Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, & Joshua B. Tenenbaum*

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Cambridge, MA, 02139

Abstract

Social cognition depends on our capacity for *mentalizing*, or explaining an agent's behavior in terms of their mental states. The development and neural substrates of mentalizing are well-studied, but its computational basis is only beginning to be probed. Here we present a model of core mentalizing computations: inferring jointly an actor's beliefs, desires and percepts from how they move in the local spatial environment. Our Bayesian theory of mind (BToM) model is based on probabilistically inverting AI approaches to rational planning and state estimation, which extend classical expected-utility agent models to sequential actions in complex, partially observable domains. The model accurately captures the quantitative mental-state judgments of human participants in two experiments, each varying multiple stimulus dimensions across a large number of stimuli. Comparative model fits with both simpler "lesioned" BToM models and a family of simpler non-mentalistic motion features reveal the value contributed by each component of our model.

Keywords: theory of mind, mentalizing, Bayesian models of cognition

Rational quantitative attribution of beliefs, desires, and percepts in human mentalizing

Humans are natural mind readers. The ability to intuit what others think or want from brief nonverbal interactions is crucial to our social lives. If someone opens a door, looks inside, closes it, and turns around, what do we think they are thinking? Humans see others' behaviors not just as motions, but as intentional actions: the result of plans seeking to achieve their desires given their beliefs; and when beliefs are incomplete or false, seeking to update them via perception in order to act more effectively. Yet the computational basis of these mental state inferences remains poorly understood.

The aim of the present work is to reverse-engineer human mental state inferences in their most elemental form: the capacity to attribute beliefs, desires, and percepts to others which are grounded in physical action and the state of the world. Our goal is a formal, computational account, analogous in scope and explanatory power to computational accounts of visual perception [32, 24, 50] that represent some of the greatest successes of model-building in cognitive science. Here we report a key step in the form of a model of how humans attribute mental states to agents moving in complex spatial environments, quantitative tests of the model in parametrically controlled experiments, and extensive comparisons with alternative models. Taken together, this work brings us closer to understanding the brain mechanisms and developmental origins of theory of mind. It could also enable us to engineer machines which interact with humans in more fluent, human-like ways.

Mental state inference (or "mentalizing") in adults likely draws on a diverse set of representations and processes, but our focus is on a capacity that appears in some form in infancy [36, 53, 8, 17, 28, 6] and persists as a richer theory of mind develops through the first years of life [52, 51]. What we call core mentalizing is grounded in perception, action, and the physical world: It is based on observing and predicting the behavior of agents reaching for, moving toward, or manipulating objects in their immediate spatial environment, forming beliefs based on what they can see in their line of sight, and interacting with other nearby agents who

have analogous beliefs, desires, and percepts. In contrast to more explicit, language-based ToM tasks, which are only passed by older children, these core abilities can be formalized using the math of perception from sparse noisy data and action planning in simple motor systems. Hence core mentalizing is an aspect of social cognition that is particularly likely to be readily explained in terms of rational computational principles that make precise quantitative predictions, along the lines of what cognitive scientists have come to expect in the study of perception and motor control [50, 25, 3].

We will contrast two general approaches to modeling human core mentalizing, which can be broadly characterized as "model-based" versus "cue-based". The model-based approach says that humans have an intuitive theory of what agents think and do – a generative model of how mental states cause actions – which gets inverted to go from observed actions to mental state inferences. The cue-based approach assumes that mentalizing is based on a direct mapping from low-level sensory inputs to high-level mental states via statistical associations, e.g. "you want something because you reach for it". Although a cue-based, heuristic approach is unlikely to provide a satisfying account of full theory of mind, it may be sufficient to explain the simpler forms of action understanding at work when we see people reaching for or moving to objects in their immediate spatial environment. However, we contend that to explain even these basic forms of mentalizing requires a model-based, generative account.

Previous work has proposed both model-based [37, 2, 31, 38, 21, 20] and cue-based [4, 55] models of how both children and adults infer one class of mental states: desires, and associated notions such as goals, intentions, and preferences. Other model-based frameworks have considered inference of knowledge about world states and causal structure [15, 43, 21, 41], inference of beliefs based on unobserved events [18], or joint inference of knowledge and intentions in the context of epistemic trust and coordination [42, 5]. However, these models are unable to reason jointly about beliefs and percepts as well as desires, as core mentalizing requires. Our work addresses these limitations, and prior models can be seen as important special cases of

the model-based and cue-based models we formulate and test here.

To make our focus concrete, consider the scenario in Fig. 1a: a hungry student leaves his office looking for lunch from one of three food trucks: Korean (K), Lebanese (L), or Mexican (M). The university provides only two parking spots, so at most two trucks can be on campus on any given day; parking spots can also remain empty if only one truck comes to campus that day. When the student leaves his office (Frame 1), he can see that the Korean truck is parked in the near spot in the Southwest corner of campus. The Lebanese truck is parked in the far spot in the Northeast corner of campus, but he cannot see that, because it is not in his direct line of sight. Suppose that he walks past the Korean truck and around to the other side of the building, where he can now see the far parking spot: He sees the Lebanese truck parked there (Frame 2). He then turns around and goes back to the Korean truck (Frame 3). What can an observer infer about his mental state: his desires and his beliefs? Observers judge that he desires Mexican most, followed by Korean, and Lebanese least (Fig. 1a: Desire bar plot). This is a sophisticated mentalistic inference, not predicted by simpler (non-mentalistic) accounts of goal inference that posit goals as the targets of an agent's efficient (shortest path) reaching or locomotion. Here, the agent's goal is judged to be an object that is not even present in the scene. The agent appears to be taking an efficient path to a target that is his mental representation of what is behind the wall (the Mexican truck); and when he sees what is actually there, he pauses and turns around. Consistent with this interpretation, observers also judge that the student's initial belief was most likely that the Mexican truck was in the far parking spot (Fig. 1a: Belief bar plot).

These inferences have several properties that any computational model should account for. First, our inferences tacitly assume that the agent under observation is approximately rational [13] – that their behavior will employ efficient means to achieve their desires while minimizing costs incurred, subject to their beliefs about the world, which are rational functions of their prior knowledge and their percepts. Second, these inferences are genuinely metarepresentational [39, 28] – they represent other agents' models of the world, and their beliefs

about, and desires toward actual and possible world states. Third, these inferences highlight the three crucial causal roles that define the concept of *belief* in ToM [51, 6]: Beliefs are the joint effects of (1) the agent's percepts and (2) their prior beliefs, and also (3) the causes of the agent's actions (Fig. 1b). These multiple causal roles support multiple routes to inference: beliefs can be inferred both forward from inferences about an agent's percepts and priors, or backward from an agent's observed actions (and inferred desires), or jointly forward and backward by integrating available information of all these types. Joint causal inferences about the situation, how an agent perceives it, and what the agent believes about it are critical: Even if we couldn't see the far side of the building, we could still infer that some truck is located there if the student goes around to look and doesn't come back, and that whichever truck is there, he likes it better than the K truck. Finally, core mentalizing inferences are not simply qualitative and static but are quantitative and dynamic: the inference that the student likes Mexican after Frame 2 is stronger than in Frame 1, but even stronger in Frame 3, after he has turned around and gone back to the Korean truck.

We explain these inferences with a formal model-based account of core mentalizing as Bayesian inference over generative models of rational agents perceiving and acting in a dynamic world. In the remainder of the paper, we first describe the basic structure of this BToM (Bayesian Theory of Mind) model, along with several candidate alternative models. We then present two behavioral experiments showing that the BToM model can quantitatively predict people's inferences about agents' mental states in a range of parametrically controlled scenarios similar to those in Fig. 1a. Experiment 1 tests people's ability to jointly attribute beliefs and desires to others, given observed actions. Experiment 2 tests whether people can use their theory of mind to reason jointly about others' beliefs, percepts, and the state of the world.

Computational models

The Bayesian Theory of Mind (BToM) model formalizes mentalizing as Bayesian inference over a generative model of a rational agent. BToM defines the core representation of rational

agency (Fig. 1b) using partially observable Markov decision processes (POMDPs): an agent-based framework for rational planning and state estimation [22], inspired by the classical theory of decision-making by maximizing expected utility [49], but generalized to agents planning sequential actions that unfold over space and time with uncertainty due to incomplete information. POMDPs capture three central causal principles of core mentalizing highlighted by Fig. 1b: A rational agent (I) forms percepts that are a rational function of the world state, their own state, and the nature of their perceptual apparatus – for a visually guided agent, anything in their line of sight should register in their world model (*perception*); (II) forms beliefs that are rational inferences based on the combination of their percepts and their prior knowledge (*inference*); and (III) plans rational sequences of actions – actions that, given their beliefs, can be expected to achieve their desires efficiently and reliably (*planning*).

BToM integrates the POMDP generative model with a hypothesis space of candidate mental states, and a prior over those hypotheses, to make Bayesian inferences of beliefs, desires and percepts, given an agent's behavior in a situational context. More formally, a POMDP agent's beliefs are represented by a probability distribution over states derived by logically enumerating the space of possible worlds, e.g, in the food truck setting, the set of assignments of trucks to parking spaces (see SI Appendix: Beliefs). The agent's belief updates, given their percepts and prior beliefs, are modeled as rational Bayesian state estimates (see SI Appendix: Bayesian Belief Updating). A POMDP agent's desires are represented by a utility function over situations, actions, and events; in the food truck setting, agents receive a different real-valued utility for eating at each truck (see SI Appendix: Desires). The agent's desires trade off against the intrinsic cost, or negative utility of action; we assume the agent incurs a small constant cost per step, which penalizes lengthy action sequences. The BToM prior takes the form of a probability distribution over beliefs and desires – a distribution over POMDPs, each parameterized by a different initial probability distribution over world states and utility functions. The hypothesis spaces of desires and initial beliefs are drawn from discrete, approximately uniform grids (see SI Appendix: Belief

and Desire Priors). The agent's desires are assumed to be constant over a single episode, although their beliefs may change as they move through the environment or the environment itself changes.

Starting from these priors, BToM jointly infers the posterior probability of unobservable mental states for the agent (beliefs, desires, and percepts), conditioned on observing the agent's actions and the situation (both the world state and the agent's state) evolving over time. By using POMDPs to explicitly model the observer's model of the agent's perceptual, inferential and planning capacities, BToM crucially allows the situation to be partially observed by either the agent, the observer, or both. The joint system of the observer and the agent can also be seen as a special case of an interactive POMDP (or I-POMDP [14]), a generalization of POMDPs to multi-agent systems in which agents recursively model each other in a hierarchy of levels; in I-POMDP terms, the observer builds a non-recursive Level-1 model of a Level-0 observer (see SI Appendix: Rational Observer Model).

To give a flavor for how BToM computations work as Bayesian inferences, we sketch the model inference for a single observed event in which the agent forms a percept of their current situation, updates their beliefs from an initial belief B_0 to a subsequent belief B_1 and then chooses an action A. (The full BToM model generalizes this computation to a sequence of observed actions with recursive belief updating over time; see Methods: Eq. 2). In the single-action case, given the prior $Pr(B_0, D, S)$ over the agent's initial beliefs B_0, desires D and the situation S, the likelihoods defined by principles (I-III) above, and conditioning on observations A of how the agent then acts in that situation, the BToM observer can infer the posterior probability $Pr(B, D, P, S|A)$ of mental states (belief states $B = \{B_0, B_1\}$, desires D, and percepts P), and the situation S given actions A using Bayes' rule:

$$Pr(B, D, P, S|A) \propto Pr(A|B_1, D) \cdot Pr(B_1|P, B_0) \cdot Pr(P|S) \cdot Pr(B_0, D, S). \quad (1)$$

The likelihood factors into three components. $Pr(P|S)$ (corresponding to principle I) represents the observer's expectations about what the agent sees in a given situation. This model of an

agent's (visual) perception is based on the *isovist* from the agent's location: a polygonal region containing all points of the environment within a 360-degree field of view [10, 34] (see SI Appendix: Percept Distribution). $Pr(B_1|P, B_0)$ (corresponding to principle II) represents the observer's model of the agent's belief update from initial state B_0 to B_1. $Pr(A|B_1, D)$ (corresponding to principle III) represents the observer's model of the agent's efficient planning process. To capture ways in which rational agents' behavior may deviate from the ideal, and BToM observers' inferences may be correspondingly weaker or more graded, we assume that the agent acts by sampling actions with probability proportional to their exponentiated expected-utility (a softmax function with parameter β). The value of β is a parameter fit to participant judgments. Under this formulation, agents typically choose the highest-utility action at each time step but sometimes choose a non-optimal action.

Our goal is not to simply present and test this one model, but also to quantitatively contrast BToM with alternative accounts of core social perception. We compare BToM with three models inspired by previous work, including two model-based alternatives, and one cue-based alternative (see SI Appendix: Alternative Models). Earlier model-based accounts [2, 21, 20] used various adaptations of Markov Decision Processes (MDPs), special cases of POMDPs, which assume that the world state is fully observed and known to the agent. MDP-based models embody the core notion that intentional agents act efficiently to achieve their goals [8], but are limited by the assumption of a fully-observable world – they cannot represent beliefs which differ from the true world state, and they capture only the planning capacities of rational agency (i.e., only the bottom section of Fig. 1b), neglecting the perceptual and inferential capacities, as well as their interaction. We demonstrate these limitations by formulating an MDP-based alternative model called TrueBelief, and showing that it is unable to model the joint inferences about beliefs, desires, percepts and world states that are at the heart of core mentalizing, and that our BToM model captures. A second alternative model, called NoCost, establishes the need for the principle of efficiency in BToM and MDP-based accounts of people's belief and desire attributions by

assuming the agent's actions are cost-free, and therefore unconstrained by the tradeoff between effort and desire. We formulate both of these alternative models as "lesioned" special cases of the full BToM model.

Several cue-based accounts have used motion features extracted from visual input to model human inferences about intentional behavior [4, 55, 48]. Non-formal cue-based accounts of infant false belief reasoning have also been proposed [40], which argue that learned associations between agents, objects and scenes underlie classic demonstrations of infant false belief reasoning [36]. To test these accounts, we formulate a motion-based heuristic alternative, called MotionHeuristic, which maps cues extracted from the agent's motion and environment directly onto people's judgments of agent's beliefs, desires and percepts of the world. For Experiment 1, MotionHeuristic fit five linear weights for desires, and five for beliefs, for a total of ten weights. The first and second weights captured the statistical association between the agent's motion (1) toward each potential goal or (2) toward an alternative goal, and attributions of desire for that goal or belief that it was present. The last three weights fit the *a priori* bias toward each desire and belief rating. For Experiment 2, MotionHeuristic fit eight linear weights for each of six possible world ratings, for a total of 48 weights. Here, the first three weights captured the association between the agent's motion toward each spot and the rating that a more highly-desired cart was located there. The remaining five weights captured the *a priori* bias toward each possible world rating.

Results

We tested BToM and these alternative modeling approaches against human judgments in two experiments. In Exp. 1, participants saw a large number of dynamic "food truck" stimuli (as in Fig. 1a), and made quantitative inferences about agents' beliefs and desires given their observable actions. Belief inferences were made retrospectively, about what the agent believed was in the far parking spot before they set off along their path, given the information from the rest

of the agent's path. Exp. 2 used similar stimuli, but participants made inferences about agents' percepts and aspects of the world that only the agent could perceive. Both experiments manipulated key variables that should affect mental state attribution according to the BToM model: the structure and constraints of the environment (agent and object locations, physical barriers), the actions observed (their cost, whether they are ongoing or completed, the space of alternative actions), the set of goals (their number and presence, their utility, their availability), and the observability of the state of the world (what is or is not perceptible due to occlusion).

Experiment 1: Predicting human inferences about beliefs and desires

In each scenario of Experiment 1, participants observed a unique action-situation pair, and rated the agent's desire for each goal object (food trucks: Korean (K), Lebanese (L), and Mexican (M)), and the agent's initial belief about the state of the world (possible occupant of the far parking spot: L, M, or nothing (N)). BToM predictions were obtained by computing the posterior expectation of the agent's utility for each goal object, and the posterior expectation of the agent's degree of belief in each possible world state (see Methods: Eq. 2).

Model predictions were compared with participants' judgments on 73 distinct scenarios generated through a factorial design (see Methods: Experiment 1)[1], which can be organized into 7 basic scenario types (Fig. 2a-g) based on the environment's structure and the agent's actions. The scenario types differ in the number of trucks present: two trucks in a-d; one truck in e-g. The high-level structure of the agent's actions varies between types: initially, the agent can go either to the truck visible in the near parking spot (a, e) or go behind the building to see which truck (if any) is in the far spot (b, f). After checking the far spot, the agent can either return to the first truck (c, g), or continue to the far truck, if it is present (d). In all scenarios where the agent goes

[1]These scenarios were generated through a factorial experimental design that produced 78 scenarios in total, five of which were not consistent with an assumption of rational agency. We characterize these "irrational" scenarios in SI Appendix: Experiment 1 and analyze only the 73 rationally interpretable scenarios here.

around the building, at the moment when they can first see the far parking spot the agent either pauses for one frame before they continue to one of the trucks (c, d, g), or the trial ends with an incomplete path (b, f). Our model predictions also assume a one-frame pause at this moment. We first present an analysis of the model's quantitative predictive power over all scenarios, and then highlight the most revealing qualitative predictions across these scenario types.

Fig. 2h shows the quantitative desire and belief fits of BToM, averaged within the seven scenario-types defined above (Desire judgments r_{BSCV}=0.97 (95% CI $0.95, 0.98$), Belief judgments r_{BSCV}=0.91 (95% CI $0.87, 0.98$)). Fig. 3a shows the quantitative desire and belief fits of BToM at the level of all 73 individual scenarios (Desire judgments r_{BSCV}=0.91 (95% CI $0.89, 0.92$), Belief judgments r_{BSCV}=0.78 (95% CI $0.72, 0.85$)). These correlations and 95% confidence intervals (CIs) were computed using bootstrap cross-validation (BSCV; see Methods: Statistics), and all were highly significant.

Although BToM quantitatively predicted both belief and desire judgments, belief judgments were fit less well, and were also intrinsically more variable than desire judgments in ways that BToM predicted. Desire judgments varied primarily between the seven scenario types, but minimally within scenarios of the same type. This shows that small differences in scene geometry, which varied within scenario types, had minimal impact on desire judgments. Consistent with this finding, BToM predictions averaged within scenario types showed a high correlation with human desire judgments (r=0.95 (95% CI $0.94, 0.96$)), while BToM predictions at the individual scenario level showed no partial correlation with human judgments after controlling for scenario type (*partial* r=0.066 (95% CI $-0.067, 0.20$)). Human belief inferences varied in more complex ways – in particular, they varied both between and within the seven scenario types. BToM predictions averaged within scenario types, combined with individual scenario BToM predictions, explain 75 percent of the variance in human belief judgments (r=0.88 (95% CI $0.84, 0.90$)). Moreover, both types of predictions yielded significant partial correlations with human belief judgments when controlling for the other (individual-scenario:

partial r=0.28 (95% CI $0.15, 0.39$); type-averaged: *partial* r=0.63 (95% CI $0.54, 0.70$)). Consistent with the greater empirical variability in human belief inferences relative to desire inferences, the BToM model showed three times greater variance within scenario types for beliefs ($\sigma^2 = 0.34$) than for desires ($\sigma^2 = 0.11$; $F(218, 218) = 3.24$ (95% CI 2.59, *inf*, one-sided)). In short, people's belief judgments were much more affected (relative to desire judgments) by the small variations in scene geometry that varied within scenario types, and this overall trend was also predicted by the BToM model.

We also compared three alternative models with BToM, in terms of how well they could predict human belief and desire judgments across all 73 individual scenarios (see SI Appendix: Experiment 1 for details). Fig. 3b,c show that both TrueBelief and NoCost were able to predict desire judgments to some extent but significantly less well than BToM (r_{BSCV}=0.72 (95% CI $0.68, 0.77$), r_{BSCV}=0.75 (95% CI $0.69, 0.81$), respectively). Fig. 3b,c show that neither TrueBelief nor NoCost could predict belief judgments at all ($r_{BSCV} = -0.022$ (95% CI $-0.16, 0.11$), r_{BSCV}=0.10 (95% CI $0.045, 0.16$), respectively). The motion-based heuristic was able to predict belief inferences as well as BToM (r_{BSCV}=0.77 (95% CI $0.69, 0.83$)) but fared worse than all models on desire inferences (r_{BSCV}=0.62 (95% CI $0.51, 0.70$)). Fig. 3d shows that although the motion-based heuristic correlates relatively highly with the human data, it is qualitatively poorly calibrated to human judgments – the range of model predictions is compressed, and the predictions mostly fall into two clusters which are aligned with the data, but which have little variance internally. These results illustrate the value of the full POMDP architecture underlying the BToM model, and more generally the need to model joint inferences about beliefs and desires, even if we only want to predict one of these two classes of mental states.

A more qualitative analysis of specific scenario types illustrates how BToM captures many subtleties of human mentalizing. Fig. 2a-c show that both BToM and human judgments are consistent with the intuitive inferences about beliefs and desires sketched in the Introduction. BToM closely predicts the differences between these scenario types, and also between these

scenarios and analogous ones in which no truck is present in the other spot (Fig. 2e-g). For instance, in scenarios with two trucks present (Fig. 2a-d), BToM correctly predicts stronger inferences when the agent checks which truck is parked in the far spot (Fig. 2c,d) as opposed to going straight to the K truck in the near spot (Fig. 2a): only in Fig. 2c,d can we clearly distinguish the strengths of the agent's desire for all three trucks, and the strengths of the agent's initial beliefs for all three possible worlds. When there is no truck parked in the far spot, BToM correctly predicts how inferences become weaker when the agent goes to check that spot (compare how belief and desire inferences for M and L trucks become indistinguishable in Fig. 2f,g, relative to 2b,c), but not when the agent goes straight to the near spot without checking (observe no effect of the second truck's presence in Fig. 2a vs. 2e). BToM also predicts stronger inferences from complete paths as opposed to incomplete paths (compare both the belief and desire inferences in Fig. 2c,d with 2b), and the subtle differences in people's judgments about the agents' beliefs and desires in the two incomplete path scenarios, varying in whether a second truck is present: When a truck is present in the far spot, the agent's brief pause at the end of the incomplete path is interpreted as weak evidence that the second truck might not be what the agent was hoping to see (Fig. 2b), while if there is no truck parked in the far spot, the same brief pause is uninformative about which of the other two trucks the agent was hoping to see (Fig. 2f). These are just a few examples of the qualitative predictions that the BToM model makes in accord with human intuition – predictions that are not specifically or explicitly wired in, but that fall out naturally from the general principle of Bayesian inference over generative models of rational agents' planning and state estimation.

Experiment 2: Reasoning about others' percepts from their actions

From early childhood, mentalizing is useful not only in explaining people's behavior, but also in learning about unobserved aspects of the world by observing other actors and inferring what they must have seen and believed in order to explain the way they

acted [8, 9, 33, 15, 43, 21, 18]. Our BToM model was not developed to account for such social inferences, but if it is really capturing core mentalizing abilities, it should generalize to handle them naturally. Exp. 2 tested this hypothesis, using similar scenarios to Exp. 1, in which an agent searched for his favorite food cart in a spatially complex environment that constrained movements and visibility. Now, however, participants could not observe the locations of three food carts, and were tasked with inferring these locations on the basis of the agent's actions. The carts served Afghani (A), Burmese (B), and Colombian (C) food, and they could be in any of three locations: North, West, and East spots (see Fig. 4a). Participants were told that the agent preferred A over B, and both A and B over C, and would always search through the environment until he found the most preferred cart that was open. To add further complexity, carts A and B could be either open or closed, while C was assumed to always be open (so the agent always had at least one available option). The agent thus could be in one of 24 possible worlds ($24 = 6 \times 2 \times 2$, for 6 assignments of carts to locations, and 2 states (open, closed) for each of the A and B carts). Although the cart locations and availabilities (specifying which of the 24 possible worlds applied) were hidden from participants, they were observable to the agent – though only within line of sight. Based only on the agent's search behavior, participants were asked to infer the locations of all three carts.

We generated 19 experimental scenarios (including three simplified introductory scenarios; see Methods: Experiment 2), varying the agent's path and including both complete paths (when the agent had successfully found the best available cart) and incomplete paths (showing only the first part of a complete path). Fig. 4a shows the environment and one representative complete path from the experiment: initially only the North location is within the agent's line of sight; after taking several steps, the agent also sees what is present in the West location; finally, the agent returns to the starting point and chooses the cart in the North location. After observing this path, participants rated the probability of all six possible spatial configurations of the three food carts. Participants overwhelmingly judged one configuration as most probable, and the BToM model agrees: cart B is in the North location, cart A is in the West, and cart C is the East. The model

captures human performance by first generating a contingent POMDP plan for each of the 24 possible worlds, and for each initial belief the agent could hold (see SI Appendix: Belief and Desire Priors), then computing a likelihood assignment for the agent's action conditioned on each possible cart configuration (and marginalizing over initial belief and whether the carts were open or closed; see Methods: Eq. 3). Assuming equal prior probabilities on all possible worlds and initial beliefs, and applying Bayes' rule, these likelihoods determine the relative posterior probabilities on possible worlds that participants are asked to judge.

Analogous judgments to those shown in Fig. 4a were made in all 19 scenarios for all six cart configurations, for a total of 114 judgments per participant. Fig. 5a shows that BToM accurately predicted participants' mean judgments (Fig. 3g, r_{BSCV}=0.91 (95% CI $0.86, 0.94$)). We also compared the performance of our three alternative models (SI Appendix: Experiment 2). Fig. 5d shows that the motion-based heuristic correlates only weakly with human judgments (r_{BSCV}=0.61 (95% CI $0.10, 0.83$)), arguing for the necessity of mental-state reasoning even in a task that does not directly ask for it. Fig. 5b,c show that both TrueBelief and NoCost also fit poorly, suggesting that joint reasoning about beliefs, percepts, and efficient action-planning is essential for this task (r_{BSCV}=0.63 (95% CI $0.24, 0.83$), r_{BSCV}=0.46 (95% CI $0.17, 0.79$), respectively).

Fig. 4b-g illustrate the BToM model's ability to capture analogous judgments for more and less complex paths, including incomplete paths. In Fig. 4b, the agent goes directly to the North location, suggesting that they saw cart A there (and A was open), but leaving the locations of B and C unknown. Fig. 4d shows a path that begins like Fig. 4a but terminates in the West location. Here, both people and BToM infer that the agent probably saw A at the West spot, but it is also possible that they saw A in the North location, and it was closed, leading them to go West where they found the B cart open. Fig. 4g shows the longest trajectory from this experiment, with the agent first seeing the North location, then checking West, then East, before returning to the West location. Although the path is longer than that in Fig. 4a, people's inferences are less certain

because the multiple reversals could be explained by several different cart configurations depending on which carts are open or closed; BToM captures this same ambiguity. Fig. 4c,e show incomplete paths, which leave both people and BToM more uncertain about the world configuration in ways that reflect rational expected values: locations that the agent has seen but moved away from are most likely to contain his least preferred option C; if he has seen and moved away from two different locations (as in Fig. 4e), most likely they contain his two least preferred options B and C (although in unknown order). Fig. 4f shows a continuation of Fig. 4e which terminates at the East location: people's inferences are similar in both scenarios, which BToM explains by predicting the likely outcome of the path as soon as the agent turns away from the West location; the additional steps in Fig. 4f provide little additional information beyond the partial path in Fig. 4e. As with Exp. 1, these and many other qualitative predictions consistent with human intuitions fall out naturally from BToM, simply from the principle of mentalizing based on Bayesian inference over models of rational agents and the constraints and affordances of the situation.

Discussion

We proposed that core mental state inferences can be modeled as Bayesian inversion of a probabilistic state-estimation and expected-utility-maximizing planning process, conditioned on observing others' actions in a given environment. Our BToM model quantitatively predicted many social inferences in complex novel scenarios, varying both environmental contexts and action sequences, and including both inferences about others' beliefs, desires and percepts, as well as unobservable world states posited to explain how others explore and exploit their environment. Alternative models which did not represent others' costs of action or uncertain world beliefs consistently diverged from human judgments, as did combinations of special-purpose motion features which did not model mental states and had to be custom-fit to each experiment. That people's judgments require joint reasoning about beliefs, desires, and

percepts is further supported by the failure of models which lesioned any one of these representations: these models show a deficit not only in the missing representations, but also in the remaining mental state inferences with which they are causally entwined. Bayesian inversion of models of rational agents thus provides a powerful quantitative model of how people understand the psychological and social world.

It is important to clarify what we mean when we say that participants, like the BToM model, are performing joint inferences about an agent's beliefs, desires and percepts. To us, joint inference is about representing a joint hypothesis space of the agent's beliefs and desires, such that in explaining a complete action sequence, the observer's posterior distributions over the agent's beliefs and desires are coupled; inferences about the agent's beliefs inform inferences about the agent's desires, and/or vice versa. In the Marr hierarchy of levels of explanation [32], this is a computational-level claim. It does not require that algorithmically, at each point in time, the observer is simultaneously considering the full joint space of all possible belief-desire combinations and updating their inferences about beliefs and desires simultaneously. The algorithmic implementation of our BToM model in fact works this way, but this could be intractable for more complex settings, and indeed there are other inference algorithms that people could use to perform joint belief-desire inference more efficiently by alternating between updating belief inferences given current desire inferences and updating desire inferences given current belief inferences. For instance, in Experiment 1, observers could initially posit a uniform distribution for the agent's beliefs, then infer the agent's desires from their full trajectory while tracking their belief updates based on isovists, and finally use the inferred desires to retrospectively infer the agent's most likely initial beliefs. Developing such an algorithmic account of BToM inferences and testing it on a more general set of stimuli and inferences is an important direction for future work.

Similarly, inverse rational POMDP planning is a computational-level theory of human core mentalizing. Although optimal POMDP planning is computationally intractable in general,

optimal POMDP solutions can be tractably approximated in certain cases [19], and modern solvers can scale to problems with millions of states [44, 45]. In the domains we study here, near-optimal solutions can be computed efficiently using approximate solvers: for Experiment 1 we used a grid-based approximation [30], and for Experiment 2 we used a point-based algorithm [27]. Testing a broader range of approximate solvers within BToM will be critical for developing algorithmic theories of human core mentalizing, and for scaling the framework to more complex domains.

The POMDP formulation we adopt here is at best only a first approximation to the true agent model in core mentalizing, but its value is in giving an elegant integrated account of three crucial functions that beliefs should fulfill in any intuitive theory (Fig. 1b) – rational updating in response to both the agent's *perception* of their environment and *inference* based on their other beliefs, and rational action *planning* to best achieve the agent's desires given their beliefs – in a form that embeds naturally inside a Bayesian cognitive model to capture judgments from sparse, incomplete data. A complete account of mentalizing will likely invoke both less and more sophisticated agent models. At one extreme, entirely model-free approaches based on motion features failed to explain judgments in our tasks, but the deep network architectures that have driven recent successes in computer vision could help to speed up routine BToM computations by learning to generate fast approximate inferences in a bottom-up, feed-forward pass [26]. At the other extreme, theory of mind in adults draws on language to represent recursive beliefs and desires, with propositional content that goes well beyond what infants can entertain [11]. Consider the belief that "Harold believes that the morning star is beautiful, but not as beautiful as the evening star, and not nearly as beautiful as Julie wants him to think she is." It is an open question whether BToM models can be extended to such cases.

BToM models can be extended to include richer environment and action models sensitive to intuitive physics [3], and multi-agent planning to parse competitive or cooperative social interactions such as chasing and fleeing [1] or helping and hindering [16]. Generative models of

multiple agents' interactive behavior can be expressed as Markov games [29], and simpler game-theoretic models have already been useful in modeling other theory of mind tasks [54]. Extending these models to capture (as BToM requires) agents' subjective beliefs about the world, and nested beliefs about other agents' beliefs [47, 12], is an important direction for future research.

Another aspect of theory of mind that our model does not fully address is the distinction between instrumental (expected utility maximizing) goals and epistemic (information seeking) goals. People intuitively make this distinction: for instance, in Fig. 1a, if asked why did the agent go around the wall, people might reply that his preferred truck is M, his goal was to get to that truck, and he was hoping M would be there, but one might also say that his immediate goal was to see what truck was parked on the other side, and with the intention of going to that truck if it turned out to be his preferred one. The latter explanation posits an epistemic subgoal as part of a larger instrumental plan. Extending our model to include explicit epistemic goals is an important direction for future work. However, it is interesting that even without explicit epistemic goals, BToM is able to explain a wide range of information-seeking behaviors as implicit epistemic goals that emerge automatically in the service of an instrumental plan. For instance, imagine that the wall in the scenarios in Exp. 1 has a window that allows the agent to look but not pass through to the other side (extending only the isovist, but not the potential routes). In some cases, BToM would predict that the agent should first go to the window, rather than moving around the wall, provided the window is closer to its start position[2].

Although our work was motivated by action understanding abilities that are present in young children, we evaluated our model only against adult judgments. It is thus an open question at what age children become able to make the kinds of inferences our experimental tasks tap into, and what if any stage in children's development of mentalizing capacities our model might be capturing. Our definition of core mentalizing is not meant to imply that BToM is an account of

[2]We thank an anonymous reviewer for this scenario.

infants' innate capacities for understanding agents, what Spelke and Carey have called "core knowledge" [46]. Our use of the term "core" is meant to imply that our model builds only on representations of the world – specifically, representations of space and spatial constraints on visibility, objects as goals, and actions as movement guided by efficiency – that are part of early-emerging human knowledge, and metarepresentations of beliefs and desires defined over those representations. In our view there is much room for core mentalizing capacities to develop and change through experience; regardless of the extent to which they build on innate capacities, they need not be hard-wired.

Finally, it is worth commenting on the contrast between BToM's framing of human planning as approximately maximizing expected utility, and prominent experiments in psychology suggesting the opposite [23]. In part this may reflect the limited domain where core mentalizing operates, relative to studies in behavioral economics: Moving through space to reach concrete sources of reward (such as food), where costs are due primarily to energy expended (or distance traveled), is a developmentally and evolutionarily ancient setting where humans may well plan efficiently, have finely-tuned expectations that others will behave likewise, and make approximately optimal Bayesian inferences subject to these assumptions. In these cases, mechanisms of human action planning and action understanding may converge, yielding mental-state inferences via BToM that are not only rational but veridical. But humans could also overextend their core mentalizing capacities to settings where people do not in fact plan well by expected-utility standards: BToM-style models could be correct in positing that people assume others act rationally in some domain, even if modeling people as rational actors is not correct there. This tension could explain why demonstrations of people violating expected-utility norms are often so compelling. They are counter-intuitive, in domains where our intuitions over-attribute rationality to ourselves and others. And the fact of their counter-intuitiveness may be the best evidence we have that intuitive theories of minds – if not always actual human minds – are rational to the core.

Methods

Computational Modeling

Full technical details of BToM are available in SI Appendix: Computational Modeling. Here we outline the basic technical information underlying Eq. 1. The BToM observer uses partially observable Markov decision processes (POMDP) to represent agents' beliefs, desires, percepts, actions, and environment. A POMDP [22] represents a state space $\mathcal{S}$, a set of actions $\mathcal{A}$, a state transition distribution $\mathcal{T}$, a reward function $\mathcal{R}$, a set of observations Ω, and an observation distribution $\mathcal{O}$. We decompose the state space $\mathcal{S}$ into a fully observable state space, $\mathcal{X}$ (the agent location), and a partially observable state space $\mathcal{Y}$ (the truck locations and availability)[3], such that $\mathcal{S} = \langle \mathcal{X}, \mathcal{Y} \rangle$.

The BToM observer's belief and desire inferences (Exp. 1) are given by the joint posterior marginal over the agent's beliefs b_t and rewards r at time t, conditioned on the state sequence $x_{1:T}$ up until $T \geq t$, and the world state y:

$$P(b_t, r | x_{1:T}, y). \tag{2}$$

The BToM observer's inferences of world states (Exp. 2) are given by jointly inferring beliefs, desires, and world states, and then marginalizing over the agent's beliefs and desires:

$$P(y | x_{1:T}) = \sum_{b_t, r} P(b_t, r | x_{1:T}, y) P(y). \tag{3}$$

Experiment 1

Experimental Design. Fig. 6 shows our factorial design, which varied four factors of the situation and action: (1) goal configuration, (2) environment configuration, (3) initial agent

[3]Technically, this is a mixed-observability MDPs (MOMDP) [35], an extension of POMDPs in which portions of the state space are fully observable, as in MDPs, and portions of the state space are partially observable, as in POMDPs. However, we will refer to the model as a POMDP for consistency and clarity, as this term is more widely known.

location, and (4) agent's high-level path. Of the scenarios generated by varying these factors, 78 were valid scenarios in which the actions obeyed the constraints of the environment, i.e., not passing through obstacles, and ending at a present goal. For example, combinations of Environment 1 with Agent path 7 were invalid, because the path passes through the obstacle. Combinations of Goal configuration 2 with Agent path 7 were also invalid, because the path ends at a spot with no goal present. The full set of experimental scenarios is shown in SI Appendix: Experiment 1 Scenarios and Results.

Five factors were randomized between subjects. Truck labels were randomly scrambled in each scenario (for clarity we describe the experiment using the canonical ordering Korean (K), Lebanese (L), Mexican (M)). Scenarios were presented in pseudo-random order. Each scenario randomly reflected the display vertically and horizontally so that subjects would remain engaged with the task and not lapse into a repetitive strategy. Each scenario randomly displayed the agent in 1 of 10 colors, and sampled a random male or female name without replacement. This ensured that subjects did not generalize information about one agent's beliefs or desires to agents in subsequent scenarios.

Stimuli. Stimuli were short animations displayed at a frame-rate of 10 Hz, depicting scenarios featuring an agent's path through a static environment. Three frames from an example stimulus are shown in Fig. 7a.

Procedure. Subjects first completed a familiarization stage that explained all details of our displays and the scenarios they depicted. To ensure that subjects understood what the agents could and couldn't see, the familiarization explained the visualization of the agent's isovist, which was updated along each step of the agent's path. The isovist was displayed during the testing stage of the experiment as well.

The experimental task involved rating the agent's degree of belief in each possible world (Lebanese truck behind the building (L); Mexican truck behind the building (M); or nothing behind the building (N)), and rating how much the agent liked each truck. All ratings were on a

7-point scale. Belief ratings were made retrospectively, about what agents thought was in the far parking spot at the beginning of the scenario, based on their subsequent path. The rating task counterbalanced the side of the monitor on which the "likes" and "believes" questions were displayed.

Participants. Participants were 17 members of the MIT Brain and Cognitive Sciences subject pool, 6 female, and 11 male. One male subject did not understand the instructions and was excluded from the analysis. All gave informed consent, and were treated according to protocol approved by MIT's Institutional Review Board.

Experiment 2

Experimental Design. Scenarios involved 24 possible worlds (6 possible permutations of the carts' locations multiplied by 4 permutations of carts A and B being open or closed), and were generated as follows. We assumed that the agent always started at the entrance of the North hallway, and chose between entering that hall, going to the West hall, or going to the East hall. An exhaustive list of possible paths was constructed by listing all possible combinations of short-term goals of the agent (go to entrance of W hall, go to entrance of N hall, or go to entrance of W hall), assuming that the first time a hall is selected it is for the purpose of exploration, and any selection of a hall that had been selected previously is for exploitation, meaning the agent has chosen to eat there. From the eleven exhaustively enumerated paths, two paths that only produced permutations of beliefs were removed, leaving a total of 9 complete paths. In addition, 7 incomplete paths (subsequences of the 9 complete paths) which produce different judgments were selected. Lastly, three of these paths were duplicated in initial displays in which all carts are assumed to be open, shown to familiarize subjects with the task. This produced a total of 19 different paths (see SI Appendix: Experiment 2 Scenarios and Results) for which each subject rated six possible configurations of carts, for a total of 114 judgments per subject. Food cart names as well as stimulus order were randomized across subjects (for clarity we describe the experiment using the

canonical cart names and ordering: Afghani (A), Burmese (B), and Colombian (C)).

Stimuli. Stimuli were static images depicting scenarios featuring an agent's path through a static environment. Example stimuli from three scenarios are shown in Fig. 7b.

Procedure. Subjects first completed a familiarization stage, which explained the basic food cart setting, then provided judgments for three introductory scenarios where the food carts were assumed to always be open. Next, the possibility that carts could be closed was introduced with a step by step example. The remaining 16 experimental scenarios immediately followed.

In each scenario, subjects were shown either a complete or an incomplete path. They were then asked to rate on a scale from 0 to 10 (with 0 meaning "Definitely Not"; 10 "Definitely"; and 5 "Maybe") how likely each of six possible cart configurations was to be the real one.

Participants. 200 U.S. residents were recruited through Amazon Mechanical Turk. 176 subjects were included in the analysis, with 24 excluded due to server error. All gave informed consent, and were treated according to protocol approved by MIT's Institutional Review Board.

Statistics

Bootstrap Cross-Validation (BSCV). Bootstrap Cross-Validation is a non-parametric technique for assessing goodness of fit [7]. BSCV is useful when comparing different models with different numbers of free parameters, as we do here, because it naturally controls for possible overfitting.

For each experiment, we generate $100,000$ random splits of the total set of individual scenarios into non-overlapping training and test sets. Identical training and test sets are used to evaluate each model. We then compute the predictive accuracy (r, or Pearson correlation coefficient) of each model on each test set, using parameters fit to the corresponding training set. The statistic r_{BSCV} denotes the median value, and confidence intervals span 95% of the $100,000$ sampled values. Bootstrapped hypothesis tests compute the proportion of samples in which the r value of one model exceeds that of another.

BSCV analyses for BToM, TrueBelief, and NoCost selected best-fitting parameters on each iteration from the discrete ranges shown in SI Appendix: Experiment 1 and SI Appendix: Experiment 2. For MotionHeuristic, best-fitting parameters were selected on each iteration from a continuous range using linear regression.

It may be surprising that BSCV correlations often exceed overall correlations. This happens because the Pearson r statistic involves estimating slope and intercept values to optimize the model fit to each test set. However, because we use the same bootstrapped training and test sets to evaluate each model, the effect does not favor any particular model.

Data Availability

The data that support the findings of this study are available at https://github.com/clbaker/BToM.

Code Availability

The code for all models and analyses that support the findings of this study are available at https://github.com/clbaker/BToM.

References

[1] C. L. Baker, N. D. Goodman, and J. B. Tenenbaum. Theory-based social goal inference. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1447–1455, 2008.

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

[3] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *P. Natl. Acad. Sci. USA*, 110(45):18327–18332, 2013.

[4] P. W. Blythe, P. M. Todd, and G. F. Miller. How motion reveals intention: categorizing social interactions. In G. Gigerenzer, P. M. Todd, and the ABC Research Group, editors, *Simple heuristics that make us smart*, pages 257–286. Oxford University Press, New York, 1999.

[5] J. Butterfield, O. C. Jenkins, D. M. Sobel, and J. Schwertfeger. Modeling aspects of theory of mind with markov random fields. *Int. J. Soc. Robot*, 1:41–51, 2009.

[6] S. Carey. *The Origin of Concepts*. Oxford Univ. Press, Oxford, 2009.

[7] Paul R. Cohen. *Empirical methods in artificial intelligence*. MIT Press, Cambridge, MA, 1995.

[8] G. Csibra, S. Biró, O. Koós, and G. Gergely. One-year-old infants use teleological representations of actions productively. *Cognitive Sci.*, 27:111–133, 2003.

[9] Gergely Csibra and Ágnes Volein. Infants can infer the presence of hidden objects from referential gaze information. *British Journal of Developmental Psychology*, 26(1):1–1, 2008.

[10] L. S. Davis and M. L. Benedikt. Computational models of space: Isovists and isovist fields. *Computer Graphics and Image Processing*, 11:49–72, 1979.

[11] J. G. de Villiers and P. A. de Villiers. Complements enable representation of the contents of false beliefs: Evolution of a theory of theory of mind. In S. Foster-Cohen, editor, *Language acquisition*. Palgrave Macmillan, Hampshire, United Kingdom, 2009.

[12] Prashant Doshi, Xia Qu, Adam Goodie, and Diana Young. Modeling recursive reasoning by humans using empirically informed interactive pomdps. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010.

[13] G. Gergely and G. Csibra. Teleological reasoning in infancy: the naïve theory of rational action. *Trends Cogn. Sci.*, 7(7):287–292, 2003.

[14] Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *J. Artif. Intell. Res.*, 24:49–79, 2005.

[15] N. D. Goodman, C. L. Baker, and J. B. Tenenbaum. Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2759–2764, 2009.

[16] J. K. Hamlin, T. D. Ullman, J. B. Tenenbaum, N. D. Goodman, and C. L. Baker. The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Dev. Sci.*, 16(2):209–226, 2013.

[17] J. K. Hamlin, K. Wynn, and P. Bloom. Social evaluation by preverbal infants. *Nature*, 450:557–560, 2007.

[18] D. Hawthorne-Madell and N. D. Goodman. So good it has to be true: Wishful thinking in theory of mind. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 884–889, 2015.

[19] D. Hsu, W. S. Lee, and N. Rong. What makes some pomdp problems easy to approximate? In *Advances in Neural Information Processing Systems 20*, 2007.

[20] J. Jara-Ettinger, H. Gweon, J. B. Tenenbaum, and L. E. Schulz. Children's understanding of the costs and rewards underlying rational action. *Cognition*, 140:14–23, 2015.

[21] A. Jern and C. Kemp. A decision network account of reasoning about other people's choices. *Cognition*, 142:12–38, 2015.

[22] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[23] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, 2011.

[24] D. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge Univ. Press, Cambridge, 1996.

[25] K. P. Körding and D. M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427:244–247, 2004.

[26] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: An imperative probabilistic programming language for scene perception. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[27] Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, volume 4, 2008.

[28] Alan M. Leslie, Ori Friedman, and Tim P. German. Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, 8(12):528–533, 2005.

[29] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163, 1994.

[30] W. S. Lovejoy. Computationally feasible bounds for partially observed markov decision processes. *Operations Research*, 39(1):162–âĂŞ175, 1991.

[31] C. G. Lucas, T. L. Griffiths, F. Xu, C. Fawcett, A. Gopnik, T. Kushnir, L. Markson, and J. Hu. The child as econometrician: A rational model of preference understanding in children. *PLoS ONE*, 9(3), 2014.

[32] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

[33] Henrike Moll and Michael Tomasello. 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, 7(1):F1–F9, 2004.

[34] V. I. Morariu, V. S. N. Prasad, and L. S. Davis. Human activity understanding using visibility context. In *IEEE/RSJ IROS Workshop: From sensors to human spatial concepts (FS2HSC)*, 2007.

[35] S. C. W. Ong, S. W. Png, D. Hsu, and W. S. Lee. Pomdps for robotic tasks with mixed observability. In *Robotics: Science and Systems*, volume 5, 2009.

[36] K. H. Onishi and Renée Baillargeon. Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258, 2005.

[37] Erhan Oztop, Daniel Wolpert, and Mitsuo Kawato. Mental state inference using visual control parameters. *Cognitive Brain Research*, 22:129–151, 2005.

[38] Peter C. Pantelis, Chris L. Baker, Steven A. Cholewiak, Kevin Sanik, Ari Weinstein, Chia-Chen Wu, Joshua B. Tenenbaum, and Jacob Feldman. Inferring the intentional states of autonomous virtual agents. *Cognition*, 130:360–379, 2014.

[39] J. Perner. *Understanding the representational mind*. MIT Press, Cambridge, MA, 1991.

[40] J. Perner and T. Ruffman. Infants' insight into the mind: How deep? *Science*, 308(5719):214–216, 2005.

[41] A. N. Rafferty, M. M. LaMar, and T. L. Griffiths. Inferring learners' knowledge from their actions. *Cognitive Science*, 39:584–618, 2015.

[42] P. Shafto, B. Eaves, D. J. Navarro, and A. Perfors. Epistemic trust: modeling children's reasoning about others' knowledge and intent. *Developmental Science*, 15(3):436–447, 2012.

[43] P. Shafto, N. D. Goodman, and M. C. Frank. Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, 7(4):341–351, 2012.

[44] D. Silver and J. Veness. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems 23*, 2010.

[45] A. Somani, N. Ye, D. Hsu, and W. S. Lee. Despot: Online pomdp planning with regularization. In *Advances in Neural Information Processing Systems 26*, 2013.

[46] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

[47] A. Stuhlmüller and N. D. Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *J. Cognitive Systems Research*, 2013.

[48] P. D. Tremoulet and J. Feldman. The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, 29:943–951, 2006.

[49] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton Univ. Press, Princeton, NJ, 1953.

[50] Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nat. Neurosci.*, 5(6):598–604, 2002.

[51] H. M. Wellman. *Making Minds: How Theory of Mind Develops*. Oxford Univ. Press, Oxford, 2014.

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

[53] A. L. Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69:1–34, 1998.

[54] W. Yoshida, R. J. Dolan, and K. J. Friston. Game theory of mind. *PLoS Comput. Biol.*, 4(12):1–14, 2008.

[55] Jeffrey M. Zacks. Using movement and intentions to understand simple events. *Cognitive Science*, 28:979–1008, 2004.

Correspondence. Correspondence and material requests concerning this article should be addressed to Joshua B. Tenenbaum, MIT Department of Brain and Cognitive Sciences, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: `jbt@mit.edu`

Acknowledgments. This work was supported by the Center for Brains, Minds & Machines (CBMM), under NSF STC award CCF-1231216; by NSF grant IIS-1227495 and by DARPA grant IIS-1227504. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank three anonymous reviewers for helpful suggestions.

Author Contributions. C.L.B., R.S., and J.B.T. designed Experiment 1. C.L.B. ran Experiment 1. C.L.B. implemented the models and performed the analyses of Experiment 1. J.J.E., C.L.B., and J.B.T. designed Experiment 2. J.J.E. and C.L.B. ran Experiment 2. J.J.E and C.L.B. implemented the models and performed the analyses of Experiment 2. C.L.B. and J.B.T. wrote the manuscript.

Competing Interests. The authors declare that they have no competing interests.

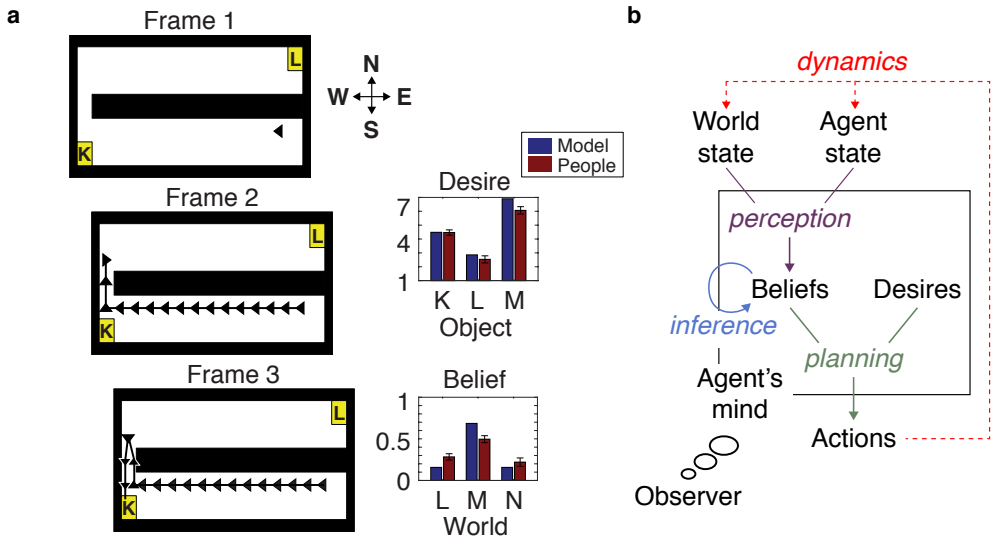

Figure 1. **a** "Food-trucks" scenario, using animated two-dimensional displays of an agent navigating through simple grid-worlds. The agent is marked by a triangle, three trucks are marked by letters (Korean (K), Lebanese (L), and Mexican (M)), parking spaces are marked by yellow regions, and buildings (which block movement and line of sight visibility) are marked by black rectangles. Frames 1-3 show several points along a possible path the agent could take, on a day when the K and L trucks are present but the M truck has not come to campus: The agent leaves his office where he can see the K truck (Frame 1), but walks past it to the other side of the building where he sees the L truck parked (Frame 2); he then turns around and goes back to the K truck (Frame 3). Which is his favorite truck? And which truck did he believe was parked on the other side of the building? Red bar plots show mean human judgments about these desires and beliefs, with standard error bars after viewing the agent's path up to frame 3. Desire ratings were given for each food truck (K, L, M), and belief ratings were given for the agent's initial belief about the occupant of the far parking spot (L, M, or nothing (N)). In this scenario, participants (n=16) judged that the agent most desired the M truck, and (falsely) believed it was probably present in the far spot. Our Bayesian Theory of Mind (BToM) model (blue bars) predicts these judgments and analogous ones for many other scenarios (see Figs. 2, 4). **b** Folk-psychological schema for Theory of Mind. BToM formalizes this schema as a generative model for action based on solving a partially observable Markov decision process, and formalizes mentalizing as Bayesian inference about unobserved variables (beliefs, desires, percepts) in this generative model, conditioned on observed actions.

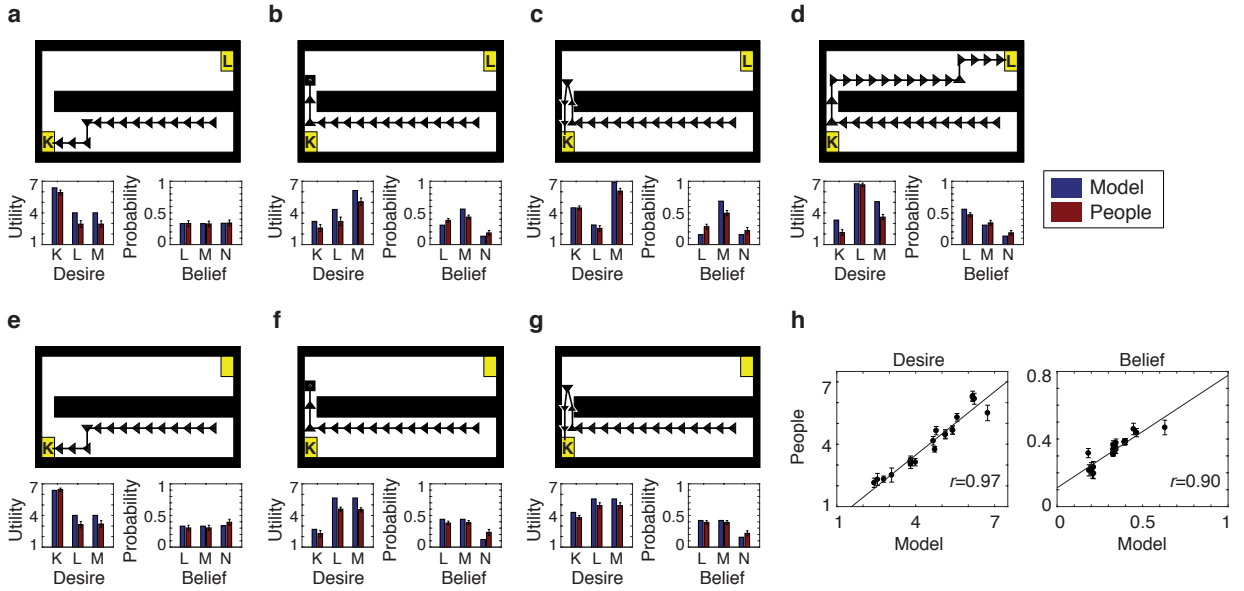

Figure 2. Exp. 1 results. Plots and correlations (Pearson r, n=21) use the best-fitting model parameterization ($\beta = 2.5$). **a-g** Comparing BToM and mean human (n=16) desire and belief inferences from seven key scenario types (error bars show s.e.m). a-d show scenarios with two trucks present; e-g, only one truck present. a,e: Agent goes straight to the nearest truck. b,f: Agent's incomplete path heads behind the building to check the far spot. c,g: Our introductory scenario: agent returns to the near truck after checking the far spot. BToM and humans attribute a desire for the missing truck, and an initial false belief that it was present. d: Agent goes to the far truck after checking the far spot. See SI Appendix: Experiment 1 for results from all scenarios and alternative models. **h** Comparing BToM and mean human (n=16) desire and belief inferences across seven scenario types (error bars show s.d. of within-trial mean judgment).

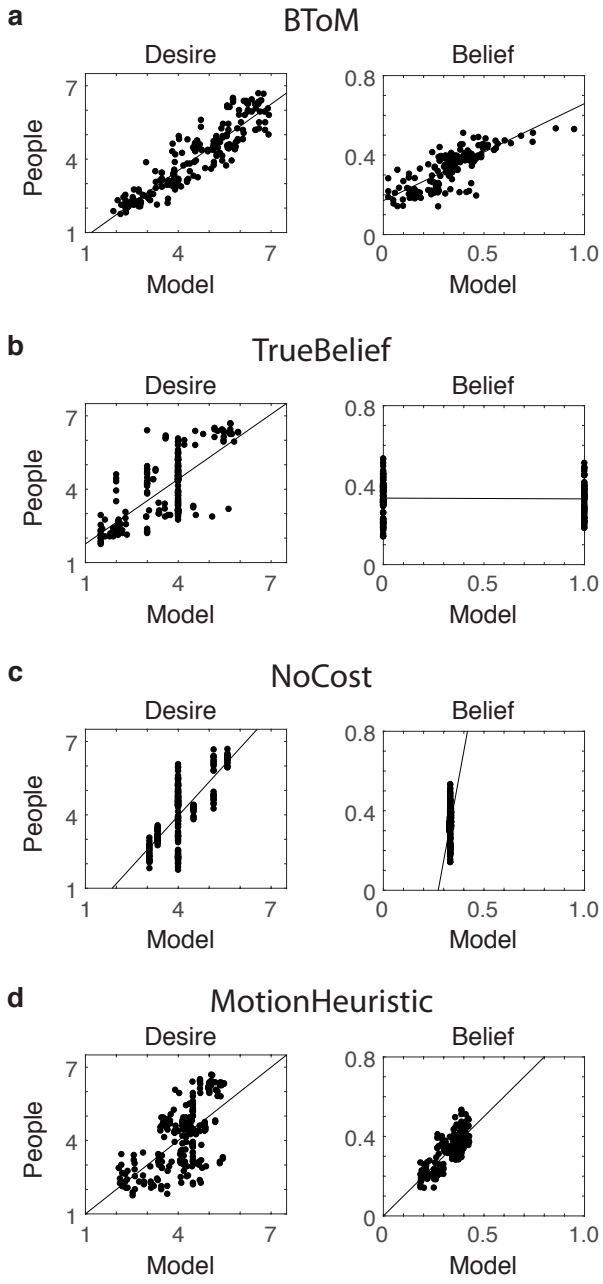

Figure 3. Comparing BToM and mean human (n=16) desire and belief inferences across all individual scenarios. Correlations (Pearson r, n=219) use the best-fitting model parameterization. **a** BToM versus people, with $\beta = 2.5$. **b** TrueBelief versus people, with $\beta = 9.0$. **c** NoCost versus people, with $\beta = 2.5$. **d** MotionHeuristic versus people.

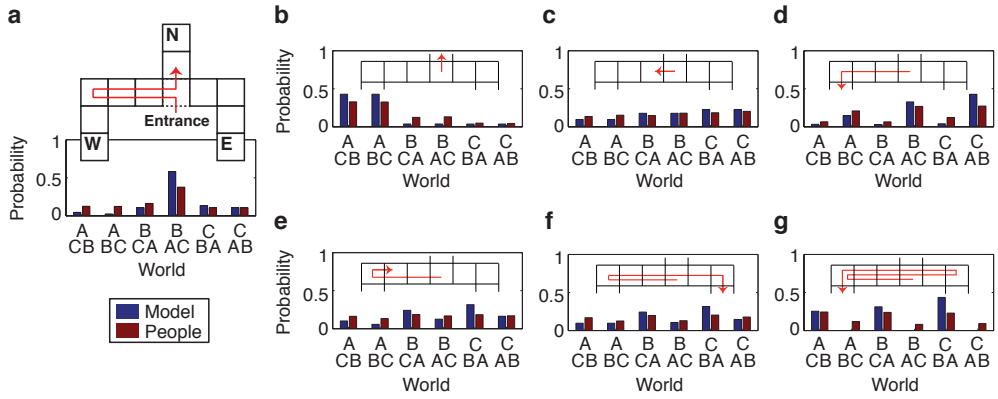

Figure 4. Exp. 2 results. **a-g** Comparing BToM and mean human (n=176) percept inferences on a range of key scenarios. Model predictions use best-fitting parameterization ($\beta = 1.5$).

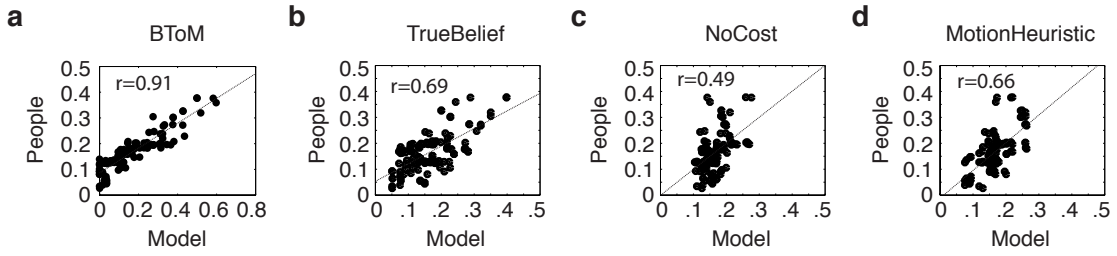

Figure 5. Exp. 2 results, comparing models and mean human (n=176) percept inferences across all individual scenarios. Correlations (Pearson r, n=114) use best-fitting model parameterization. **a** BToM versus people, with $\beta = 1.5$. **b** TrueBelief versus people, with $\beta = 0.25$. **c** NoCost versus people, with $\beta = 5.0$. **d** MotionHeuristic versus people.

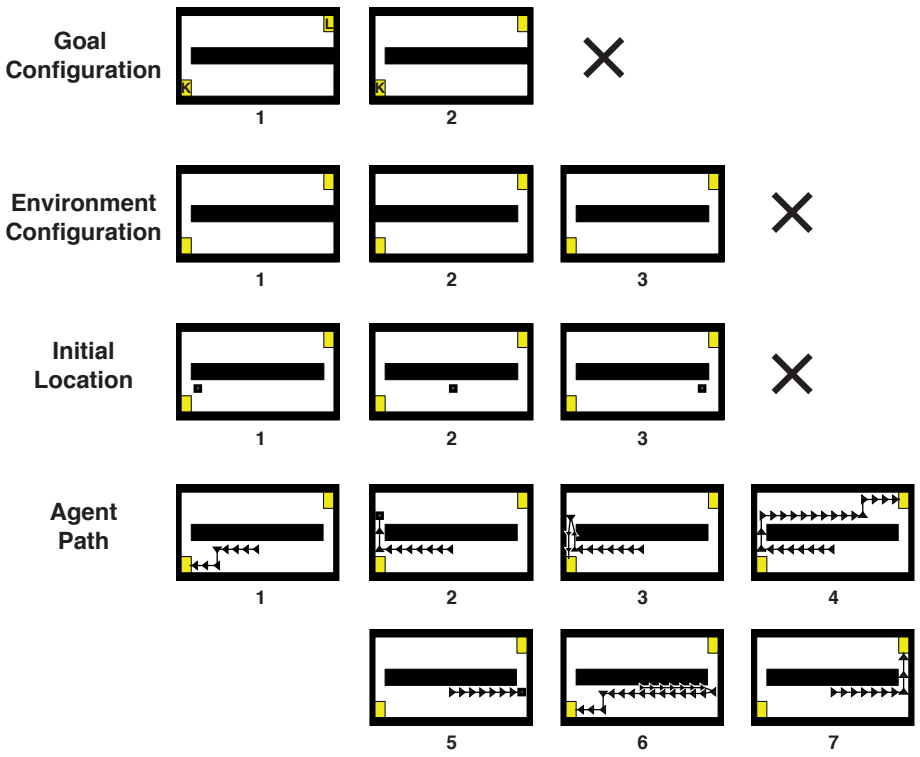

Figure 6. The factorial design of Experiment 1 varied four factors: (1) goal configuration, (2) environment configuration, (3) initial agent location, and (4) agent path.

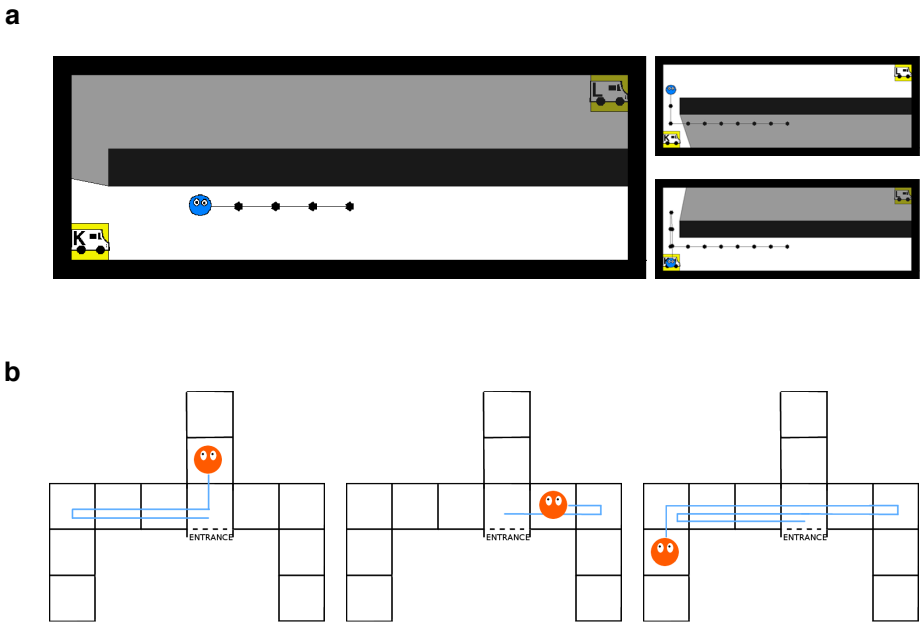

Figure 7. Example experimental stimuli. **a** Three frames from an example Experiment 1 scenario. **b** Three example scenarios from Experiment 2.