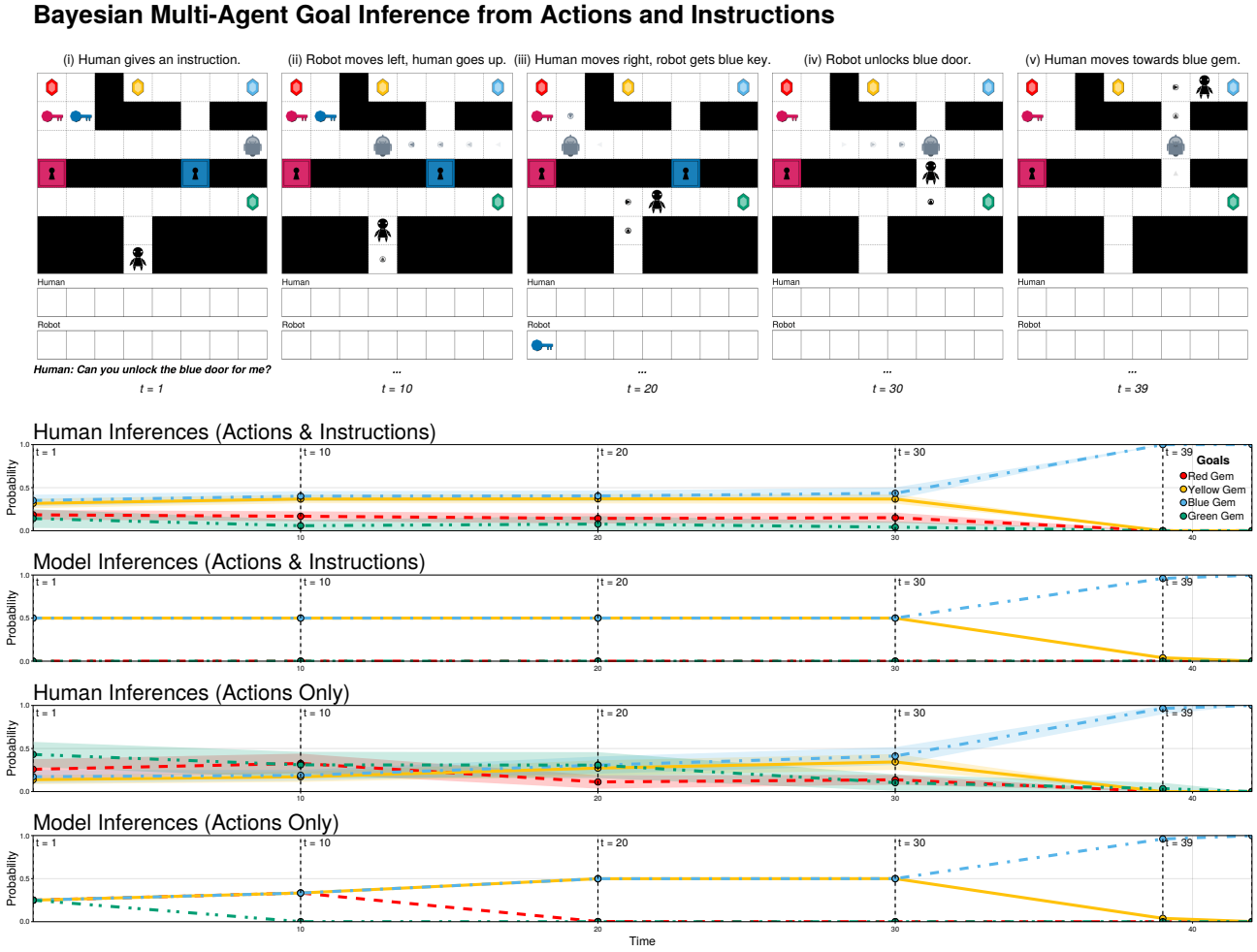


Figure 3: Goal inferences over time from the actions and instructions of a human-robot team, where the team's goal is for the human to pick up one of the four colored gems. *(Row 1)* Frames from an illustrative state-action trajectory $(s_{1:t}, a_{1:t})$, with the initial utterance $u =$ *"Can you unlock the blue door for me?"* shown below the first frame. *(Row 2)* Average human inferences (w. 95% CI) given both the instruction and actions, elicited at the selected frames. *(Row 3)* Model inferences via Bayesian inverse planning from instructions and actions. *(Row 4)* Average human inferences (w. 95% CI) given actions only, without any instructions provided. *(Row 5)* Model inferences via Bayesian inverse planning from actions only.

USD per hour. Each participant provided goal inferences for 10 out of the 20 stimuli. In total, each stimulus was rated by approximately 30 participants.

Before viewing the stimuli, participants went through a tutorial and answered five comprehension questions, which all participants passed. Participants also earned points proportional to their Brier skill score, a measure of well-calibrated prediction (Weigel et al., 2007), and were paid a bonus for earned points ($1 to $3 per participant on average) to incentivize good-faith effort at inferring the goal. One participant in the with-instruction condition was excluded from our analysis due to an low total point score below the first quartile (Q1) minus the inter-quartile range (IQR).

3.5. Computational Experiments and Model Fitting

We ran Bayesian goal inference with our model on the same set of stimuli we provided to humans. We fixed the probability of communicating an utterance for plans with salient actions to $p_{\text{communicate}} = 0.95$, and also fixed the set of 7 few-shot examples $\mathcal{E}$ used by our utterance model, leaving no other free parameters besides the Boltzmann policy's temperature T, which we varied from 0.0625 to 16 in powers of two. Our model was run on both the stimuli including the instructions and without the instructions, with the latter serving as a baseline model that only computes goal inferences from action observations. To fit the model, we computed Pearson's correlation coefficient R between model inferences at each judgment point vs. average human inferences

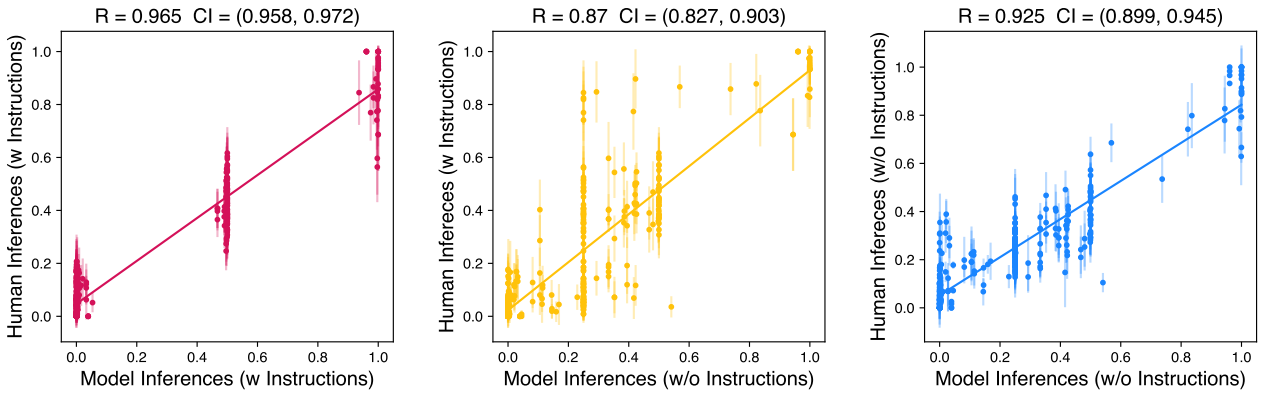


Figure 4: Correlation plots between average human inferences and model inferences, in the with-instructions condition *(left)* and without-instructions condition *(right)*. Our model displays a good fit in both conditions. In contrast, the without-instructions model correlates less well with human inferences from both actions and instructions *(middle)*.

at the same judgment point. We found that $T = 1.0$ led to high correlation in both conditions, achieving the highest geometric mean of R across conditions. As such, the following results all use $T = 1.0$ unless stated otherwise.

3.6. Results

Qualitative Analysis Figure 3 shows an illustrative example of multi-agent goal inference from actions and instructions over time. In this example, the human principal is able to directly reach the green gem, but the other three gems, along with the robot, are locked behind a red door and a blue door. As such, the human principal requires the robot's help to unlock one of the doors in order to reach one of those gems. The red gem is closer to the red door, while the yellow and blue gems are closer to the blue door. Without instructions (i.e. observing actions only), both humans and our model place a (close-to)-uniform prior over the four gems. But with instructions, both our model and humans place higher probability on the yellow and blue gems at $t = 1$, while down-weighting the red and green gems. This is because a rational agent would not utter an instruction for the green gem, and would instruct the robot to unlock the red instead of blue door if the goal was the red gem.

As the example progresses through $t = 10$ (robot moves toward the key) and $t = 20$ (robot picks up the blue key), the team's actions gradually provide information about the goal, such that the action-only model eventually stops considering the green gem and red gem as live possibilities. Humans are slightly more uncertain when shown only actions, down-weighting the red gem, but maintaining the possibility that the green gem might be the goal even at $t = 20$. In contrast, this goal uncertainty is considerably reduced when instructions are provided at the start, with lower uncertainty in humans manifesting not just as faster convergence to the true goal, but also smaller confidence intervals (reflecting lower population variance). Only at $t = 30$ (robot unlocks the door) do human inferences from actions alone converge towards human inferences with language instructions. From this point onward, the team's actions have revealed enough information that goal inferences are the same regardless of the initial information provided by language.

Correlational Analysis To quantitatively evaluate the fit between our model's inferences and human participants' inferences, we run a correlational analysis and show the results across the two experimental conditions in Figure 4. As can be seen, our model's goal inferences are strongly correlated with human judgments in both experimental conditions, with a Pearson's R of 0.965 (95% CI of [0.958, 0.972]) in the with-instructions condition (left plot), and a Pearson's R of 0.925 (95% CI of [0.899, 0.945]) in the without-instructions condition (right plot). The 95% confidence intervals are calculated through bootstrapping with 1000 samples.

Figure 4 also shows that the correlation coefficient is higher in the with-instructions condition. We suspect that this is because language instructions reduce the uncertainty and variance in human inferences, leading to a better fit. The left-most plot of Figure 4 also shows that the probability ratings form small clusters around values of 0, 0.5, and 1, whereas the data points in the without-instruction condition have a much wider spread. This indicates that instructions help the observer effectively reduce the set of possible goals to just two or one goals. As a baseline comparison, we plot the correlation between our model's inferences without instructions and human inferences with instructions (middle plot). This resulted in a lower correlation coefficient of 0.87, with a 95% CI of [0.827, 0.903], demonstrating the importance of modeling inference from instructions to explain human goal inferences.

	t = first		t = median		t = last	
	$P(g_{\text{true}})$	Brier Score	$P(g_{\text{true}})$	Brier Score	$P(g_{\text{true}})$	Brier Score
Humans (with instructions)	0.51 (0.06)	0.10 (0.06)	0.56 (0.21)	0.10 (0.04)	0.94 (0.08)	0.01 (0.01)
Humans (without instructions)	0.23 (0.05)	0.20 (0.03)	0.44 (0.15)	0.13 (0.05)	0.92 (0.02)	0.00 (0.00)
Model (with instructions)	0.64 (0.23)	0.09 (0.06)	0.65 (0.23)	0.09 (0.06)	0.99 (0.02)	0.00 (0.00)
Model (without instructions)	0.25 (0.00)	0.19 (0.00)	0.55 (0.17)	0.11 (0.04)	0.98 (0.04)	0.00 (0.00)

Table 1: Goal inference metrics for humans (averaged across subjects) and our model across both experimental conditions. We report the probability assigned to the true goal $P(g_{\text{true}})$ and Brier score (lower is better) at the initial, median, and final judgment points ($t \in \{\text{first}, \text{median}, \text{last}\}$). Values are averaged over all stimuli, and standard deviations are in brackets.

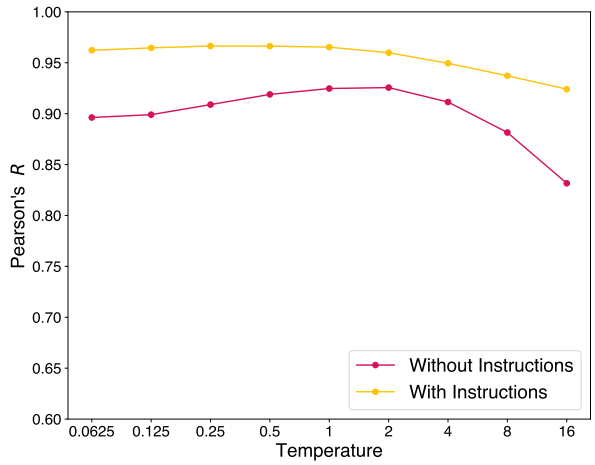


Figure 5: Correlation between model and human inferences under different model temperatures. In both conditions, the models show good fit but the correlation coefficients start to decline with temperature values greater than 2.

Sensitivity Analysis To evaluate the robustness of our correlational results to changes in model parameters, we also performed a sensitivity analysis, computing correlations for every value of T we tested. Results of this analysis are shown in Figure 5. We find that our model correlates highly with humans under many temperature settings, though correlations eventually decrease once temperature becomes sufficiently large. In the without-instructions condition, the correlation increases initially and peaks at temperature value of 2 before declining, while in the with-instructions condition, the performance starts to decline after a temperature value of 1. This suggests that humans tend to expect a small to moderate amount of action noise in this setting.

Goal Inference Accuracy Finally, we analyze the performance of both humans and our model in terms of accurately inferring the team's true goal. At each judgment point (which includes the initial state where no actions are shown), we compute the probability assigned to the true goal $P(g_{\text{true}})$ by both our model and the average human, along with the Brier score $\sum_i (P(g_i) - \mathbf{1}[g_i = g_{\text{true}}])^2$, a measure of well-calibrated inference. We then average these values across the entire dataset at each judgment point. We show these results for the initial, median, and final judgment points in Table 1. (In cases where a stimulus has no midpoint, we average the true goal probabilities for the middle two judgment points.)

We find that, on average, our model assigns slightly higher probability to the true goal than humans ($p < 0.001$ in both experimental conditions via a paired t-test), indicating that our model is able to effectively infer the team's goals for the tested stimuli . Notably, there is a very clear difference between experimental conditions at the initial timestep: When instructions are provided at the start, both humans and the model assign much higher probability to the true goal than when no instructions are observed. This illustrates the importance of language in conveying useful information rapidly, as compared to inferring goals from actions alone.

Since instructions provide a lot of information, we also see that the true goal probability at the median judgment point only increases marginally in the with-instructions condition, with the actions observed between the first and median points providing limited extra information. In contrast, when only actions are observed, the increase in $P(g_{\text{true}})$ is much more pronounced. However, instructions alone are not enough to disambiguate the goal: As inference progresses on to the last judgment point, goal accuracy is significantly higher than at the first judgment point (corresponding to an "Instructions Only" baseline) and median judgment point.

Lastly, we analyze how instructions affect variance in goal inferences among our participants. By computing the standard deviation of at each judgment point among human participants, we observe a lower standard deviation in human ratings (SD = 0.150) when instructions are provided, compared to the condition without instructions (SD = 0.188), and find that this difference is statistically significant ($t = 7.71$, $p < 0.001$). This indicates that instructions do not only improve average accuracy, but also reduce individual variability when humans infer others' goals.

4. Related Work

In addition to Bayesian theory-of-mind, rational speech act theory, and the Imagined We approach to cooperation, our model is closely related to several other lines of research:

Multimodal Goal Inference and Reward Learning. Inferring goals can be framed as online inverse reinforcement learning (IRL), where the aim is to infer a reward function that explains an agent's behavior in a *single* episode (Jara-Ettinger, 2019). Our model is hence related to but distinct from IRL methods that learn reward functions from paired datasets of instructions and demonstrations (Williams et al., 2018; Tung et al., 2018; Fu et al., 2019). Most closely related is *reward-rational implicit choice* (Jeon et al., 2020), a framework for multimodal Bayesian reward learning from multiple types of human feedback, including demonstrations and language. Indeed our architecture can be viewed as a practical instantiation of this framework for those modalities, using our LLM utterance model (Figure 2) as the (inverse) grounding function between utterances and trajectories.

Value Alignment and Assistance Games. The principal-assistant team setting that we study is inspired by the formalization of human-AI value alignment as *assistance games* (Hadfield-Menell et al., 2016). While our focus here is on modeling how an external observer would infer the goal of such a team, following the Imagined We approach, the same group agent model might also be used by the assistant to infer the *joint* goal and plan that the principal has in mind[3]. As such, our model can be seen as an alternate approach to approximately solving assistance games, requiring less recursion than iterated best-response (Hadfield-Menell et al., 2016), and more tractability than solving for the the pragmatic-pedagogical equilibrium (Fisac et al., 2017).

Instruction Following with Language Models. Our work builds upon a long tradition of grounded instruction following from natural language (Tellex et al., 2011), leveraging an LLM utterance likelihood to achieve wide coverage over utterances without task-specific training. Whereas numerous papers have used LLMs to translate natural language to actions (Ahn et al., 2022) or task specifications (Kwon et al., 2023; Yu et al., 2023; Liu et al., 2022), our approach is best viewed as a successor to Bayesian approaches such as Squire et al. (2015), using LLMs in place of classical BoW models for natural language commands.

5. Discussion and Future Work

In this paper, we extended prior models of Bayesian goal inference to a multi-modal, multi-agent setting. Experiments demonstrate that our model can explain human inferences of team goals in a range of different cooperative scenarios, and that both our model and humans can effectively infer goals of cooperating agents in a multi-agent setting. Importantly, linguistic communication provides highly useful information that enables observers to more reliably infer a team's goal. As communication is ubiquitous in multi-agent contexts, our contributions provide a means to better modeling, understanding, and inferring shared plans and goals, which are crucial for applications in human-AI collaboration.

However, although our model shows human-like performance on our tested stimuli, it is still limited to relatively simple multi-agent scenarios. For instance, our model assume that agents follow a Boltzmann-rational policy, which does not account for bounded rationality (Alanqary et al., 2021) or other systematic deviations from optimality (Shah et al., 2019). As a result, we have anecdotally found that goal inference can be sensitive to some deviations from optimal behavior, such as the human simply staying idle for extra timesteps. In real-life scenarios, humans may not act optimally, or may have certain preferences when dividing tasks among team members (e.g. they may prefer minimizing their movements and assigning more work to the assistive agent). To address these issues, we aim to make our model more robust to boundedly-rational behavior, and to account for different team structures and preferences.

Along similar lines, our current model of communicative team utterances is at once somewhat heuristic and overly rational. The heuristic aspect is that we have manually defined which actions are salient in order to approximate pragmatic communication, but ideally this could be done in more principled manner that accounts for a wider diversity of utterances. For example, a more thoughtful speaker might come up with multiple plans that a listener might be expecting, and then communicate those actions that are *most unique* to the plan that the speaker actually has in mind. The overly-rational aspect is that we have modeled utterances as communicating actions that come from *only* the optimal lowest-cost plan. But real-world speakers are boundedly-rational at best, and may either accidentally omit salient actions from the plan, or communicate actions from a less optimal plan. Figuring out how to model these boundedly-rational speech acts is an important line of future work.

In sum, we are still some ways away from a complete model of rational communicative cooperation, not least due the many theoretical and technical challenges involved in defining and implementing reasonably optimal communicative behavior. Nonetheless, we have made a number of important steps in this direction by combining aspects of both Bayesian theory of mind and rational speech act theory into a combined model of sequential action and communication, while also using the power of probabilistic programming, model-based planning, and neural language models to go beyond the toy models that approaches like ours have previously been limited to. Using the technical infrastructure we have introduced in this paper, we look forward to addressing the challenges we have laid out above, and many more still.

[3]With the modification that the assistant should not condition on their own actions to infer the shared goal.

Acknowledgements

This work was funded in part by the DARPA Machine Common Sense, AFOSR, and ONR Science of AI programs, along with the MIT-IBM Watson AI Lab. Tan Zhi-Xuan is funded by an Open Phil AI Fellowship.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., and Tenenbaum, J. B. Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*, 2021.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, 2017.

Berke, M., Belledonne, M., and Jara-Ettinger, J. Learning a metacognition for object perception. *arXiv preprint arXiv:2011.15067*, 2020.

Boutilier, C. Planning, learning and coordination in multiagent decision processes. In *TARK*, volume 96, pp. 195–210. Citeseer, 1996.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cusumano-Towner, M. F., Radul, A., Wingate, D., and Mansinghka, V. K. Probabilistic programs for inferring the goals of autonomous agents. *arXiv preprint arXiv:1704.04977*, 2017.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., and Mansinghka, V. K. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 221–236. ACM, 2019.

Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R. G., Wu, Y., Michalewski, H., Saurous, R. A., Sohl-Dickstein, J., et al. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.

Evans, O., Stuhlmüller, A., Salvatier, J., and Filan, D. Modeling Agents with Probabilistic Programs. http://agentmodels.org, 2017. Accessed: 2019-10-25.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., and Dragan, A. D. Pragmatic-pedagogic value alignment. *arXiv preprint arXiv:1707.06354*, 2017.

Fu, J., Korattikara, A., Levine, S., and Guadarrama, S. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

Gergely, G. and Csibra, G. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, 7(7):287–292, 2003.

Goodman, N. D. and Frank, M. C. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Goodman, N. D. and Stuhlmüller, A. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154, 2008.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pp. 3909–3917, 2016.

Jara-Ettinger, J. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29: 105–110, 2019.

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

Koenig, S. and Likhachev, M. Real-Time Adaptive A*. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pp. 281–288, 2006.

Kwon, M., Xie, S. M., Bullard, K., and Sadigh, D. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=10uNUgI5Kl.

Lew, A. K., Tessler, M. H., Mansinghka, V. K., and Tenenbaum, J. B. Leveraging unstructured statistical knowledge

in a probabilistic language of thought. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2020.

Li, B. Z., Chen, W., Sharma, P., and Andreas, J. Lampp: Language models as probabilistic priors for perception and action. *arXiv e-prints*, pp. arXiv–2302, 2023.

Lieder, F. and Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

Liu, J. X., Yang, Z., Schornstein, B., Liang, S., Idrees, I., Tellex, S., and Shah, A. Lang2LTL: Translating natural language commands to temporal specification with large language models. In *Workshop on Language and Robotics at CoRL 2022*, 2022.

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

Seaman, I. R., van de Meent, J.-W., and Wingate, D. Nested reasoning about autonomous agents using probabilistic programs. *arXiv*, pp. arXiv–1812, 2018.

Shafto, P., Goodman, N. D., and Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89, 2014.

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. D. On the feasibility of learning, rather than assuming, human biases for reward inference. *arXiv preprint arXiv:1906.09624*, 2019.

Shum, M., Kleiman-Weiner, M., Littman, M. L., and Tenenbaum, J. B. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6163–6170, 2019.

Squire, S., Tellex, S., Arumugam, D., and Yang, L. Grounding english commands to reward functions. In *Robotics: Science and Systems*, 2015.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., and Gao, T. Modeling communication to coordinate perspectives in cooperation. *arXiv preprint arXiv:2106.02164*, 2021.

Tang, N., Stacy, S., Zhao, M., Marquez, G., and Gao, T. Bootstrapping an imagined we for cooperation. In *CogSci*, 2020.

Tang, N., Gong, S., Zhao, M., Gu, C., Zhou, J., Shen, M., and Gao, T. Exploring an Imagined "We" in human collective hunting: Joint commitment within shared intentionality. In *Proceedings of the annual meeting of the cognitive science society*, volume 44, 2022.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., and Roy, N. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 1507–1514, 2011.

Tung, H.-Y., Harley, A. W., Huang, L.-K., and Fragkiadaki, K. Reward learning from narrated demonstrations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7004–7013, 2018.

Weigel, A. P., Liniger, M. A., and Appenzeller, C. The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135(1):118–124, 2007.

Williams, E. C., Gopalan, N., Rhee, M., and Tellex, S. Learning to parse natural language to grounded reward functions with weak supervision. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4430–4436. IEEE, 2018.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., and Kleiman-Weiner, M. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.

Yu, W., Gileadi, N., Fu, C., Kirmani, S., Lee, K.-H., Arenas, M. G., Chiang, H.-T. L., Erez, T., Hasenclever, L., Humplik, J., et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.

Zhi-Xuan, T. GenParticleFilters.jl, 2020. URL https://github.com/probcomp/GenParticleFilters.jl.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., and Mansinghka, V. Online bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems*, 33, 2020.

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

A. Experiment Interface

Figure 6 shows the interface for our human experiments, which is adapted from Alanqary et al. (2021). The interface displays animated stimuli which pause at specific judgment points. At each judgment point, participants provide goal inferences by selecting the gems(s) they believe to be the team's most likely goal, then proceed to next segment. Responses are converted to uniform distributions over the selected goals (e.g., if three goals were selected, 33.3% probability is allocated to each goal). If no goal seems more likely than the others, participants can choose the *All Equally Likely* option.

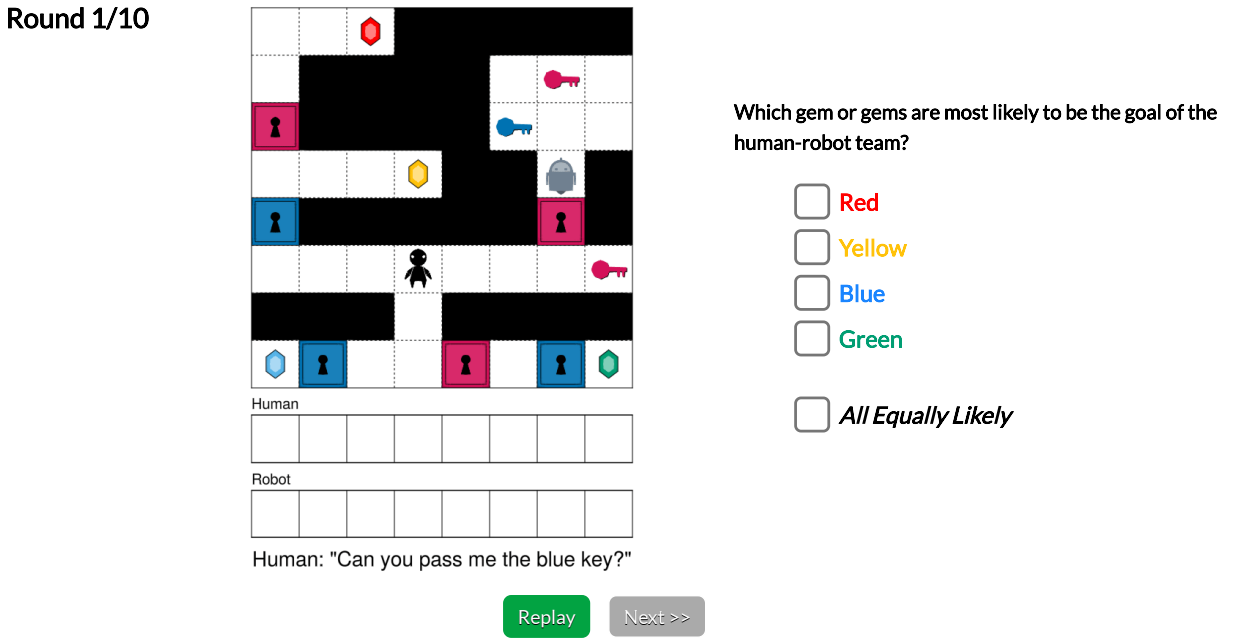


Figure 6: Interface for the "with instructions" condition of our human experiments.

B. Experiment Stimuli and Additional Results

In the supplementary information, we provide both storyboard plots and animated GIFs showing the 20 stimuli we used in our human experiment. We also provide goal inference storyboards over time for each stimulus, similar to Figure 3. This information can also be accessed at `https://osf.io/gh758/`.