Mental Inference: Mind perception as Bayesian model selection

Lukas Burger

Advised by Julian Jara-Ettinger,
Assistant Professor of Psychology

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for the degree of Bachelor of Science

Yale University
April 24th, 2020

Abstract

Beyond an ability to represent other people's mental states, people can also represent different types of minds, like those of babies, pets, and even wildlife that we rarely interact with. While past research has shown that people have a nuanced understanding of how minds vary, little is known about how we infer what kind of mind different agents have. Here we present a computational model of mind attribution as Bayesian inference over a space of generative models. We tested our model in a simple experiment where participants watched short videos in the style of Heider & Simmel, 1944, and had to infer the representations in the agent's mind. We find that, from just a few seconds, people can make accurate inferences about agents' mental capacities, suggesting that people can quickly infer an agent's type of mind based on how they interact with the world and with others.

Contents

1. Introduction

People's ability to understand each other's behavior rests on an assumption that agents are, broadly speaking, rational (Dennet, 1989). If you learned that Charlie wants to grab his favorite toy, and that he believes that someone put it in a drawer, you can predict that he'll walk towards the drawer, open it, and take his toy. Conversely, if you watched Charlie walk straight towards a drawer, open it, and retrieve his toy, you would immediately recognize that he wanted his toy and knew where to find it (why else would he have acted in this way?). This capacity to transform people's actions into judgments about their mental states, called a *Theory of Mind* (Gopnik et al., 1997; Wellman, 2014), is the basis of human social intelligence, allowing us to explain other people's behavior (Malle, 2006), share what we know (Bridgers et al., 2016), distinguish those who are nice from those who are mean (Jara-Ettinger et al., 2015; Hamlin et al., 2013), and communicate with each other (Sedivy, 2003; Grice et al., 1975).

1.1 Other Minds

Consider, however, what would happen if you found out that Charlie is not actually a person, but a golden retriever. Intuitively, Charlie could still want his favorite toy and know where to find it. Yet, we would not always expect him to be able to get it—most obviously, because Charlie's physical constraints are different from our own, making it difficult for him to open drawers and retrieve objects. But, we might also expect Charlie to fail for a deeper reason: his inability to devise complex action plans from his beliefs and physical constraints which can fulfill his desires.

Classical research in cognitive science has found that people perceive a wide range of types of minds, roughly organized around two dimensions: agency and experience (H. M. Gray et al., 2007). Intuitively, agency corresponds to an agent's cognitive activity—the complexity of their representations and the sophistication of the computations that they perform. Experience corresponds to an agent's subjective ability to sense the world and their own mental states—experiences like seeing and hearing, and emotions like joy, jealousy, anxiety, and pain. For example, a frog may be perceived as high in experience because it has rich sensory representations of its environment but low in agency because it is limited to simple cognitive operations. Robots, on the other hand, may be classified as high in agency because of their complex planning mechanisms, but low in experience because of their limited sensory connection to the world. The degree to which we ascribe agency and experience to a mind captures a wide range of phenomena, from our perception of the 'uncanny valley' (K. Gray & Wegner, 2012) to the type of moral responsibility that we think a creature can receive (K. Gray et al., 2012).

1.2 Perceiving Other Minds

Despite evidence that people distinguish between myriad types of minds, several major questions remain. First, how do people acquire this 'mental space'? Does it emerge through a slow process requiring years of experience? Or can people infer a type of mind as quickly as they infer mental states? Second, how do inferences about minds relate to inferences about mental states? Are the computations behind mind inference similar to the ones at work when we infer beliefs and

desires? Or do they follow radically different inferential principles? And third, how can we formalize agency and experience in precise computational terms?

In this paper, we provide a first step towards answering these questions. Our goal is to develop a computational model of mind perception that clarifies how we infer what type of mind we are observing, and how these inferences relate to the computations we undergo when reasoning about mental states. By comparing the predictions of our model to those of human participants we can quantify how well our model represents human perception of other minds. Our approach builds on previous work that models mental-state attribution as Bayesian inference over a generative model of rational action, and extends it to the perception of other minds represented as a collection of generative models.

While much work has attempted to formalize in precise computational terms how people infer *beliefs and desires* from observable action (Jern et al., 2017; Lucas et al., 2014; Jara-Ettinger et al., under review; Baker et al., 2017; see Jara-Ettinger 2019 for review), to our knowledge, no similar effort exists for the problem of inferring types of minds. Inspired by classical work that showed how simple two-dimensional displays can elicit rich mental-state inferences (Heider & Simmel, 1944), we compare our model to humans in a simple task where participants have to infer the mental structure of a "guard" attempting to capture a "thief", in simple grid-world environment, using continuous confidence measures that allow us to obtain graded quantitative inferences about the mental structure of the "guard".

2. Computational Framework

At a high level, our computational model can be thought of as searching across a space of possible minds, to find one which, under the right beliefs and desires, explains the agent's observed behavior. We thus begin by briefly reviewing models of mental-state inference, and then turn to how our framework expands on this approach.

2.2 Mental-State Inference

When inferring mental states, research suggests that we do so by assuming that agents act rationally to fulfill their desires given an agent's beliefs (Dennet, 1989; Gopnik et al., 1997). This idea can be formalized as an expectation that agents act to maximize the subjective rewards that they obtain while minimizing the costs that they may incur (Jara-Ettinger et al., 2016; Lucas et al, 2014; Jern et al., 2017). Through this assumption, mental-state attribution can be achieved by applying Bayesian inference to a generative model that produces action plans which maximize the agent's expected utilities, as determined by the beliefs and desires. Formal implementations of this idea—typically done through Markov Decision Processes, a framework for computing utility-maximizing plans—capture with quantitative accuracy how people infer other people's competence, preferences, beliefs, percepts, and moral standing (Jara-Ettinger et al., under review; Baker et al., 2017, 2009; Ullman et al., 2009; Lucas et al., 2014; Jern et al., 2011, 2017).

Inferences around an expectation that agents maximize utilities, however, depend not only on an assumption of rationality, but also on the structure of the generative model. Returning to the example in our introduction, if Charlie wanted to grab his favorite toy, we would expect that the way he attempts to maximize his utility (namely, by getting his toy while incurring the

lowest necessary cost) will depend on whether Charlie can hold beliefs, whether he can sustain his desire to get the toy over extended periods of time, and whether he can infer the location of the toy from some indirect evidence.

2.2 Mental Inference

Building on previous work, we define a mind M as a generative model that transforms mental states into observable actions (see Figure 1). Given some observed actions a, the posterior probability that an agent has mind M is given by

$$p(M|a) \propto p(a|M)p(M). \qquad (1)$$

Because the relationship between a type of mind and observed behavior is mediated by the mental states, we compute the likelihood function by integrating over the potential mental states that the agent might have, such that

$$p(a|M) = \sum_{s \in S_M} p(a|g,M)p(g|s,M)p(s|M) \qquad (2)$$

where S_M is the space of all mental states that a mind M can have (i.e. the space of all possible inputs to the generative model), $p(s|M)$ is the prior probability that an agent with mind M would have mental states s, $p(g|s,M)$ is the probability that the agent would have goal g under mind M with mental states s, and $p(a|g,M)$ is the likelihood that an agent pursuing goal g would take actions a.

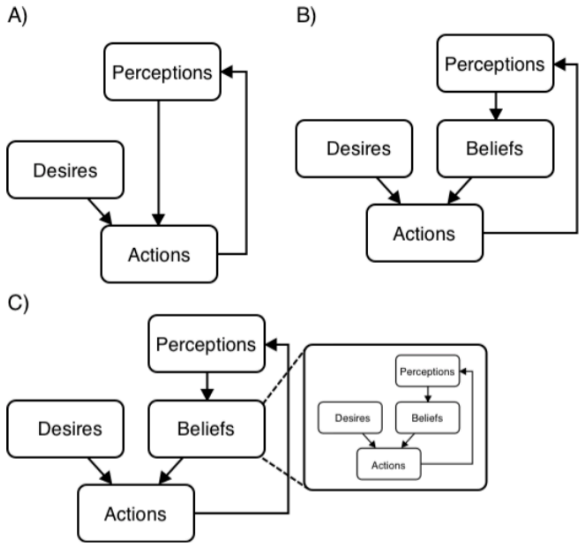

Figure 1: Three example mental models that our approach considers. A) mind with no Beliefs. B) mental model with Beliefs but no Theory of Mind. C) mind with a Theory of Mind, capable of understanding that other agents have minds of their own.

2.3 Space of Minds

Modeling the full space of possible minds is a daunting task and beyond the scope of our project. Our goal instead is to test the plausibility of this approach; as such, we imposed two simplifying assumptions. First, we only considered a small family of types of minds, generated by four discrete criteria (see Fig. 1). Minds were constructed by parameterizing whether (1) the agent had belief representations (Fig. 1a-b; *Belief* component; determining whether the agent's actions were the product of an intermediate mental representation, or the result of a direct mapping from percept to action), (2) whether the agent's belief representations were stable or whether they decayed over time (*Forgetting* component; leading the agent to lose its representations over time; set to probabilistically occur after approximately 1.5 to 2.5 seconds), (3) whether the agent could represent the mental states of other agents (Fig. 1b-c; *Theory of Mind* component; allowing it to

predict other agents' trajectories and intercept them along their path), and (4) whether the agents' perceptual system only consisted of seeing, or if it consisted of seeing and hearing (*Hearing* component).

Our second assumption was that agents' desires are known, making Eq. 2 more tractable. In the context of our experiment (see Procedure), participants had to infer the mind of a guard trying to catch a thief, and thus always knew that the guard's desire was to capture the thief. To fulfill this desire, the guard is given a variety of intermediate goals, which can affect the agent's behavior in different ways. We considered a relatively simple space of goals that the guard could pursue in the service of its desires: 'guarding' (consisting of standing still until seeing the thief), 'chasing' (moving directly towards the last position the thief was seen in), 'intercepting' (moving to a position that would intercept the thief), 'searching' (moving randomly around the map in the hope of locating the thief), and 'patrolling' (repeating a waypoint-marked route over and over until the thief is detected). These goals allowed the guard to display a rich spread of behaviors while pursuing its eventual goal of capturing the thief.

Because the generative models of minds specify the representations in an agent's mind, they also indirectly determine the space of goals that agents can pursue. For instance, an agent with no beliefs can chase an agent, but will stop doing so as soon as the agent is out of sight. In contrast, an agent with beliefs can continue searching for an agent (although they may eventually forget about the thief's existence), or move to where they predict the agent was going (if they have a Theory of Mind). Finally, as the agent navigates, agents who can hear can also update their representations based on auditory evidence.

To summarize, in our framework, a parameter space determines the space of possible minds (instantiated as generative models); the generative model determines the space of goals that the agent can pursue given its (known) desires and internal representations; and, finally, the goals specify how the agent plans to move to different locations (using a probabilistic Markov Decision Process where we softmax the value function to produce a probabilistic policy, in line with past work on action understanding; Baker et al. 2009, 2017; Jara-Ettinger et al. under review). Given this entire forward process we can then compute the posterior distribution over types of minds given some observed actions through Eq. 1, using a uniform prior over the space of minds and the space of goals.

3. Methods

To test our model, we ran a simple experiment where participants watched 2D videos of a thief trying to steal a treasure, which was guarded by another agent. After watching each video, participants were asked to infer the what type of mind the guard has.

3.1 Participants

90 U.S. participants (as determined by their IP address; with ages M=36.77; SD=12.41) were recruited through Amazon's Mechanical Turk platform.

3.2 Stimuli

Stimuli consisted of fifteen videos of approximately 10 seconds (range = 2 - 22 secs; see

bit.ly/2O2nyUX for videos). Figure 2 shows schematics of these videos. In each video the thief navigated towards the treasure along a different route. The thief's behavior was hard-coded with the goal of eliciting different behaviors from the guard, but the guard's behavior was obtained programmatically by sampling from different generative mind models. Guard paths were then adjusted to make the videos more concise and to elicit different inferences (e.g., aligning the guard's search path so that it would miss the thief). Below we briefly describe the key components of each video.

3.3 Stimuli description

In Trial 1, the guard is initially positioned immediately behind the thief, and chases him all the way to the treasure. Because even the simplest model can produce this behavior, the trajectory did not reveal any aspects of the guard's mind. In Trial 2, the guard is inside a room, and exits as soon as the thief walks nearby, revealing that the guard can hear and react to the thief's presence even when separated by a wall. In Trial 3, the guard sees the thief walk, chases after him, and then begins to search the map upon losing him, thus revealing that the guard has beliefs, but no Theory of Mind because the guard did not intercept the thief along his path to the treasure. Trial 4 shows the guard using Theory of Mind to predict the thief's location and intercept him on his way to the treasure (rather than going to where the guard last saw the thief). Trial 5 shows a guard with no beliefs, who first chases after the thief (as the thief slows down), but stops moving after the thief is out of sight, because the guard has no belief capacity to store the position of the thief.

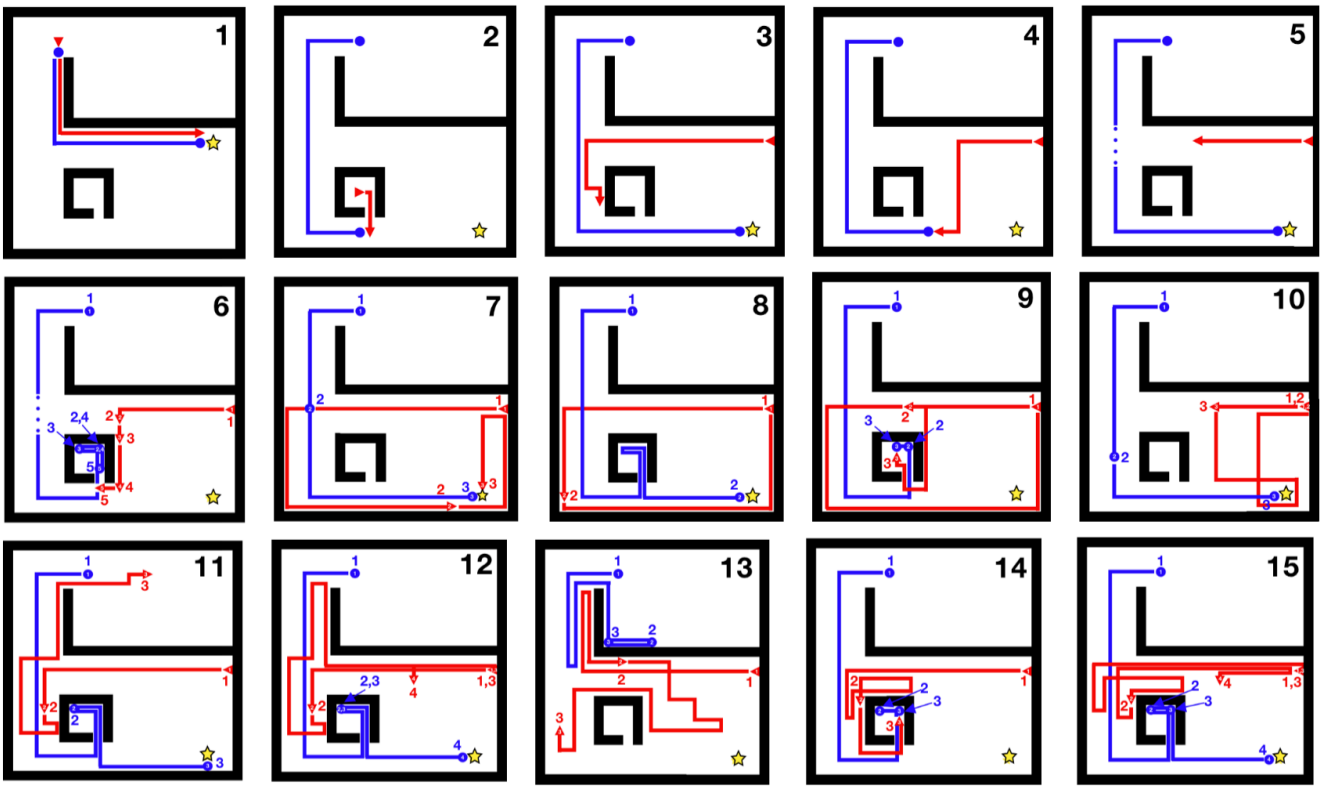

Figure 2: Schematics of the fifteen trials in our experiment. In each figure, the blue line indicates the thief's trajectory, the red line indicates the guard' trajectory, and the golden star indicates the treasure. Dotted lines indicate slower movement, and the numbers in each trajectory correspond to matched time points in the video.

Trial 6 begins in the same way as Trial 5, revealing that the guard has no belief representations. However, the thief then enters the room and begins moving around, prompting the guard to move each time he hears a new sound. This continually interrupted behavior occurs because the guard has no beliefs and is only able to react when he hears the thief move intermittently within the room. In Trial 7, the guard begins patrolling the area and then goes straight towards the treasure as soon as he sees the thief, using the *Theory of Mind* component of his mental model. Trial 8 is similar, with the difference that the guard never sees the thief and does not hear him as he moves around inside the room, giving little information about the guard's mind. Trial 9 is the same as Trial 8, but the guard does hear the thief moving around in

the room, and so switches his route to find the thief. Trial 10 shows a guard that spots the thief and then turns around and goes to the treasure after he stops seeing the thief, revealing that the guard has a Theory of Mind.

The last five trials show more complex trajectories. In Trial 11, the guard chases the thief, and then searches around as the thief moves inside the room (revealing that the guard does not hear and also does not forget about the thief's existence because of the duration of his search). Trial 12 is similar to Trial 11, with the difference that the guard eventually forgets about the thief after a short period of time and returns to his original position. In Trial 13, the thief retraces his steps after the guard sees him. The guard then continually searches for the thief, revealing he has stable belief representations, but also fails to hear the guard moving around on the other side of the wall. Finally, in Trials 14 and 15, the guard first spots and loses the thief. As the guard searches, he either hears the thief's movements (Trial 14) or does not and eventually returns to his original position after forgetting (Trial 15).

3.4 Procedure

Participants first read a short tutorial that explained the logic of the task. Participants then completed a questionnaire that ensured they had read the instructions and only participants who answered all questions correctly were given access to the task. The rest of the participants were told they had answered at least one question wrong and were given the chance to read back through the instructions and complete the questionnaire again.

Each participant was assigned five randomly-assigned videos (counterbalanced to get an equal number of participants in each trial). Each trial showed the video on repeat and four

questions: A *Belief* question asking “Does the guard have a memory? (does not immediately forget)”, a *Forgetting* question asking "Does the guard forget that the thief exists after a period of time? (approx. 2 seconds)", a *Hearing* question asking "Can the guard hear?", and a *Theory of Mind* question asking "Can the guard predict where the thief is going?". Each of these sliders had labels "Definitely No", "Unsure", and "Definitely Yes" at the left, middle, and right of the slider, respectively.

4. Results

Judgments were z-scored within participants and then averaged across trials. Figure 3 shows the results from the study. Each sub-plot illustrates the model and participant inferences about each type of mind (arranged in the same order as Figure 2). Overall, our model showed a correlation of $r = 0.70$ ($CI_{95\%}$: 0.54 - 0.81) against participant judgments.

Trial 12 shows a case where participant inferences closely followed those of our model. After losing the thief, the guard began searching in the wrong area, revealing that he had belief representations but no Theory of Mind. The fact that the agent failed to detect the thief as it moved inside the room reveals that he lacked hearing, and his eventual return to the starting point suggested that he forgot about the thief (note that, because values are z-scored, such that 0 indicates average inference value).

In Trial 2, because the guard was clearly unable to see the thief, but nonetheless reacted at the right moment, both participants and our model inferred that the agent could hear. Participants and our model were equally uncertain judgements about whether the guard *Forgets* or has a

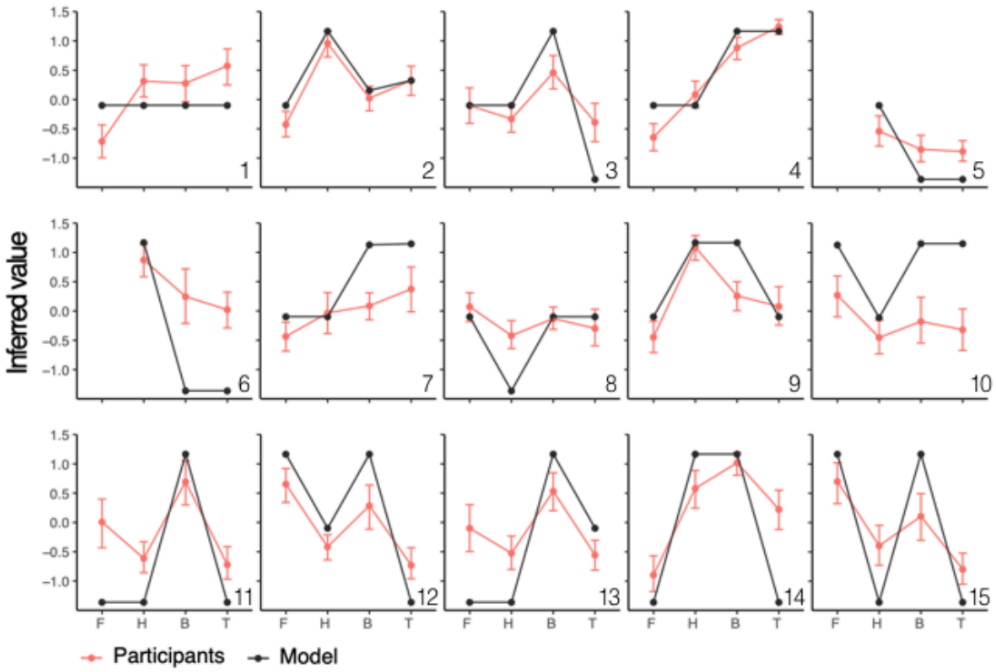

Figure 3: Results from the experiment. Each plot shows the results from the corresponding trial in Figure 2. F (forgetting) corresponds to the probability that agent's memory decays, H (hearing) to the probability that the agent can detect sounds, B (beliefs) to the probability that the agent has beliefs, and T (Theory of Mind) to the probability that the agent can predict other agents' goals and plan accordingly. The black lines show z-scored model predictions, and red lines show average z-scored participant judgments with 95% bootstrapped confidence intervals. Our model does not make predictions about forgetting when it infers that the agent lacks beliefs and we thus do not include those judgments in trials 5 and 6.

Theory of Mind, which can be attributed to the short duration of the video. There is not enough information in these few frames to make strong inferences about any of the traits besides *Hearing*.

Trial 1 shows a case where participants and our model disagree. Here, the guard was always one step behind the thief, not revealing any of its capacities. Nonetheless, participants were more likely to think that the agent could hear, had beliefs and Theory of Mind, and did not forget. Interestingly, these attributes correspond to the typical way we represent other agents, suggesting that participants had a prior distribution over what capabilities are likely which our model did not consider.

Looking at each individual capability, our model had a correlation of r = 0.86 ($CI_{95\%}$: 0.62 - 0.95) for *Hearing* against participant judgements. This is, unsurprising, given the visual nature of hearing inferences. *Theory of Mind* had a correlation of r = 0.62 ($CI_{95\%}$: 0.16 - 0.86) against participant judgements, suggesting that humans can recognize Theory of Mind in others rather easily. Forgetting showed a correlation of r = 0.62 ($CI_{95\%}$: 0.10 - 0.87). This capability can be difficult to infer because of its temporal nature and because our model may have had more precise estimates of the memory decay (see Discussion). Finally, *Beliefs* showed the lowest correlation, r = 0.58 ($CI_{95\%}$: 0.10 - 0.84). This was unexpected given the large effect *Beliefs* have on the guard's behavior. This could be due to a failure in conveying the meaning of beliefs to participants in the experiment, or because agents with no beliefs are rare, making them harder to reason about.

5. Discussion

Here we proposed a computational model of mind attribution as Bayesian inference over a family of generative models that transform mental states into observable actions. In a simple task showing two-dimensional demonstrations of a guard trying to catch a thief, we found that people can infer the structure of the underlying generative model, even from a few seconds of video. This is early evidence that humans use statistical inference to select a generative mental model when perceiving different types of agents.

Our work connects Bayesian models of action understanding with research in cognitive science that shows that people conceptualize different types of agents as having different types of minds (K. Gray et al., 2012). Although past work has argued that minds are structured around

two dimensions, agency and experience, no work, to our knowledge, has attempted to formalize these dimensions in precise computational terms. Our work is a first step in this endeavor. In our approach, experience can be considered the sensory component of the generative model—what the agent sees and hears—and agency can be considered the cognitive components—its beliefs, memory decay, and ability to mentalize about other agents. A challenge for future work is finding a full characterization of how each component in models of Theory of Mind relates to these two dimensions of mind perception.

5.1 Future Work and Limitations

Although participants performed surprisingly well in this task, particularly when considering that they had to infer an agent's mind from just a few seconds of video, they still showed some notable disagreements with our model (Figure 3). While more research is needed, at least two possibilities may help explain why this happened. A first possibility is that the videos contained too much information for participants to absorb in such a short time frame. If so, longer videos with multiple events that reveal agents' cognitive capacities may help participant inferences align more tightly with our model.

A second possibility is that searching over a space of generative models is difficult. In our experiment, the generative models that we considered may not directly map to the ones that we use when we reason about agents in the natural world—such as beetles, birds, squirrels, and scallops. As such, it is possible that we inadvertently increased task demands by asking participants to reason about a space of minds that they are not accustomed to reasoning about. Alternatively, it is possible that our generative model included too many details about the

domain, relative to what participants knew (e.g., our model had more precise estimates of agents' hearing radius, memory decay, etc). Thus, it is possible that a generative model with less information may show less confidence in a human-like way. We are currently exploring this possibility. Nonetheless, the fact that participants were able to reconstruct big components of agents' minds, suggests that people can indeed perform quick and flexible mind inferences, even in unusual situations.

A related limitation in our model is that we used a uniform prior over the space of possible minds. It is likely that people come with strong priors about what types of minds are more likely than others. For instance, participants may find it a priori plausible that an agent lacks a Theory of Mind, but not that an agent lacks an entire belief representation which may be more essential to our idea of mind. In current work we are estimating participants' priors empirically and integrating them into our model.

One outstanding question is how to formalize the complete space of minds that people can reason about. Our approach of instantiating minds as generative models allows us to ask this question in a more formal way. Under our framework, the problem is reduced to constructing a space of generative models that capture how we can reason about agents which contain or lack different representations and reasoning capabilities. With a sufficiently complex space of mental models, we may be able to make predictions about a wide array of minds in rich detail. In future research we will investigate this question.

5.2 Concluding Remarks

In our study, both participants and our model knew the agent's goal, making Eq. 2 easier to compute. In more realistic situations, observers have to compute an agent's type of mind, its mental states, and goals, all at once. Thus, it is possible that with this added uncertainty, learning the variability in minds that we encounter in the world may require more data than our task suggests, taking years to learn.

On the other hand, our experiment intuitively suggests that people might have more sophisticated capacities than what we tested. While our task focused on inferring a single mind, people might be able to infer multiple types of minds at once. In Trial 5, for instance (Figure 2), the guard's behavior reveals that it lacks belief representations. At the same time, the fact that the thief strategically slowed down to get the guard to move away from the treasure, suggests that the thief (1) knew that the guard lacked beliefs, (2) had a stable representation of the guard, and (3) could predict the guard's behavior. This intuition is consistent with classical work showing that we can read complex social interactions between multiple agents (Heider & Simmel, 1944). In future work we may test for this possibility.

Altogether, our work shows how, beyond an ability to infer the contents of other people's minds, people can also infer the type of mind behind an agent's behavior. Unlike other mental-state inference models, our work accounts for the fact that Charlie may be a different sort of agent and operate under vastly different mental and physical constraints than people do. This work is a first step towards a computational understanding of how we infer types of minds, and sheds light on how people can search through and attribute different mental models, based on how agents act and plan to fulfill their goals.

Acknowledgments

Thank you to:

Julian Jara-Ettinger for his invaluable time, support and mentorship;

The Computational Social Cognition Lab for their advice and guidance;

Natalia Córdova Sánchez for leading us through this process;

The Yale Cognitive Science program;

and my fellow peers;

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*.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2016). Children consider others' expected costs and rewards when deciding what to teach. *In Proceedings of the 38th annual conference of the cognitive science society* (pp. 559–564).

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

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

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315 (5812), 619–619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125 (1), 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*.

Grice, H. P., Cole, P., Morgan, J., et al. (1975). Logic and conversation. 1975, 41–58.

Hamlin, K., 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.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*.

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

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

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (under review). The naive utility calculus as a unified, quantitative framework for action understanding.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science*, 26 (5), 633–640.

Jern, A., Lucas, C. G., & Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. *In Advances in neural information processing systems*.

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

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.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

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