Grounding Language about Belief in a Bayesian Theory-of-Mind

Lance Ying[*12], **Tan Zhi-Xuan**[*1], **Lionel Wong**[1], **Vikash Mansinghka**[1], **Joshua B. Tenenbaum**[1]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Harvard University, Cambridge, MA, USA

Abstract

Despite the fact that beliefs are mental states that cannot be directly observed, humans talk about each others' beliefs on a regular basis, often using rich compositional language to describe what others think and know. What explains this capacity to interpret the hidden epistemic content of other minds? In this paper, we take a step towards an answer by grounding the semantics of belief statements in a Bayesian theory-of-mind: By modeling how humans jointly infer *coherent* sets of goals, beliefs, and plans that explain an agent's actions, then evaluating statements about the agent's beliefs against these inferences via epistemic logic, our framework provides a functional role semantics for belief, explaining the gradedness and compositionality of human belief attributions, as well as their intimate connection with goals and plans. We evaluate this framework by studying how humans attribute goals and evaluate belief sentences while watching an agent solve a doors-and-keys gridworld puzzle that requires instrumental reasoning about hidden objects. In contrast to pure logical deduction, non-mentalizing baselines, and mentalizing that ignores the role of instrumental plans, our model provides a much better fit to human goal and belief attributions, demonstrating the importance of theory-of-mind for modeling how humans understand language about beliefs.

Keywords: theory-of-mind, belief modeling, social cognition, epistemic language, semantics

Introduction

Language about belief is pervasive in social life. Whether one is telling a friend about a mutual friend's aspirations (*"She seems to think she'll get the role she's applying for..."*), speculating about the insider knowledge behind an investment decision (*"Meta probably thinks this technology is going to be a big deal..."*), or listening to news commentators explain the actions of governments (*"The UK believes that this policy will improve..."*), language provides an incredibly rich medium for communicating about our mental models of the world. Remarkably, we often talk about belief despite not knowing what other people truly think, instead inferring beliefs from what they say or do. Belief, after all, is a guide to rational action, and there are many things a person would or would not do if they believed some fact to be true.

How do people understand language about belief, in light of this intimate connection? Put another way, what do people take as the *meaning* of belief statements, given the tight links between language, thought, and action? While philosophers, logicians, and linguists have long studied the semantics of belief (Gochet & Gribomont, 2006; Chisholm, 1955; Loar, 1981), many of these inquiries have focused on the relationship between a belief sentence and the proposition it embeds (Partee, 1973; Stalnaker, 1987) or how belief sentences can be nested and updated (Bolander, 2017), not how people use and interpret such sentences in everyday social contexts. On the other hand, computational cognitive scientists have developed numerous models that explain how people attribute beliefs and goals to other agents (Baker et al., 2009; Jara-Ettinger et al., 2019; Houlihan et al., 2023; Pöppel & Kopp, 2019; Zhang et al., 2023). Building upon the work of Baker et al. (2017), these models formulate mental state attributions as the product of a Bayesian theory-of-mind (BToM): Since beliefs and goals play a *functional* role in guiding plans and actions (Dennett, 1989), it is possible to *interpret* an agent's actions as driven by particular goals and beliefs through Bayesian inference. However, prior work in this paradigm has not studied the rich compositional nature of belief sentences, or how humans might interpret them.

In this paper, we ground *natural language statements* of belief in the functional role semantics (Harman, 1982) afforded by a Bayesian theory-of-mind. In our approach, belief statements are not just claims about which propositions are thought to be true by an agent — which we capture with epistemic logic (Gochet & Gribomont, 2006) — but also imply *rationality constraints* on how the agent is likely to plan and act given their goals (Loar, 1981) — as captured by our Bayesian agent model. As such, our framework explains how humans can assess the likeliness of a belief statement based on how well it coheres with an agent's actions, along with the goals and intentions those actions imply. To implement this framework, we combine the strengths of machine learning methods with the coherence and precision of Bayesian and logical reasoning (Wong et al., 2023; Ying et al., 2023), using a large language model (LLM) as a tool to automatically translate natural language statements into logical form (Shin et al., 2021), then evaluating these statements with respect to the inferences produced by a probabilistic programming architecture for Bayesian inverse planning (Zhi-Xuan et al., 2020).

To evaluate this framework, we design an experiment where participants are shown animations of a player navigating a gridworld puzzle, which requires the player to collect one of four valuable gems (see Figure 1). These gems

[*]Equal Contribution

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

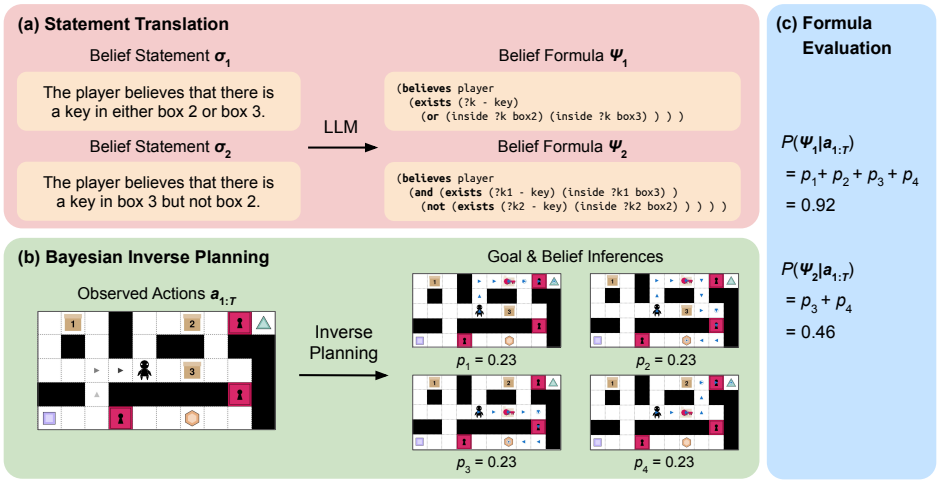

Figure 1: Overview of our proposed model. **(a)** We translate belief statements σ_i from English into first-order epistemic logic (using a large language model). **(b)** Given a series of observed actions $a_{1:T}$, we simulate possible mental states (combinations of the agent's goals, beliefs, and plans), assigning a posterior probability p_i to each hypothesis via inverse planning. **(c)** Each logical formula ψ_i can then be evaluated against the inferred mental states to produce a probability rating $P(\psi_i|a_{1:T})$.

are sometimes locked behind doors, so the player may have to collect keys located inside opaque boxes in order to reach their desired gem. For simplicity, we assume that the player *knows* where the keys are located. However, participants observing the scenario do not. As such, they have *infer* the player's goals and beliefs from the actions they observe, then provide their inferences as ratings. By comparing our model's outputs to human ratings, we evaluate how well it explains people's interpretations of language about belief.

Computational Model

To explain how human observers interpret and evaluate language about belief, our model makes use of: (i) epistemic logic as a formal compositional representation of belief statements; (ii) a Bayesian generative model that encodes the functional role that belief plays *vis a vis* an agent's goals, plans, and actions (i.e., a Bayesian ToM). These two components allow us to (iii) infer likely goals and beliefs from an agent's actions, then (iv) assess the likeliness of a belief statement with respect to those inferences.

Representing Belief Statements

As an expressive representation of belief statements and the environments they describe, we add epistemic modalities to the Planning Domain Definition Language (PDDL), a first-order language for model-based planning and reasoning (McDermott et al., 1998). In a PDDL domain, a set of predicates $\mathcal{P}$ are used to describe objects O, each of which has a type $\tau \in \mathcal{T}$. For example, to represent the fact that `key1` is red in color, we write `(iscolor key1 red)`, with predicate `iscolor` $\in \mathcal{P}$ and objects `key1`, `red` $\in O$, with types `key`, `color` $\in \mathcal{T}$ respectively. Each environment state s is essentially a set of such predicate terms, stating which relations are true or false between objects.

Since PDDL is first-order, we can evaluate the truth value of a sentence ϕ in a state s, where ϕ is compositionally defined in terms of logical operations and quantifiers, and predicates can take objects $o \in O$ or variables `?v` as arguments x_i:

$$\phi ::= (P\ x_1\ \ldots\ x_n) \mid \texttt{(not}\ \phi\texttt{)} \mid \texttt{(and}\ \phi_1\ \phi_2\texttt{)} \mid \texttt{(or}\ \phi_1\ \phi_2\texttt{)} \mid \texttt{(exists (?v -}\ \tau\texttt{)}\ \phi\texttt{)} \mid \texttt{(forall (?v -}\ \tau\texttt{)}\ \phi\texttt{)}$$

This expressivity allows us to determine the truth value of not just individual relations, but also general queries about whether some property holds for some class of objects.

However, PDDL alone cannot express claims about what an agent *believes*. To do this, we follow work in epistemic planning (Bolander, 2017; Muise et al., 2022), introducing a `believes` operator: Given some regular PDDL sentence ϕ and agent x, `(believes` x ϕ`)` means that agent x believes ϕ. As such, a statement like *"The player believes that there is a red key in box 1"* can be represented as `(believes player (exists (?k - key) (and (iscolor ?k red) (inside ?k box1))))`. This extended language corresponds to a restricted fragment of epistemic first-order logic (Gochet & Gribomont, 2006).

Modeling the Functional Role of Belief

Epistemic logic allows us to express compositional belief statements. But what do such beliefs imply about an agent's likely behavior? We formally model this functional connection with a probabilistic generative model:

$$\textit{Goal Prior:} \quad g \sim P(g) \tag{1}$$
$$\textit{State Prior:} \quad s_0 \sim P(s_0) \tag{2}$$
$$\textit{Belief Update:} \quad b_t \sim P(b_t|s_t) \tag{3}$$
$$\textit{Action Selection:} \quad a_t \sim P(a_t|b_{t-1}, g) \tag{4}$$
$$\textit{State Transition:} \quad s_t \sim P(s_t|s_{t-1}, a_t) \tag{5}$$

In this model, we assume that agents are usefully described as having beliefs b_t and goals g insofar as they take actions a_t towards the goal that are (approximately) rational given their beliefs. As such, our model defines a *functional role* for belief. Notably in our setting, we focus on the simpler case where the agent has both *complete knowledge* and *veridical perception*, such that their belief at each step t is always equivalent to the environment state s_t. Nonetheless, an observer will still have uncertainty over what the agent believes, since they have uncertainty over the initial state s_0.

What does it mean for an agent's actions to be rational with respect to their beliefs (and goals)? Following the principle of rational action (Gergely & Csibra, 2003), this means that actions should lead efficiently to a goal in the world that the agent believes it is in. In a multi-step task like ours (see Figure 1), an agent should thus *plan* to reach any instrumental subgoals necessary for their final goal (e.g. keys in boxes), then act roughly according to that plan. We capture this probabilistically by adopting a Boltzmann-rational model of action selection, given the agent's goal g and their belief b_{t-1}:

$$P(a_t|b_{t-1},g) = \frac{\exp\left(-\beta \hat{Q}_g(b_{t-1},a)\right)}{\sum_{a'} \exp\left(-\beta \hat{Q}_g(b_{t-1},a')\right)} \tag{6}$$

Here $\hat{Q}_g(b_{t-1},a)$ represents the estimated cost of the optimal plan to achieve goal g after taking action a starting at the believed state b_{t-1}. β is a parameter controlling action optimality (higher is more optimal). To compute $\hat{Q}_g(b_{t-1},a)$ efficiently, we leverage recent advances in sequential inverse planning (Zhi-Xuan et al., 2020, 2024), using real-time heuristic search (Koenig & Likhachev, 2006; Hernández et al., 2011) to rapidly estimate the cost to the goal g.

Joint Inference of Goals and Beliefs

Similar to Baker et al. (2017) and others (Farrell & Ware, 2020), we model observers as performing joint inference over the agent's goal g and belief history $b_{0:T}$ given a series of observed actions $a_{1:T}$, The corresponding posterior $P(g,b_{0:T}|a_{1:T})$ factorizes as follows:

$$P(g,b_{0:T}|a_{1:T}) \propto \textstyle\sum_{s_{0:T}} P(g)P(s_0)P(b_0|s_0) \prod_{t=1}^{T} P(a_t|b_{t-1},g)P(s_t|s_{t-1},a_t)P(b_t|s_t) \tag{7}$$

Since we assume the agent has perfect knowledge of the environment ($b_t \equiv s_t$), this reduces to inferring g and $s_{0:T}$:

$$P(g,s_{0:T}|a_{1:T}) \propto P(g)P(s_0)\textstyle\prod_{t=1}^{T} P(a_t|s_{t-1},g)P(s_t|s_{t-1},a_t) \tag{8}$$

To compute this distribution, we initialize a set of weighted samples $\{(g^i, s_0^i, w^i)\}_{i=1}^N$ by enumerating over all possible combinations of goals and initial states, where N is the number of combinations, and w^i is initialized to $P(g^i)P(s_0^i)$. Then we sequentially update each sample i, multiplying w^i by the likelihood $P(a_t|s_{t-1}^i, g^i)$ of observing action a_t, and simulating the next state s_t that results from a_t:

$$w^i \leftarrow w^i \cdot P(a_t|s_{t-1}^i, g^i) \qquad s_t \sim P(s_t^i|s_{t-1}^i, a_t)$$

Each weight w^i represents the *unnormalized* probability of the pair $(g^i, s_{0:t}^i)$ at step t. Normalizing the weights gives us the probability $P(g^i, s_{0:t}^i|a_{1:t}) = w^i / \sum_{j=1}^N w_j$

In stochastic environments, this procedure is a sequential Monte Carlo algorithm (Del Moral et al., 2006). However, since our environment is deterministic, this algorithm reduces to *exact* Bayesian filtering, which we implement efficiently as a variant of Sequential Inverse Plan Search (Zhi-Xuan et al., 2020) using the Gen probabilistic programming system (Cusumano-Towner et al., 2019).

Quantitatively Interpreting Belief Statements

Having computed the posterior $P(g,b_{0:T}|a_{1:T})$, we model the interpretation of belief statements in two steps: First, we *translate* the natural language statement σ into a a formula ψ in epistemic logic. Second, we *evaluate* the expected truth value of the translated sentence ψ with respect to inferred distribution over belief states b_T, thereby grounding the meaning of the sentence in our BToM model.

To perform the translation itself, it is possible to use a domain-specific grammar that maps English words to predicates and operators in our epistemic extension of PDDL. However, this can typically only handle a restricted fragment of natural language, while requiring significant engineering effort. Hence, we opt to use large language models (LLMs) as general purpose semantic parsers (Wong et al., 2023), which can translate natual language statements σ to symbolic forms ψ after being provided with only a few example translations (Shin et al., 2021). While not the focus of our present study, we expect this approach will enable us to approach human-level flexibility when interpreting language about belief.

Next, to evaluate the observer's credence in $\psi :=$ `(believes x ϕ)`, we extract the sentence ϕ attributed to the agent x. We then compute the expected value of ϕ being true in the belief state b_T:

$$\begin{aligned} P(\psi|a_{1:T}) &= \mathbb{E}_{b_T \sim P(b_T|a_{1:T})}[\phi \text{ is true in } b_T] \\ &= \textstyle\sum_{i=1}^N w_i \cdot [\phi \text{ is true in } b_T^i] / \sum_{j=1}^N w_j \end{aligned}$$

One notable aspect of this formulation is that the value of $P(\psi|a_{1:T})$ depends on the prior over belief states $P(b_0) = P(s_0)$. However, it is not obvious what priors $P(b_0)$ human observers bring to bear. As such, we experiment with two possibilities: $U_{S_0}(s_0)$ is uniform over the set of possible initial states $s_0 \in S_0$, and $U_\psi(s_0)$ induces a uniform prior $P(\psi) = 0.5$ over the *statement* being true. In the second case, $P(\psi|a_{1:T})$ becomes a kind of (normalized) likelihood $\bar{L}(\psi|a_{1:T}) = \frac{P(a_{1:T}|\psi)}{P(a_{1:T}|\psi)+P(a_{1:T}|\neg\psi)}$, which rates how much more evidence there is for the statement ψ as opposed to its negation $\neg\psi$. In the absence of evidence, $\bar{L}(\psi|a_{1:T}) = 0.5$.

In settings where there are too many possibilities to enumerate over, we can sample a representative set of goals and states instead.

Experiments

To evaluate our model of how humans interpret belief statements, we designed an experiment where participants had to infer the goals and beliefs of an agent solving a gridworld puzzle called Doors, Keys, & Gems (Zhi-Xuan et al., 2020). In these puzzles, an agent has to pick up keys and unlock doors to reach a valuable gem. Doors can only be opened by keys of the same color and each key can be used once. To introduce partial observability, we also added boxes, each of which might be empty or contain exactly one key.

Scenario Generation

We constructed 18 scenarios in the Doors, Keys, & Gems environment with different maze designs and item locations (see Figure 2). In each scenario, there were 4 goal gems of different shapes (triangle, square, hexagon, circle), and 2 to 3 boxes. Each scenario was paired with two English statements describing the agent's beliefs about the contents of the boxes, such as *"The player believes that there is a red key in box 1."* To test compositional language understanding, we also created negated, conjunctive, and disjunctive belief statements (e.g. *"The player believes that there is a blue or red key in either box 2 or box 3."*)

Experiment Design

Our study was conducted online through a custom interface. Participants first completed a tutorial and a comprehension quiz. During each scenario, participants rated the goals and beliefs of the observed agent at several judgement points. For goal ratings, participants were presented with a checkbox for each gem, and asked to select all gems that were likely to be the agent's goal. For belief ratings, participants were shown a scale from 1 to 7 below each belief statement, with 1 representing "Definitely False", 7 representing "Definitely True", and 4 representing "Equally Likely to be True or False". We normalized these ratings to lie between 0 and 1.

Participants

We recruited 100 US participants through Prolific (mean age = 39.57, 50 female, 49 male, 1 agender). Each participant rated 9 out of the 18 scenarios. Participants were paid US$15/hr, and received a bonus for correctly inferring the agents' goals ($\$0.05/n$ for each judgment point where they selected n goals, one of which was correct).

Alternative Models

Since our account of the semantics of belief crucially requires the ability to infer the mental states of other agents, we compare our model against baselines that either do not perform such mentalizing at all, or perform limited versions:

Omniscient Observer: Possesses complete knowledge of the validity of the belief statements. This models what epistemic logic would deduce about the agent's beliefs, given the premise that the agent knows everything about the world, and given that the observer does too.

Ignorant Observer: Rates all belief statements as false due to insufficient premises. This models what epistemic logic would (fail to) conclude about the agent's beliefs, following a semantics of negation as failure to deduce facts.

(Uncertain) Non-Mentalizing Observer: Rates all belief statements according to the prior probability that they are true, according to the uniform prior U_{S_0} over states.

Heuristic Mentalizer: Assumes that the agents will always move physically closer to their goal, ignoring instrumental actions like picking up keys. This baseline tests the importance of accounting for means-ends coherence when inferring an agent's beliefs from their actions.

Results

Qualitative Analysis

Figure 2 shows the inferences of both humans and our model in 6 illustrative scenarios. These examples demonstrate qualitatively how our model fits human data. Due to space constraints, we describe only the first 3 examples.

In the first two examples, once the player picks up a red key, both human participants and our model rate the circle to be the only likely goal. This is because the circle is the only goal that requires picking up a red key. Once the player walks down the corridor past box 1, both humans and our model rate statement 1 (*"The player believes that there is a blue key in box 1."*) to be highly improbable. This is because the player needs a blue key to reach the inferred goal and a rational agent would not walk past box 1 if they knew it contained a blue key. As the agent walks towards box 2 in Example 1, both humans and our model find it likely that there is a blue key in box 2, whereas in Example 2, once the player walks past box 2, both human participants and our model assign a higher likelihood for statement 2 (*"The player believes that there is a blue key in box 3."*)

In Example 3, once the player picks up a blue key, both humans and our model infer the square and the hexagon to be the likely goals, since they are locked behind a blue door nearby. As the player walks past the blue door and towards boxes 2 and 3, these inferences veer towards the circle. It also becomes more likely for there to be a red key than a blue key among boxes 2 and 3, since the player needs a red key to reach the goal.

Quantitative Analysis

Correlation plots between human and model belief judgments are shown in Figure 3, and a summary of key results is presented in Table 1. We find that our full BToM model has a much better fit with human ratings than the baselines. This demonstrates the importance of a theory-of-mind for a semantics of belief statements: Without accounting for the coherence between an agent's beliefs, goals, and actions, observers cannot infer an agent's beliefs, and hence cannot evaluate the probability of such belief statements. Our results also demonstrate the importance of a *probabilistic* theory-of-mind. Human judgments about what agents believe are not

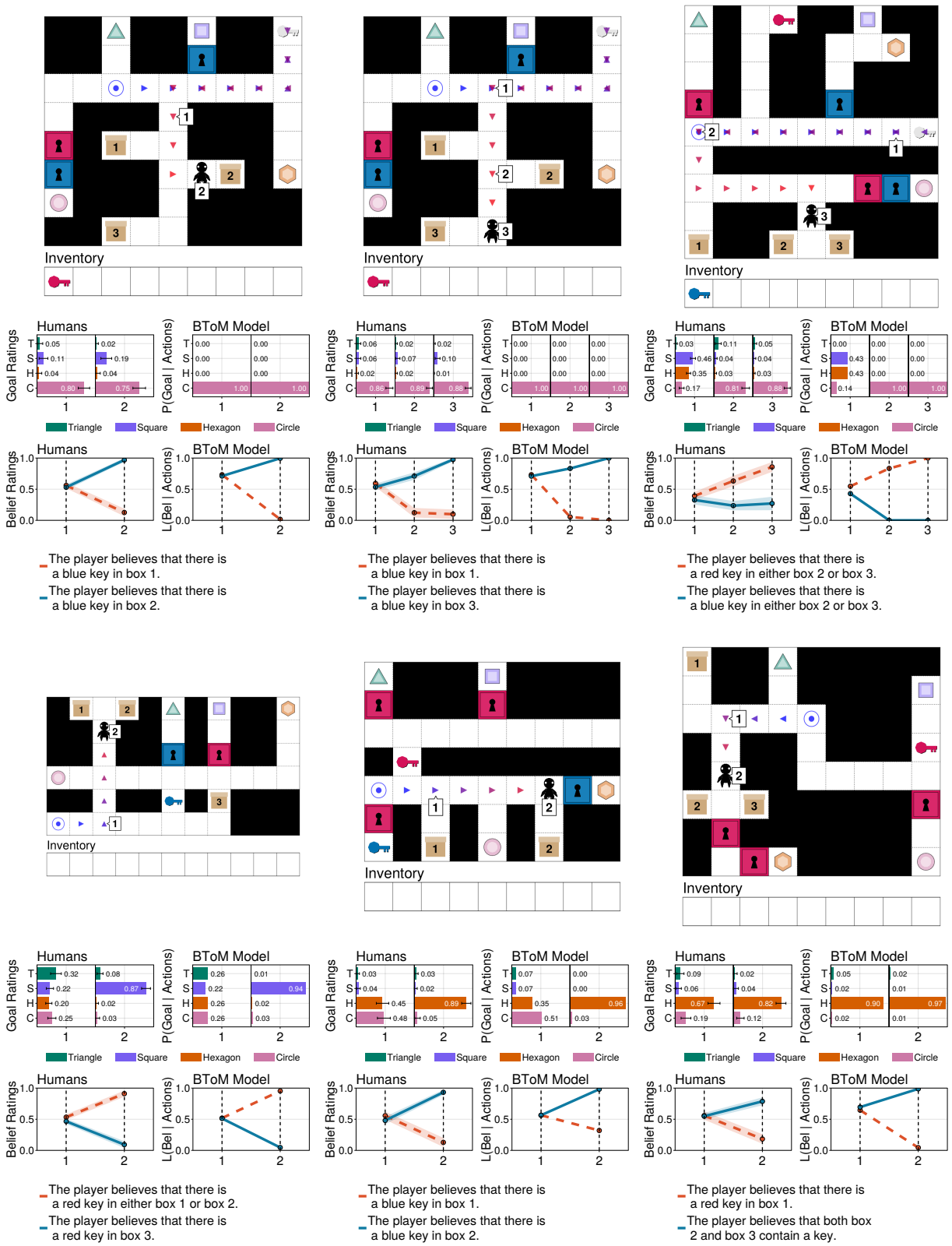

Figure 2: Human and model inferences across 6 illustrative scenarios. Judgement points are shown in the map as tooltips annotating the trajectory. Keys picked up along the trajectory are shown in gray. Below each map, we show goal inferences, followed by belief inferences (assuming U_ψ as our belief prior). Overall, our model closely matches human responses.

Table 1: Human-model correlations for both goal attributions and belief statement ratings. Bootstrap confidence intervals with a 95% confidence level are reported in brackets.

Model	Belief Prior	Goal Attributions	Belief Statements
Full BToM	Uniform over Statements (U_ψ)	0.93 [0.91, 0.94]	0.92 [0.91, 0.93]
Full BToM	Uniform over States (U_{S_0})		0.86 [0.85, 0.87]
Heuristic Mentalizer	Uniform over Statements (U_ψ)	0.17[0.13, 0.19]	0.04 [0.02, 0.06]
Heuristic Mentalizer	Uniform over States (U_{S_0})		0.19 [-0.03, 0.40]
Non-Mentalizer	Uniform over States (U_{S_0})	—	0.19 [-0.03, 0.40]
Omniscient Observer	—	—	0.64 [0.48, 0.75]
Ignorant Observer	—	—	—

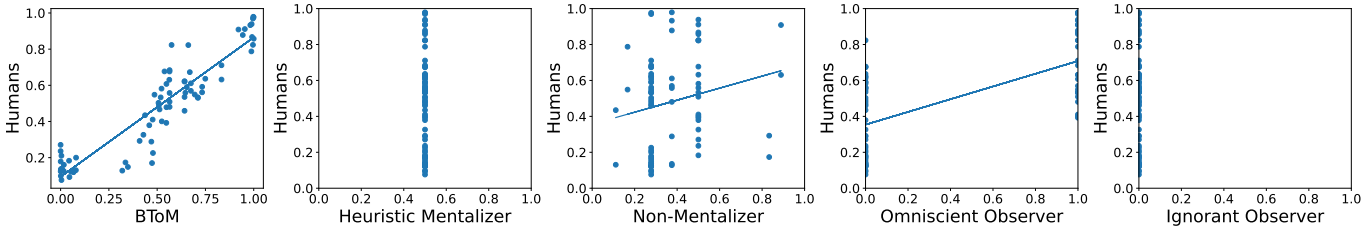

Figure 3: Correlation plots comparing belief judgments from humans (y-axis) against models (x-axis). Our full BToM model shows a significant better fit to human belief ratings than alternative models

all-or-nothing phenomena. Our account provides a semantics for such graded judgments, unlike logical models which judge statements to be either false or true.

Interestingly, although both versions of our full model correlated well with human belief judgments (Table 1, rows 1–2), the model assuming a uniform statement prior U_ψ had a significantly better fit. We found that this was because certain statements are true in *more possible worlds* (e.g. *"The agent believes that there is a red key in box 1 or box 2."* vs *"The agent believes that there is a red key in box 1."*), such that they have a higher *base rate* of being true if all states s_0 are equally probable. Our participants did not seem to take into account these base rates when providing responses. Instead, they tended to rate a statement more highly only if there was more *evidence* for that statement, consistent with our earlier suggestion that belief ratings might correspond to a normalized likelihood $\bar{L}(\psi|a_{1:T})$. While this is not Bayesian in the orthodox sense, it is semantically and pragmatically quite reasonable — it suggests that people are only willing to say that an agent believes ϕ if there is evidence for that belief.

Discussion

In this study, we explored how humans reason about other agents' beliefs and interpret beliefs communicated in natural language. Our experiment shows the deep connection between how we interpret language about belief and how we understand other agents' minds: People are able to interpret and adjust their evaluations of natural language statements as new actions come to light, demonstrating the importance of grounding such language in a theory of how agents' beliefs and goals are connected to their plans and actions. Such a theory provides a functional role for belief as a concept.

All that said, there is much work to be done before a model like ours can account for all the ways in which people use and interpret language about beliefs. A key limitation of our current model is that it only explains the *deterministic* beliefs of other agents, under the assumption that what other agents believe is what they *know*. But of course, one of the hallmarks about belief — and how we reason and talk about it — is that it can come apart from reality in all sorts of ways. People have uncertain beliefs (just like our observer), which we express in language using modals like "might" or "unlikely". People have false beliefs, and we form sentences describing them all the time (*"He thinks that ..., but actually..."*). People might be ignorant or possess only partial knowledge, and we describe this differently from mere uncertainty. Relatedly, people often think and talk about belief in abstract terms, without concretizing those beliefs into concrete worlds like our model does. If we are to explain such phenomena, we will need a much richer theory-of-mind than our model currently offers: One that models our uncertainty about other's uncertainty (Baker et al., 2017; Gmytrasiewicz & Doshi, 2005), as well as the abstract ways we represent both the world and each others' minds (Bigelow et al., 2023).

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, and gifts from Reid Hoffman and the Siegel Family Foundation. Tan Zhi-Xuan is supported by an Open Philanthropy AI Fellowship.

We thank our colleagues Brian Leahy, Cedegao Zhang, and Megan Wei for helpful discussions and suggestions during the development of this project.

References

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

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

Bigelow, E. J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. *Cognition*, *238*, 105498.

Bolander, T. (2017). A gentle introduction to epistemic planning: The del approach. *arXiv preprint arXiv:1703.02192*.

Chisholm, R. M. (1955). Sentences about believing. In *Proceedings of the aristotelian society* (Vol. 56, pp. 125–148).

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). 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).

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(3), 411–436.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Farrell, R., & Ware, S. (2020). Narrative planning for belief and intention recognition. In *Proceedings of the aaai conference on artificial intelligence and interactive digital entertainment* (Vol. 16, pp. 52–58).

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

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24*, 49–79.

Gochet, P., & Gribomont, P. (2006). Epistemic logic. In *Handbook of the history of logic* (Vol. 7, pp. 99–195). Elsevier.

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*(2), 242–256.

Hernández, C., Sun, X., Koenig, S., & Meseguer, P. (2011). Tree adaptive a. In *The 10th international conference on autonomous agents and multiagent systems-volume 1* (pp. 123–130).

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220047.

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. (2019). The naive utility calculus as a unified, quantitative framework for action understanding. *PsyArXiv*.

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

Loar, B. (1981). *Mind and meaning*. CUP Archive.

McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., ... Wilkins, D. (1998). *PDDL - the Planning Domain Definition Language.* Technical Report CVC TR-98-003/DCS TR-1165, Yale Center for Computational ....

Muise, C., Belle, V., Felli, P., McIlraith, S., Miller, T., Pearce, A. R., & Sonenberg, L. (2022). Efficient multi-agent epistemic planning: Teaching planners about nested belief. *Artificial Intelligence*, *302*, 103605.

Partee, B. H. (1973). The semantics of belief-sentences. In *Approaches to natural language: Proceedings of the 1970 stanford workshop on grammar and semantics* (pp. 309–336).

Pöppel, J., & Kopp, S. (2019). Satisficing mentalizing: Bayesian models of theory of mind reasoning in scenarios with different uncertainties. *arXiv preprint arXiv:1909.10419*.

Shin, R., Lin, C. H., Thomson, S., Chen, C., Roy, S., Platanios, E. A., ... Van Durme, B. (2021). Constrained language models yield few-shot semantic parsers. *arXiv preprint arXiv:2104.08768*.

Stalnaker, R. (1987). Semantics for belief. *Philosophical Topics*, *15*(1), 177–190.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Ying, L., Collins, K. M., Wei, M., Zhang, C. E., Zhi-Xuan, T., Weller, A., ... Wong, L. (2023). The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*.

Zhang, C., Kemp, C., & Lipovetzky, N. (2023). Goal recognition with timing information. In *Proceedings of the international conference on automated planning and scheduling* (Vol. 33, pp. 443–451).

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

Zhi-Xuan, T., Ying, L., Mansinghka, V., & Tenenbaum, J. B. (2024). Pragmatic instruction following and goal assistance via cooperative language guided inverse plan search. In *Proceedings of the 23rd international conference on autonomous agents and multiagent systems.*