Integrating Experience into Bayesian Theory of Mind

Marlene D. Berke (marlene.berke@yale.edu), Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)
Department of Psychology, Yale University

Abstract

Other people's mental states—what they want, what they know, and how they combine the two to act—are structured by the experiences that they've had. In line with this, we propose that inferences about other people's experiences are a central, but often neglected, aspect of human Theory of Mind. We explore this idea by presenting and testing a computational model that jointly infers others' desires, knowledge, and experience. We find that, by focusing inferences on others' experience, our model can make richer inferences about other's knowledge than would be otherwise possible. Our model quantitatively fits participant judgments on two experiments above an and beyond an alternative model. Overall, our work extends the richness of human Theory of Mind judgements that can be formalized as Bayesian inference over a generative model.

Keywords: Theory of Mind; Computational modeling; Social cognition

Introduction

Imagine visiting a supermarket with your friend. Since the start of the pandemic, the supermarket has re-arranged their aisles. As you enter the store, your friend says "I'll get the vegetables" and heads off in the wrong direction. From this simple behavior you can instantly infer that your friend probably hasn't been to this supermarket since the pandemic began. This inference, although simple, can then help you build a richer model of your friend's mental representations: she probably doesn't know that the vegetables and meat sections have switched places, or that there's a new fish counter where the flowers used to be, but at least the bakery is in the same place, so they'll have no trouble finding the bread.

This capacity to build mental models of other people's minds is known as a *Theory of Mind* (ToM; Gopnik et al., 1997; Wellman, 2014). Over the last decade, behavioral, computational, and developmental research has found that people attribute mental states like beliefs and desires through an expectation that other people behave rationally (see Jara-Ettinger et al., 2016, for review). In the example above, for instance, we were able to infer that your friend had an inaccurate representation of where the produce was located, because her behaviour—heading in the wrong direction—would otherwise be irrational.

But this previous work has missed something. In our example above, we not only inferred that our friend did not know where the produce was, we also inferred that our friend lacked an *experience*: visiting the supermarket since the pandemic. And this experience inference enabled us to deduce not just where they believed the vegetables were, but also the meat, flowers, and bread. As this example illustrates, experiences structure the way that we expect agents to acquire beliefs, enabling us expand a belief that is diagnostic of an experience (e.g., your friend thinks the vegetables are over there), into a richer representation of their epistemic states (like the location of the vegetables, meat, flowers, and bread).

Past research on Theory of Mind has typically equated experience with perceptual access, treating it as an observable factor that does not require inference (i.e., perceptual access implies seeing, and seeing implies knowing; e.g., Baker et al. 2017; Onishi & Baillargeon 2005; Lin et al. 2010; Wimmer & Perner 1983. Even when we learn from a teacher, it's assumed that perceptual access to the lesson is a given; e.g., Shafto et al. 2014.). In more complex cases, however, such assumption faces two challenges. First, we are rarely privy to the vast majority of experiences that other people have had in their life. Therefore, assuming that an agent has not experienced something simply because it is not actively in their visual field would be an error. Second, even when an agent has direct perceptual access to an object or an event, this does not imply that the agent is experiencing it: people can drift off, mind wander, or simply have too much information in their visual field to process.

Based on this analysis, here we propose that representations of other people's experiences are a central component of human Theory of Mind, forming a cornerstone that helps us understand and predict other people's behavior. This view implies that (1) people should be proficient at inferring other people's potential experiences based on their behavior, and (2) people can then use these inferences to build more nuanced representations of other people's minds. In this paper we present a first approximation of this idea. We introduce a simple computational model that aims to capture how representations of experience might be integrated into computational frameworks of Theory of Mind. We also present two behavioral experiments that aim to evaluate our model and seek some initial evidence on people's capacity to infer others' experiences and use these inferences to make sense of an agent's behavior.

Computational Model

Previous research suggests that human judgements about an agent's mental states can be modeled as Bayesian infer-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

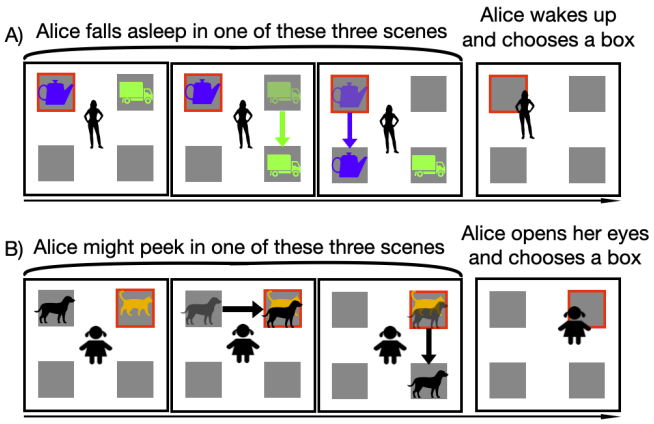

Figure 1: Timelines of example stimuli from Experiments 1 (A) and 2 (B). In A), the green truck moves downward and then the blue teapot moves downward, one at a time. Alice falls asleep in one of the three states in this sequence. Alice then wakes up and chooses a box (outlined in red) to look for the object she wants. In B), the black dog moves rightward and then downward. Alice might peek during one of the three states in this sequence. Alice then opens her eyes and chooses a box to look for the goal she wants.

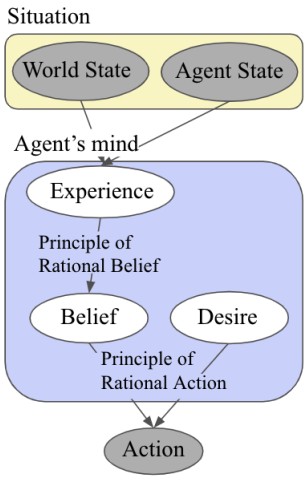

Figure 2: Conceptual model schematic. The yellow box represents the physical situation (the world state and the agent's state) and the blue box represents the agent's mind. The shaded areas denote observability.

ence over a generative model describing how mental states rationally produce actions (Jara-Ettinger et al., 2020; Jara-Ettinger, 2019; Jern et al., 2017; Lucas et al., 2014; Baker et al., 2017). We take this framework as a starting point, and we extend it to include explicit representations about agents' experiences. The key difference from previous models is that our generative model is designed to capture the acquisition of beliefs structured around experiences. This constrains inferences about an agent's beliefs to those that are compatible with one another and compatible with a possible experience that the agent might have had.

For clarity, we present our model in terms of our experiments, but the model can be applied to arbitrary world states and actions. Consider an event like the one shown in Figure 1 A. Here, Alice is in a room with four boxes, two of which contain objects. Alice knows the initial location of the objects, and she updates her beliefs as the objects move from one box to another. However, Alice falls asleep at one point in the event and no longer sees the objects moving. When Alice wakes up she moves towards one of the boxes to collect an object, and the goal is to infer (1) when Alice fall asleep, (2) which of the two objects she was looking for, and (3) her beliefs about the location of each object.

For instance, when Alice approaches the top left box in Figure 1A, her action reveals that she was probably looking for the blue teapot and that she fell asleep before the teapot moved to the bottom left location. However, this action does not reveal whether Alice fell asleep at the start, or after the green truck had moved, so we would be unable to infer exactly when Alice fell asleep or where she thinks the truck might be. These are the kinds of inferences that our model aims to capture.

Generative model

Figure 2 shows a conceptual schematic of our model. In line with previous work, we define a situation as the combination of the world state and the agent's state. The world state consists of all physical information about the scene, including the objects and their locations. The agent state is the physical location of the agent, which determines that all changes take place within the agent's field of vision. The situation is fully known to the observer. In these experiments, the world state changes as objects move, and the agent state changes when the agent moves toward a box.

Together, the world state and the agent state determine the space of information the agent might possibly experience. But the experience variable captures whether the agent receives and processes that information (e.g. whether the agent is awake, attentive, and has their eyes open). In Experiment 1, the agent start off as experiencing (i.e. awake) and switches to not-experiencing (i.e. asleep) during one of the world states. In Experiment 2, the agent is assumed to start off as not experiencing (i.e. with their eyes closed) and then briefly switches to experiencing (i.e. peek) during one of the world states. When the agent experiences a world state, they update their beliefs about the locations of the objects to match the current world state. When the agent does not experience a world state (e.g. when they are asleep or have their eyes closed), the agent does not update their beliefs about the locations of the objects and instead retains their beliefs based on the last world state they experienced. In the context of our task, a desire consists of an object that the agent seeks to obtain. None of these mental states or processes (i.e. desires, beliefs, experiences, and belief updating) are observable.

In this simple setup, the principle of rational action is implemented as an assumption that the agent will choose the box where she believes her desired object is located. This action is observable.

Inference

In both experiments, the task is to infer the agent's desire D, beliefs about the state of the world B (which consist of the location of each object), and the agent's experience E from the agent's observable action A and the observable sequence of world states W. This can be decomposed into interpretable terms as follows:

$$p(D,E,B|A,W) \propto p(A|D,B)p(B|E;W)p(E)p(D) \qquad (1)$$

where $p(A|D,B)$ is the probability of the action given a desire and beliefs about the locations of the objects, which is described by the principle of rational action. $p(B|E;W)$ is the updating of beliefs according to experiencing a world state, which is described by the principle of rational beliefs. $p(E)$ is the prior over experience (i.e. in Experiment 1, it is prior over when Alice falls asleep, and in Experiment 2, it is the prior over if and when Alice peeks), and $p(D)$ is the prior over desires (i.e. which goal Alice wants). Each of these terms are described in the generative model.

Conditioning on the sequence of world states and the action, we use the generative model to infer the agent's desire, experience, and belief about the location of each of the objects. We implemented inference via Markov Chain Monte Carlo (MCMC) using Metropolis-Hastings.

Behavioral Experiments

In Experiment 1 we present a first test designed to evaluate our computational model and people's ability to infer other people's experiences and use these inferences to build richer mental-state representations. Specifically, here we ask people to infer 1) which goal an agent desires 2) what the agent experienced and 3) where the agent believes each object is. In Experiment 2, we perform a conceptual replication of Experiment 1 in a slightly modified paradigm, changing the priors over experience (i.e. $p(E)$ in Eq. 1), so as to enable us to establish further evidence that these inferences are supported by a nuanced mental model of how experience affects behavior. Pre-registration, stimuli, instructions, model predictions, and data for both experiments available at: (https://osf.io/34nvx/?view$_only$ = 4*ca*64*e*6*a*902*d*4*b*29*bf*5*ad*89*c*85753609).

Experiment 1

Participants 120 U.S. participants (Age: mean = 29.9 years, range = 18-83 years; Gender: 51 women, 67 men, 2 non-binary) were recruited from Prolific. An additional 8 participants were recruited but not included in the study because they did not complete the experiment (n = 3) or because they failed to pass the attention check questions (n = 5; see procedure).

Stimuli Stimuli consisted of 18 short videos (see Fig. 1 for schematic). Each video showed an agent (Alice) in the center of a room with four boxes and two visible objects, each on top of a box. The video then showed up to two events, each consisting of one or both of the objects moving from one box to another. After the two events, the objects faded into the boxes and the agent approached one of the boxes.

The full parametric combination of possible starting states, object movements, and agent choices leads to 16384 possible trials. However, the set of possible inference patterns in this paradigm is discrete and much smaller. To reduce this stimuli space, we ran our model on every possible trial, and used its inferences to select a set of trials that captured a range of possible inference values that our model predicts people should be able to make.

Specifically, we first collapsed all inference values that our model produced (integrating beliefs, desires, and experiences), and then selected a combination of trials that included every possible inference value generated by the model. This resulted in 18 trials that stemmed from six different event (i.e., object movement) sequences combined with every possible box that Alice could choose (excluding cases where Alice approached a box that had never had an object). Note that each trial elicits 13 (not-independent) judgments: two goal inferences, three experience inferences, and eight belief inferences, such that this relatively small number of events enabled us to cover a wide range of possible predictions.

Each video was randomly assigned to one of three test conditions, such that there were six videos per condition (tested across participants). For each condition, the pattern of object movements in each trial was randomly rotated (0, 90, 180, or 270 degrees), flipped (no flip or horizontal flip), and assigned different object icons. This enabled us to reduce visual similarity across videos without affecting the inferences predicted by our model. Splitting the 18 trials into three conditions also limited the time it would take a given participant to complete the experiment (completing all 18 trials in one sitting would have been infeasible for an online experiment).

Procedure Participants first read a brief tutorial that explained the logic of the task in the context of a warmup video and they were asked seven simple attention check questions to ensure they understood the logic of the task (full experiment available in OSF repository). Only participants who answered all questions correctly were given access to the experiment. For each comprehension question, the participant had four opportunities to answer correctly. If they failed all four times, they were not allowed to continue into the experiment. They had the option to restart the tutorial from scratch if they chose to.

In the main phase of the task, participants were randomly assigned to one of the three conditions and watched each of the condition's six videos in a randomized order. After watching each video, participants were asked to answer "Which object is Alice looking for?" using a slider, with each end representing one object (e.g. the ends labelled "the blue teapot"

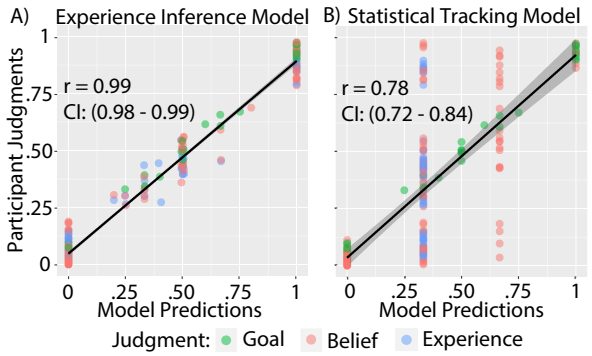

Figure 3: Experiment results 1 for A) The Experience Inference model and B) the statistical tracking alternative model. Each point represents a trial with model prediction on the x axis and participant judgment on the y axis. The black line shows best linear fit between the model and the data with 95% confidence bands (in gray). The color of the points represent the type of judgement – green represent goal attribution, red represents belief attribution (about an object in a box), and blue represents the experience attribution (about when Alice fell asleep).

and "the green truck"). Next, participants were shown three time frames from the video (as in Fig. 1) and they were asked to distribute ten clicks across the three key frames (labelled as "Scene 1", "Scene 2", and "Scene 3") to indicate their belief about when Alice fell asleep (with each click showing a red dot on the selected scene). Finally, participants were asked where Alice believes each object is located. For the first object, participants entered their beliefs by distributing ten clicks across the four boxes. This process was repeated for the second object. In each judgment, each click was treated as representing 10% of the mass of the participant's posterior distribution.

Results Each of the 18 trials produced 13 (non-independent) points: 2 goal inferences (one per object), 8 object location beliefs (4 per object), and 3 inferences about Alice's experience (i.e. when she fell asleep). As Figure 3A shows, our model showed a high quantitative fit to participant judgments ($r = 0.99$; $CI_{95\%}$: $(0.98, 0.99)$). The model fit was similar for each inference type: $r = 0.99$ for beliefs, $r = 0.98$ for experiences, and $r = 0.99$ for goals. Figure 4A shows results from three example trials, illustrating how our model captures graded inference patterns across a range of events.

One alternative possibility is that participants in our task did not use Theory of Mind and instead relied on a simple form of statistical tracking. That is, participants may have tracked the statistical distribution of the position of the objects and used this distribution to infer Alice's goals (where Alice's goal is given by the relative percentage of time that each object spent in the chosen box), beliefs (where the belief about the location of each object matches the percentage of time the object spent in each box), and experience (where Alice could have fallen asleep at any point in the video). This statistical tracking model makes theoretically identical inferences to a Bayesian ToM model lacking the principle that experience structures belief acquisition. In other words, the statistical tracking model is an algorithm that implements a Bayesian ToM model with a uniform prior, likelihood, and posterior over when Alice fell asleep. Crucially, this statistical tracking model made different attributions about experience and beliefs than did our Experience Inference model. But since in our Experience Inference Model goal attribution did not depend on experience, the statistical tracking model and our Experience Inference model made identical goal attributions.

This statistical tracking model had an overall correlation of $r = 0.78$ ($CI_{95\%}$: $(0.72, 0.84)$ (see Fig. 3B) with participant judgements, which was reliable lower than our main model ($\delta = 0.21$; $CI_{95\%}$: $(0.15, 0.26)$). These results suggest that inferences about other people's experiences through Theory of Mind support people's reasoning in our task.

Experiment 2

Participants 80 U.S. participants (Age: mean = 24.1 years, range = 18-39 years; Gender: 32 women, 47 men, 1 non-binary) were recruited from Prolific. An additional 35 participants were recruited but not included in the study because they did not complete the experiment (n = 6) or because they failed to pass the attention check questions (n = 29; see procedure).

Stimuli Stimuli consisted of 10 short videos, that showed events with the same structure as Experiment 1 (Figure 1B). To select the trials, we ran our model over every possible event (n=16384) and used its inferences to select a set of trials that captured the full range of possible inferences that people should be able to make in this task. We selected a set of trials such that they spanned every possible inference value along each dimension (goal, experience, belief) separately (11, 10, and 15 possible different, but not independent, values for goal, experience, and beliefs inferences). The final ten videos were randomly split into two conditions (n=5 videos per condition) and movements and object icons were randomized in the same way as Experiment 1.

Procedure The procedure was nearly identical to Experiment 1, with the difference that participants were now told that Alice is playing hide-and-seek and has her eyes closed but is likely to peek in half of the games that she plays, so as to set participants' prior over whether or not Alice peeks. Participants were asked eight simple attention check questions to insure that they understood the logic of the task (full experiment available in OSF repository) and the inclusion procedure was identical to Experiment 1.

Participants then completed the same questions from Experiment 1, with the difference that they were asked whether and in which scene Alice peeked (with an option for "Alice

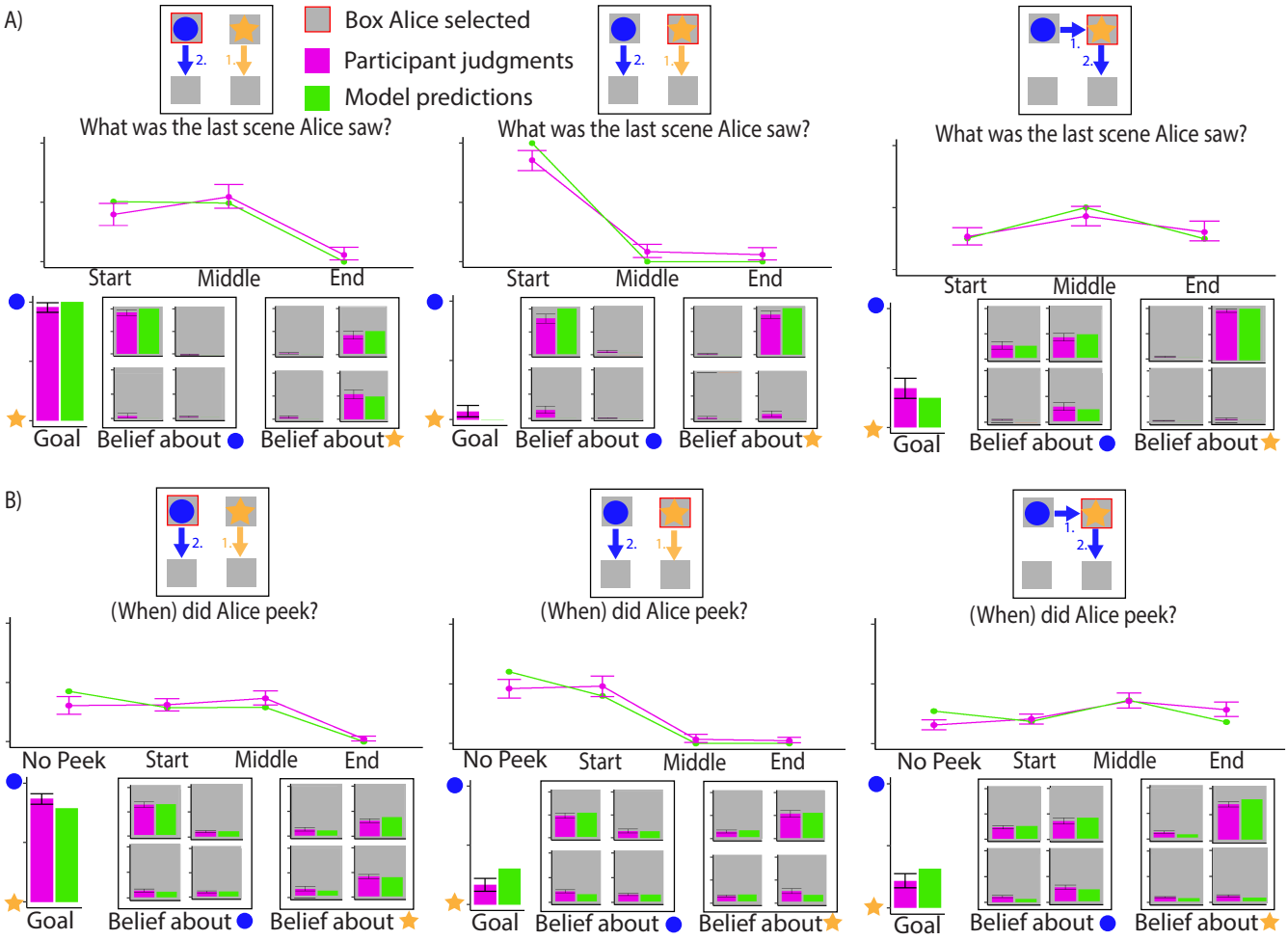

Figure 4: Detailed results for three trials example trials. A) shows results from Experiments 1, and B) shows results on the same trials in Experiment 2. Each panel presents the results from one trial. At the top of each panel is a schematic showing object movements and the box that Alice picked outlined in red. In the first two examples, the object movements are the same as in Figure 1A, but Alice's choice varies. The third example has the same object movements and choice as in Figure 1B. The lineplot graph shows the participant judgements (magenta) and model inferences (green) about Alice's experience: in A) it is which of the three scenes Alice last saw before falling asleep, and in B) it is whether and when Alice peeked. The barplot labeled "Goal" displays participant judgements and model inferences about which object Alice wants. Beliefs about each object's location are displayed using four barplots in the same arrangement of the four boxes. For example, the top left barplot under "Beliefs about the blue circle" gives participant judgements (magenta) and model inferences (green) about Alice's belief that the blue circle is in the top left box. The y-axis for every graph shown is from 0 to 1. All participant judgements have 95% bootstrapped confidence intervals.

didn't peek"), rather than in which scene Alice fell asleep.

Results Each of the 10 trials produced 14 points: 2 goal inferences (one per object), 8 object location beliefs (4 per object), and 4 inferences about Alice's experience (i.e. if and when she peeked). For detailed results for three example trials, see Figure 4B. As Figure 5 illustrates, our model showed a high quantitative fit to participant judgments ($r = 0.95$; $CI_{95\%}$: $(0.92, 0.96)$). The model fit was similar for each inference type: $r = 0.96$ for beliefs, $r = 0.93$ for experiences, and $r = 0.97$ for goals. As in Experiment 1, an alternative possibility is that participants did not use a Theory of Mind including experience and instead relied on a simple form of statistical tracking. This alternative model had a correlation of $r = 0.80$ ($CI_{95\%}$: $(0.70, 0.87)$; Fig. 5B), which was reliable lower compared to our main model ($\delta = 0.15$; $CI_{95\%}$: $(0.08, 0.23)$). These results suggest that the inferences about other people's experiences through Theory of Mind support people's reasoning in this task.

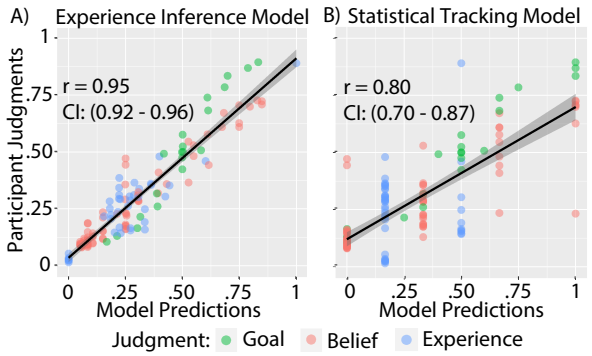

Figure 5: Experiment 2 results for A) The Experience Inference model and B) the Statistical Tracking alternative model. Each point represents a trial with model prediction on the x axis and participant judgment on the y axis. The black line shows best linear fit between the model and the data with 95% confidence bands (in gray). The color of the points represent the type of judgement – green represent goal attribution, red represents belief attribution (about an object in a box), and blue represents the experience attribution (about if and when Alice peeked).

Discussion

Here we proposed that representations of other people's experiences are a central component of human Theory of Mind. Specifically, we proposed that, by understanding how experience structures belief acquisition, people can build richer representations of others' minds. To explore this idea we presented a computational model that integrates experience into Theory of Mind. In two separate experiments we found that people, like our model, were able to jointly infer another agent's experiences and beliefs about the world. Our model captured human inferences about beliefs, desires, and experiences with quantitative accuracy. Critically, our two experiments were nearly identical, but varied subtly in the structure of experience (the agent falling asleep in Experiment 1, and sometimes peeking in Experiment 2). Participant inferences reflected a sensitivity to these differences (resulting in a good fit with the model in both cases), which suggests that people might have an ability to reason flexibly about different types of experience.

An alternative model lacking the principle that experience structure beliefs failed to capture the richness of participant judgments. Interestingly, however, this model was nonetheless able to accurately infer the agent's desires. This finding points to one potential mechanism that people could use to infer mental states: statistical tracking may sometimes enable goal inference; the way the agent pursues the goal might then reveal their experience; and the inferred experience might enable us to draw richer inferences about their mind. This is a question that we hope to explore in future work.

Our work opens at least three directions for future work. First, in this work we considered an overly simplified representation of experience—a binary variable about whether or not the agent could receive information. In real life, these representations are likely graded, capturing degrees of processing that are critical for reasoning about states such as distraction or light sleep. Even more importantly, in real life, we don't know always the relevant past history of what someone may have experienced. That hypothesis space of what someone may have experienced in their lifetime can be large, unconstrained, and individualized. Nonetheless, the types of computations that we specified here might be critical to understanding others' behavior in constrained contexts. Intuitively, some behaviors are also tightly linked with certain experiences, and these linkings might make the problem more tractable (e.g., hearing someone speak fluent French would immediately give us a guess about where they grew up and what kinds of other experiences they may have had). How these inferences might support Theory of Mind in the wild is an open question we hope to explore in future work.

A second and related limitation of our work is that our model focused only on the role of experience when reasoning about others' goals and knowledge. The inferences that we make in real life about the causes and consequences of experience are richer: we can also infer the causes behind a person's experience or lack of experience. For example, if your friend is an American History major but can't remember anything about the presidency of FDR, you might infer that they skipped or slept though that lecture, and therefore might further infer that they find Great Depression Era history boring. Related work has found that people are proficient at inferring the causes behind people's goals (Jara-Ettinger et al., 2016), opening the possibility that they might be able to do the same for the causes behind experiences.

Finally, a third direction for future work lies in the integration of metacognition. People lacking an experience are often aware of it and may adapt their behavior to account for it. For example, in Experiment 1, Alice could wake up knowing that she fell asleep, and it might be more natural for us to expect Alice to be uncertain about where the objects are since she might suspect that the objects moved while she was asleep. Similarly, in Experiment 2, we might expect Alice to use her metacognition and choose to peek near the end of the video so as to reduce her uncertainty stemming from the chance that the animals might move while her eyes are closed.

Overall, our work is a first step in positing that experience is a central component of human Theory of Mind. Beyond being able to read the mental states of another agent, we are also able to make inferences about that agent's previous experiences and how they shape their actions. This work advances our understanding of the computations behind the human ability to make rich inferences about another agent's mind and history.

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.

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

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

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

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

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

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.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, *308*(5719), 255–258.

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

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.