Hierarchical task knowledge constrains and simplifies action understanding

Isaac Davis (isaac.davis@yale.edu), Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)
Department of Psychology, Yale University, New Haven, CT 06511

Abstract

Human social interactions require understanding and predicting other people's behavior. A growing body of work has found that these inferences are structured around an assumption that agents act rationally and efficiently in space. While powerful, this view treats action understanding in a vacuum, ignoring that much social inference happens in the context of familiar, hierarchically structured events (e.g.: buying groceries, ordering in a restaurant). We propose that social and world knowledge is critical for efficiently interpreting behavior and test this idea through a simple block-building paradigm, where participants infer an agent's sub-task (study 1a), next action (study 1b), and higher-level goal (study 1c), from very sparse observations. We compare these inferences against a Bayesian model of goal inference that exploits task structure to interpret agents' actions. This model fit participant judgments with high quantitative accuracy, highlighting how world knowledge may help support social inferences in a rich and powerful way.

Keywords: Computational modeling; Social cognition

Introduction

Many of the hallmarks of uniquely-human cognition—language, social learning, and moral reasoning, to name a few—are possible thanks to our ability to represent others as having mental states that cause their behavior, a *Theory of Mind* (Gopnik et al., 1997). Making sense of others' behavior in terms of the happenings of their minds enables us to determine what they are trying to accomplish (Baker et al., 2009), what they intend to communicate (even when they're ambiguous; Jara-Ettinger & Rubio-Fernandez 2021), what they're likely to do next (Jara-Ettinger et al., 2020), and whether their intentions are praiseworthy or condemnable (Young et al., 2007).

While there is little doubt that representations of other people's mental states structure human social reasoning, questions remain about how we determine which mental states to attribute based on how people act. In the last two decades, one prominent proposal has emerged, which posits that humans understand each other by assuming that agents act to maximize utilities—the difference between the costs that they incur and the rewards they obtain (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016). Under this view, mental-state attribution is a process of identifying combinations of beliefs and desires under which the agent's behavior would maximize utilities. Consistent with this, even young children and infants rely on an expectation of utility maximization to understand other people's goals (Gergely et al., 1995; Lucas et al., 2014), infer what they know (Jara-Ettinger et al., 2017), make sense of future action (Liu et al., 2017), and make sociomoral evaluations (Kiley Hamlin et al., 2013). Moreover, formal computational implementations of this idea, known under the umbrella term of *Bayesian Theory of Mind*, reach human-level performance in simple tasks of preference and mental-state attribution (Baker et al., 2017; Jern et al., 2017).

Despite the broad success of this approach, questions remain about how such inferences might support social reasoning in more complex real-world situations, where both the space of mental states and the space of possible goals can be too large for current mental-state inference algorithms to be tractable. In response to this puzzle, some researchers have argued that BToM models provide only a computational-level description that does not necessarily capture the true algorithmic implementation in the mind (Jara-Ettinger et al., 2016), or that simpler satisficing models might be enough (Pöppel & Kopp, 2018).

Here we propose a different potential solution to this approach. To illustrate our idea, imagine watching someone throw a paper cup into the trashcan at a coffee shop. From this simple action, you might infer that the person was previously sitting at a table drinking their coffee and is now ready to leave, or that they work there and they are cleaning up trash others left behind. From a Bayesian Theory of Mind perspective, such inferences might be supported by inferring that this action is utility maximizing for agents that do not want the coffee cup to remain in the coffee shop and believe that placing it into the trashcan will help get rid of it. With these inferences in hand, we might infer that the agent's desire applies to other coffee cups in the shop (and is therefore cleaning up) or no longer has any desires that would lead them to stay in the coffee shop (and is therefore leaving). Alternatively, however, we propose that we might have learned throughout our life that people generally throw cups out when they are performing a cleaning task, which is a sub-task carried out by employees in coffee shops (which can be an intermediate step in managing a coffee shop) and by patrons of the coffee shop (where it is typically the last step they perform at coffee shops).

Under this second approach, our inferences hinged on knowing how different common events unfold, and recognizing that the observed action is a step in a broader action plan that people commonly take. Note that this second ap-

proach does *not* posit that Bayesian Theory of Mind is unnecessary. Quite the opposite. These high-level inferences about what step an agent might be completing depend on a capacity to process others' behavior and identify their immediate goals (in our example, throwing a paper cup to the trash can). Nonetheless, this proposal helps constrain the types of inferences that we may need to perform under a Bayesian analysis of utility maximization. After identifying an immediate goal through standard models, we may be able to then reason about behavior at a higher level of abstraction, in terms of our understanding of how common physical and social events unfold.

More specifically, we propose that a neglected component of Theory of Mind lies in how we rely on knowledge about the structure of how common events unfold, such as grocery shopping, having a conversation, or cooking a meal. That is, people may process body movements to identify immediate goals, but then rely on knowledge of events to make sense of the agent's broader plan and predict what they might do next. In this paper we present a simple computational interpretation of this idea with two goals in mind. First, to show how a relatively computationally simple system can support rich inferences about others when reasoning about abstract action plans. Second, to empirically evaluate people's capacity to quickly integrate knowledge about potentially novel action plans to make sense of behavior. To achieve this, we focused on a simple block-building paradigm, where participants watched simple videos of an agent completing a single goal and were then asked to infer the agent's broader goals and future actions, given some knowledge of task structure.

Computational framework

At a high level, we model scenarios in which an observer watches an agent performing a sequence of actions, and must reason about that agent's goals, sub-goals, or future actions. However, we are particularly interested in cases where the observable action data are very sparse, and the observer must leverage information about potential action plans to put the observations in context and make coherent inferences. To this end, we model scenarios in which an agent is following a set of instructions to build a structure out of blocks (Figure 1). The observer's background knowledge consists of a) the set of blocks available in the kit (capturing the physical context, which specifies the space of actions that agents can take), b) a set of "target" structures that can be built using the blocks (capturing the space of high-level plans that we expect people to typically pursue), and c) a set of step-by-step instructions for building each target structure (capturing our knowledge of how these high-level plans unfold). Each target structure is modular, consisting of several components that can be built in any order before the final structure is put together.

The observer watches the agent perform a single action out of context (picking up a single block and placing it atop another block, with no context indicating what other blocks have already been used or what components have already been built; Figure 1c). Depending on the task, the observer may also be told which of the target structures the agent is building. The observer must then infer either which component of the structure the agent is working on (Study 1a), which type of block the agent will need next (Study 1b), or which of the target structures the agent is trying to build (Study 1c).

Model

We model the observer's responses as ideal Bayesian inference. In order to infer which component C the agent is building, given the target structure S and observed action a (Task 1a), we compute the posterior probability according to Bayes' rule:

$$P(C|a,S) \propto P(a|C,S)P(C|S) \tag{1}$$

We compute the likelihood $P(a|C,S)$ as the number of times action a occurs in the instructions for component C, divided by the total number of actions in the instructions for component C. Participants are told that the agent is equally likely to be at any point in the instructions, so the prior likelihood $P(C|S)$ is equal to the total number of actions in the instructions for component C divided by the total number of actions in the instructions for structure S (across all components). We use equation (1) to generate predictions for Study 1a.

To predict which block the agent will need next, we first apply equation (1) to compute the probability that the agent is working on each component of the model. Then, for each component, we identify all possible next actions a', given the current action a, and compute, for each type of block b, the fraction of all possible next actions which require a block of type b, weighted by the posterior probability of that next action. This yields the probability $P(b|S,a)$ that the agent will require block b next, given the agent's current action and target structure, which we use to generate predictions for Study 1b.

To infer which structure the agent is building, we again compute the posterior probability according to Bayes' rule:

$$P(S|a) \propto P(a|S)P(S) \tag{2}$$

We assume a uniform prior for $P(S)$ (i.e. that the agent is equally likely to be working on any structure). To compute the likelihood of action a given the target structure S, we marginalize over possible components, i.e. $P(a|S) = \sum_{C \in S} P(a|C,S)P(C|S)$, where $P(a|C,S)$ and $P(C|S)$ are computed as in equation (1). We use equation (2) to generate predictions for Study 1c.

Experiments

We conducted three studies, corresponding to the three inference types described above. All studies and analyses were pre-registered unless explicitly noted (`osf.io/9ubqn`).

Participants

For each study, we recruited 40 adult participants with US-based IP addresses via Amazon Mechanical Turk. As per

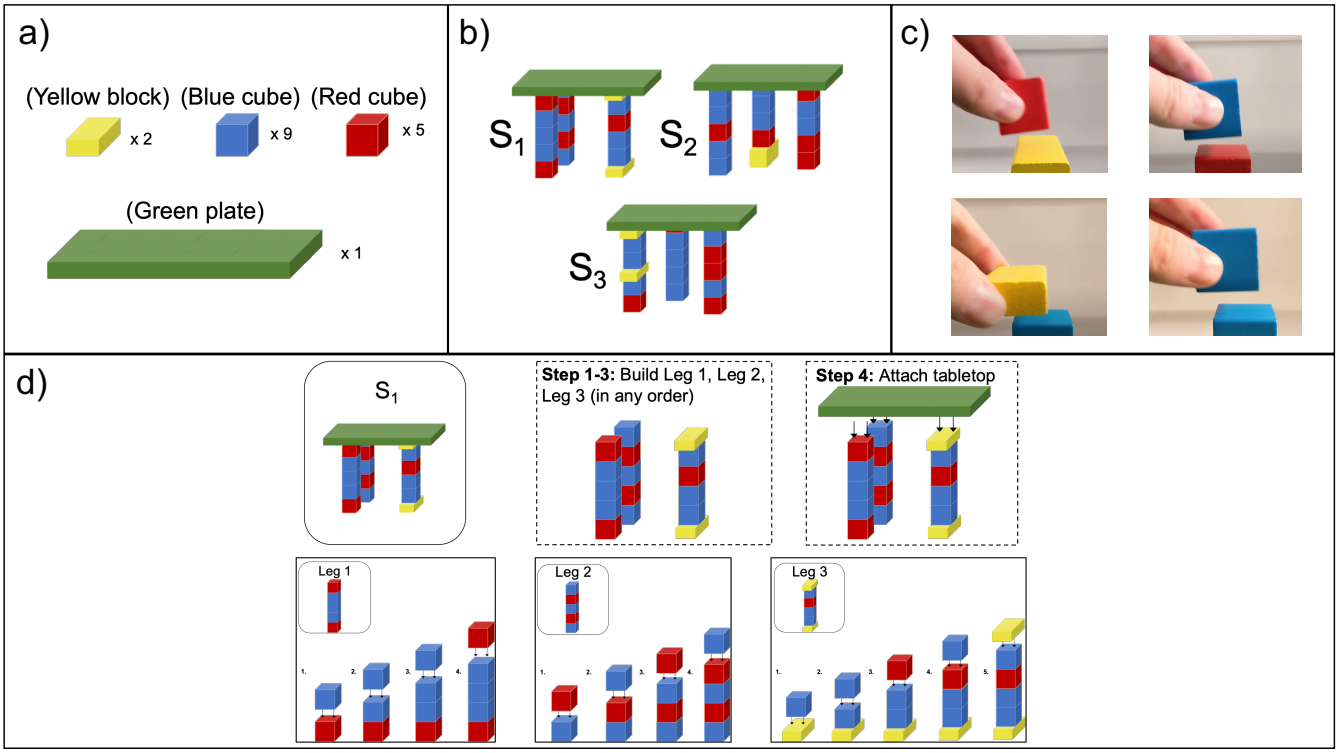


Figure 1: a) The set of blocks shown to participants. b) The three target structures participants are shown that can be all built using the exact same set of blocks. c) Still frames from 4 of the short stimuli videos used in all studies. d) Example of instructions shown for one of the target structures.

our preregistration, participants who failed one or more comprehension questions were excluded from the study, leaving N=38 participants for Study 1a (mean age=38.9, SD=11.3; n=2 exclusions), N=37 for Study 1b (mean age=37.7, SD=11.1; n=3 exclusions), and N=37 for Study 1c (mean age=35.1, SD=7.4; n=3 exclusions).

Stimuli

Participants were first shown a picture of a block-building kit containing a fixed number of blocks of four different shapes and colors (Fig. 1a), pictures of three different structures that can be built from the kit (Fig. 1b), and pictographic instructions for building each structure (Fig. 1d). For trial stimuli, we recorded 8 short video clips depicting an experimenter's hand placing one of the blocks from Fig. 1a on top of another block (see Fig. 1c for examples). Videos were recorded against a blank background, and were framed so as to make the height of the bottom block ambiguous (thereby making it unclear what had already been built prior to the action depicted in the clip, or which blocks had already been used). The 8 clips corresponded to all 8 combinations of blocks that occurred in at least one of the three instruction sets, excluding the "green plate" (as it was always the very last block to be used in each construction).

The three target structures were designed to achieve a mix of frequencies with which the actions depicted in the clips occurred. That is, within each structure, certain actions occurred only on one component, and other actions occurred on all components (possibly with varying frequencies). For example, in Structure 1, only one of the three components involves placing a blue on a yellow, and all three components involve placing a red on a blue, but the second component involves twice as many "red on blue" actions as the first or third component. This mixture meant that participants could rely on deductive inference for some trials (i.e.: those for which the depicted action occurred in only one possible component), but would have to rely on statistical reasoning for others (those for which the depicted action occurred on multiple components with different frequencies).

In Studies 1a and 1b, participants were told in each trial which of the three structures the agent in the clip is building, and then shown one of the 8 clips. Each trial corresponds to a single video/target model pairing, yielding 16 total trials (since some sequences do not occur in all three constructions-for example, Structure 1 does not involve placing a red block on a red block at any step, so there is no "red-on-red/Structure 1" trial). In Study 1c, participants are only shown the clip, and must infer which of the three models the agent is building. Study 1c therefore contains only 8 trials, one for each video clip.

Procedure

Instructions were largely identical for all three studies: participants were first shown the blocks, target structures, and instruction sets, and then told that they will be watching short video clips of someone building one of the structures. After initial instructions, participants were given two chances to pass a 4-question comprehension check, to ensure that they both understood the instructions and could visually distinguish and identify the blocks depicted in the videos. Participants who failed one or more questions on both tries were excluded from the study.

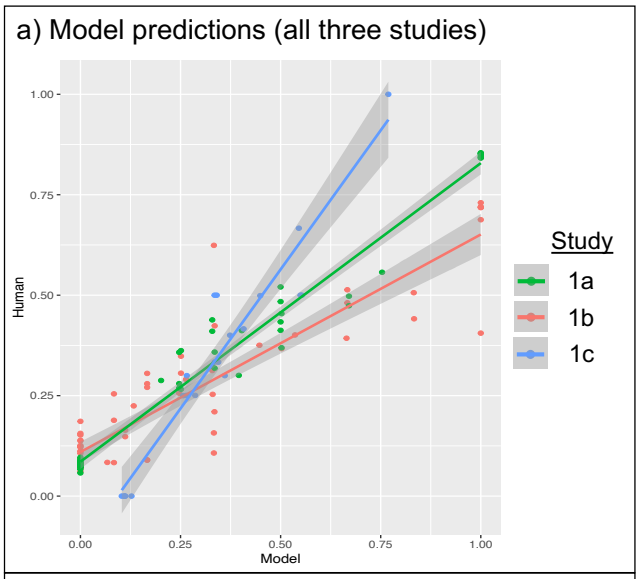
Upon passing the comprehension check, participants were then shown all 16 trials (for Studies 1a and 1b) or 8 trials (for Study 1c) in a random order. In each trial of Study 1a, participants were asked to guess which component of the target structure (Leg 1, Leg 2, or Leg 3) the agent in the clip was working on. Participants could view the instructions for each component at any time by clicking a link to open the image in a new window. Participants responded using three numerical sliders, one for each component, ranging from 0 (definitely NOT building) to 100 (definitely building).

In Study 1b, participants were shown the same stimuli as in 1a, but were instead asked to guess which of the four types of blocks the agent would most likely need next. Participants responded using four numerical sliders, one for each type of block, ranging from 0 (definitely WON'T need) to 100 (definitely WILL need). In Study 1c, participants were shown only the video clips, and then asked to guess which of the three target structures the agent in the clip was building. Participants could view the instructions for each structure at any time by clicking a link to open a full-page image in a new window. Participants responded using three numerical sliders ranging from 0 (definitely NOT building) to 100 (definitely building).

Results

For each trial in each study, participants returned three (for Studies 1a & 1c) or four (for Study 1b) numerical values ranging from 0-100, indicating the participant's estimated likelihood of each possibility given the stimulus information. As preregistered, for each participant and each trial, we normalized the participant's response to sum to 1, then averaged responses across participants within each trial. This yielded, for each trial, the average estimated probability of each possible response for that trial. We then compared these estimated probabilities against the probabilities generated by our computational model.

Figure 2 presents a scatter plot comparison between model predictions and human data (2a) and a summary of correlations and confidence intervals for each study (2b). Across all three studies, model predictions were very highly and significantly correlated with participant responses (Study 1a: $r = .98$, $95\%CI(.97, .99)$; Study 1b: $r = .91$, $95\%CI(.86, .94)$; Study 1c: $r = .96$, $95\%CI(.89, .98)$). Although our preregistration did not include a plan to present an overall model



b) Summary of results

	Study	Inference Task	r	95% CI
	1a	Sub-Step	.98	(.97, .99)
	1b	Next action	.91	(.86, .94)
	1c	Goal Structure	.96	(.89, .98)

Figure 2: Panel a) depicts a comparison of model predictions against participant data, color-coded by study. Each point represents one option probability in one trial, with model predictions on the x axis and aggregate human judgments on the y axis. Shaded bands indicate 95% confidence intervals around a linear regression. Panel b) depicts Pearson correlations between model predictions and human data for each study, along with 95% confidence intervals

evaluation across all tasks, we present one here for completeness. To achieve this, we z-scored participant judgments and model predictions (to standardize all task judgments into a common scale) and computed the overall correlation between model predictions and participants judgments using the data from all tasks. This analysis revealed a correlation of $r = .94$, $95\%CI(.92, .96)$. Importantly, participant responses closely tracked model predictions in both "obvious" trials, where the observed action was consistent with only one possible answer, and in "mixed" trials, where the observed action was consistent with multiple answers (possibly with different probabilities). To better visualize this, Figure 3 shows an example of a "mixed" trial from each study, where the observed action is consistent with all possible answers, but with different probabilities. The rightmost column depicts a comparison between model predictions and average participant responses. In all three examples, participants correctly identified both the more probable answers and the less probable answers at rates

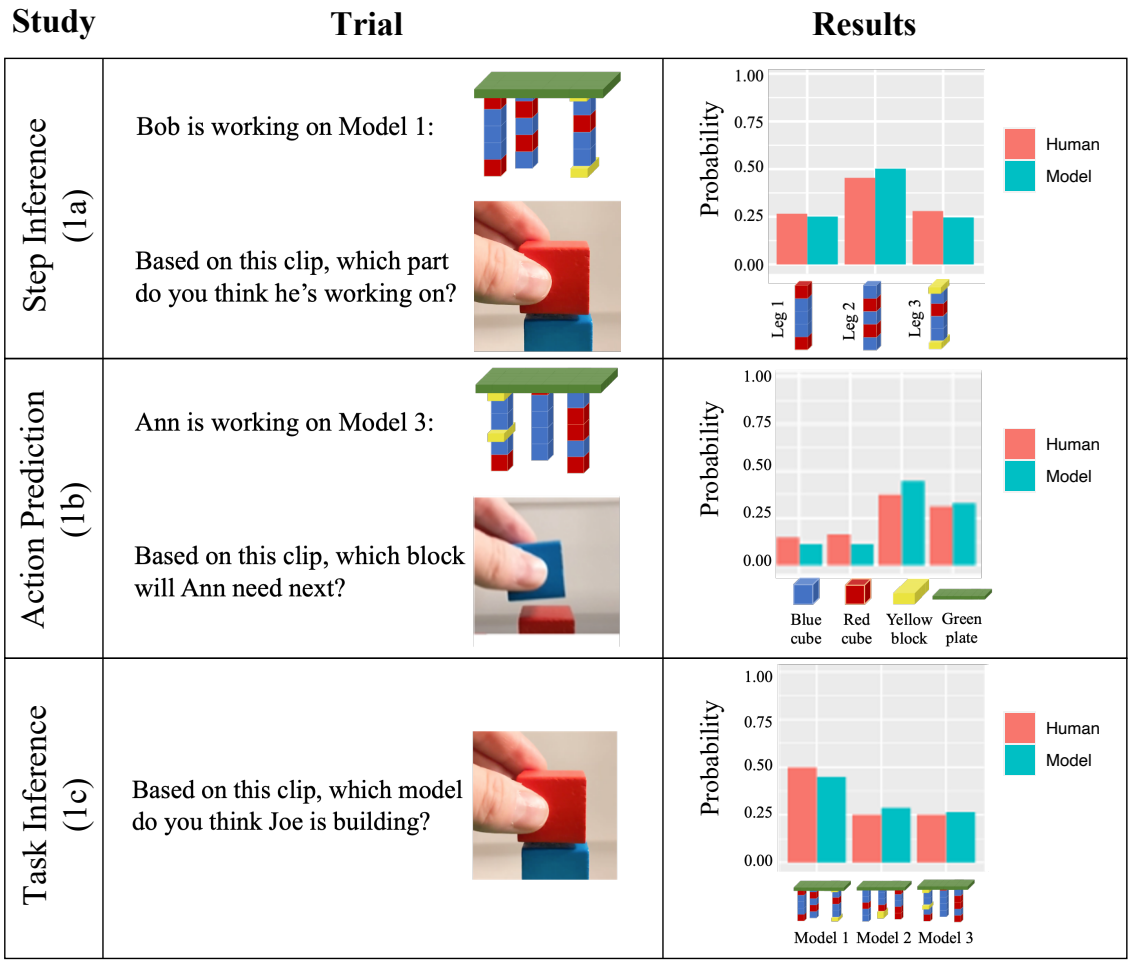


Figure 3: Example of trial stimuli and results. Each row corresponds to one study. Middle column depicts an example stimulus from each study. Third column depicts comparison of model predictions against average participant responses for each study

that closely tracked model predictions. This strongly suggests that participants are utilizing the same structural and statistical information leveraged by our computational model.

The strongest discrepancy between human data and model predictions occurred in Study 1b, where participants predicted which block the agent would need next. A closer examination of the data reveals a likely cause for this discrepancy: in certain 1b trials, the observed action is consistent with both an intermediate step of a component *and* a final step of a component. For example, if the agent is currently working on Leg 2 of Structure 1, the observation "place blue block on red" is consistent with both the third step of that component *and* the final step of that component. In such cases, the model considers both possibilities when predicting the next action: if the observation corresponds to the final step in the current component, then the next action will be the first step in one of the two remaining components, or the final step of the construction (attaching the green plate). A closer analysis of the individual responses, however, suggests that only about half of participants considered this possibility in these ambiguous trials, while the other half only considered possible next steps within the same component. This oversight appears to account for the larger discrepancy between human data and model predictions in Study 1b, though the correlations are still high and significant for that study.

Discussion

Theory of Mind (ToM)—the capacity to infer, represent, and reason about the mental states of others (Gopnik et al., 1997)—underlies many uniquely-human cognitive capacities, including language (Jara-Ettinger & Rubio-Fernandez, 2021), social reasoning (Baker et al., 2009; Jara-Ettinger et al., 2020), and moral reasoning (Kiley Hamlin et al., 2013; Young et al., 2007). Understanding how we accomplish mental state inference is therefore crucial for understanding human cognition more generally. One of the most prominent accounts to emerge in the past two decades posits that humans understand each other through a lens of utility maximization—that is, we expect others to choose actions that maximize the difference between the costs that they incur

and the rewards they obtain (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016). Following this approach, a wealth of research has shown that even children and infants rely on expectations of utility maximization when reasoning about and interpreting human behavior (Gergely et al., 1995; Jara-Ettinger et al., 2017; Liu et al., 2017; Lucas et al., 2014). This approach also lends itself to a computational implementation leveraging Bayesian inference (*Bayesian Theory of Mind*), and has yielded computational models that match human performance in a range of simple mental inference tasks (Baker et al., 2017; Jern et al., 2017).

However, in many real-world cases, the space of possible mental states to consider, relative to the sparsity of available data (i.e.: observed behavior) can lead to intractable computations under a Bayesian Theory of Mind approach. In this project, we propose that in many such cases, observers rely on detailed knowledge of the environment, and how common events frequently unfold within that environment. That is, when observed behavior is sparse enough to be consistent with a broad space of possible mental states, we propose that people rely on knowledge of structured action plans to put their observations into context and draw coherent inferences. To this end, we proposed a computational model of the process through which people integrate structured environmental knowledge into their mental inferences, and tested this model in three experiments using a simple block-building paradigm.

Our results provide converging support for our account: across all three experimental tasks, participant responses closely and consistently tracked the predictions of our computational model. Furthermore, the model was able to replicate human performance in both "obvious" tasks, where the observed action was consistent with only one possible goal or sub-goal, and more complex tasks, where the observed behavior was consistent with multiple possibilities but with different probabilities. This suggests that participants were able to leverage structured knowledge of the environment and possible action plans to make coherent inferences about agent goals and actions from very sparse behavioral data.

Our work leaves several open questions. First, our computational model assumed direct access to agent actions. This is in line with standard models of Theory of Mind, where actions and choices are assumed to be directly observable (e.g., Baker et al., 2017; Jara-Ettinger et al., 2020; Jern et al., 2017). In the real world, however, actions are discrete categories that are not observable, and must be inferred based on people's continuous body movements. Interestingly, some recent research suggests that people can infer actions and goals from body movements using Bayesian inference structured around an expectation that agents move efficiently in space (Qian et al., 2021). This further supports the idea that Bayesian inference over a generative model where agents act rationally and efficiently may form the back-bone of action understanding, which is then transformed to richer inferences based on our social and world knowledge. In future work we hope to extend our model to capture a full pipeline from limb movements to hierarchical task inferences that enables us to test our account in a more holistic way.

Relatedly, our model and experiments assumed knowledge about the types of tasks that agents generally pursue, and the steps involved in each task. While this assumption may be reasonable in many cases, it also raises the question of how people learn this word and social knowledge in the first place. This suggests that people may have an additional capacity for learning event structures from observable actions. To illustrate this idea, imagine visiting a foreign country and walking into a restaurant with unfamiliar customs. If you saw another person walk in and head directly towards an interactive screen on the wall, you might initially infer that this must be the first step in ordering food. However, if you next saw another two people walk up to the counter, you might then start to believe that this restaurant has two possible methods for ordering food, or that there are two different steps to be completed, where the order does not matter. Recent work has indeed found that people can quickly infer potential task structures from watching a small set of agents navigate an environment (Velez-Ginorio et al., 2017), suggesting that this capacity may underlie the task learning that later enables us to draw rich and powerful inferences about the social world.

Finally, another limitation of our work is that we focused on a novel task setting that is not representative of the common events where people engage in action understanding. This enabled us to have a paradigm where we had precise control over what information was available to participants and to our model. In future work, however, we hope to extend our model to more realistic events that are representative of social inferences in the real world.

To conclude, we investigated how humans can infer goals and predict actions from very sparse behavioral data by leveraging structural knowledge about the environment and frequent action sequences. We proposed a simple computational account of this inference and tested it in a simple block-building paradigm. The results demonstrate that people leverage knowledge about potential action sequences in a process very similar to how the model utilizes the same information. This constitutes an important first step towards a broader understanding of human mental inference in complex real-world environments.

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), 0064.

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

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories.* Mit Press Cambridge, MA.

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science advances*, *7*(47), eabj0970.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Pöppel, J., & Kopp, S. (2018). Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 470–478).

Qian, Y., Kryven, M., Gao, T., Joo, H., & Tenenbaum, J. (2021). Modeling human intention inference in continuous 3d domains by inverse planning and body kinematics. *arXiv preprint arXiv:2112.00903*.

Velez-Ginorio, J., Siegel, M., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires.(2017).

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240.