People can infer the magnitude of other people's knowledge even when they cannot infer its contents.

Rosie Aboody*, Isaac Davis*, Yarrow Dunham, and Julian Jara-Ettinger

Yale University

Author Note

Data availability: All stimuli, data, model, and analysis code is available at: https://osf.io/pjy6x.

Competing interests: The authors declare no competing interests.

* These authors made equal contributions.

Abstract

Inferences about other people's knowledge and beliefs are central to social interaction. However, it is often not possible to tell what exactly other people know, because their behavior is consistent with a range of potential epistemic states. Nonetheless, in many of these situations we often have coarse intuitions about how much someone knows, despite being unable to pinpoint the exact content of their knowledge. Here we sought to explore this capacity in humans, by comparing their performance to a normative model capturing this kind of broad epistemic-state inference, centered on the expectation that agents maximize epistemic utilities. We evaluate our model in a graded inference task where people had to infer how much an agent knew based on the actions they chose (Experiment 1), and joint inferences about how much someone knew and how much they believed they could learn (Experiment 2). Critically, the agent's knowledge was always under-determined by their behavior, but the behavior nonetheless contained information about how much knowledge they possessed or believed they could gain. Our model captures nuanced patterns in participant judgments, revealing that people have a quantitative capacity to infer amorphous knowledge from minimal behavioral evidence.

Keywords: Computational Modeling Social Cognition Theory of Mind

People can infer the magnitude of other people's knowledge even when they cannot infer its contents.

Introduction

Imagine going to your friend's house for dinner and, as you're cooking together, realizing that you'll need more flour. As the two of you head out, you notice that your friend immediately starts walking in the direction of a large supermarket, rather than her usual go-to bodega around the corner. From this simple decision you might quickly suspect that she knows something you don't. Perhaps the bodega doesn't carry flour; maybe it's cash only and your friend intends to use her credit card; or the supermarket might be the only place that's open late. Inferences like these not only enable us to make sense of others' behavior, but also help us decide when to share what we know, and from whom to learn what we don't, forming a cornerstone of complex social action.

The ability to interpret other people's behavior in terms of mental states, called *Theory of Mind*, has its origins in early childhood. From infancy, we interpret other people's behavior as goal-directed (Woodward, 1998) and infer others' goals and preferences by assuming that agents act to maximize utilities—the difference between the costs they incur and the rewards they obtain (Csibra, 2003; Jara-Ettinger et al., 2016; Liu et al., 2017). Throughout our life, this expectation enables us to make a variety of judgments, such as inferring what others like (Lucas et al., 2014; Jern et al., 2017), predicting how they might behave (Jara-Ettinger et al., 2020), and determining their social affiliations (Jern & Kemp, 2014; Ullman et al., 2009; Davis et al., 2023).

As the example above shows, however, inferences about others' minds are not restricted to goals and preferences: they also include judgments about what others may or may not know. Consistent with this, research in computational social cognition has found that people can make quantitative inferences about the contents of others' beliefs based on their behavior (Baker et al., 2017). This work showed that a computational model of joint belief-desire attribution, embedded in a Bayesian framework for action understanding, captures how people determine what an agent likely believes about their environment given their behavior (e.g., if an agent looking for lunch walks towards the end of the block, peeks around the corner to see a Mexican food truck, and then turns around, we can infer that the agent was hoping to see a different food truck there).

While this work shows that people can make quantitative targeted belief inferences, such as determining whether an agent knew the type of food a vendor might be selling based on their behavior, these inferences often require access to a relatively constrained hypothesis space and key actions that reveal the agent's beliefs. In many everyday situations, however, there may be a wide range of different belief states compatible with the behavior we observe, making it difficult or impossible to infer the specific contents of someone's beliefs. In cases like these, our representations of other people's epistemic states appear to consist of amorphous estimates of how much others know, without being sure exactly what it is that they know. Returning to the example in the introduction, when your friend chose to go to the supermarket, it is easy to infer that she knows more than you do, even though we might not know exactly what she knows. In cases like these, where the exact contents of an agent's beliefs are underdetermined by their behavior, can people make inferences about how much an agent knows in a precise and quantitative manner (to the degree revealed in the data)? Or are these inferences coarse and qualitative, providing no more than unreliable hints about others' knowledge?

Research investigating people's ability to quantify others' knowledge—i.e., inferences about how much people know without knowing the exact epistemic content—has generally focused on children. By early in preschool children can represent how much others know about a domain, without needing to list the full contents of their knowledge (Landrum & Mills, 2015; Lutz & Keil, 2002). However, to our knowledge, no work has explored our capacity to infer knowledge magnitude from others' actions, or specified the computations that might underlie such inferences.

Here we propose that such inferences are part of our broader quantitative inferential system within Theory of Mind, and therefore supported by an expectation that agents maximize utilities. Specifically, given the expectation that agents choose actions which (they believe) fulfill their goals as efficiently as possible, an agent's choice of action can reveal what that agent believes to be efficient, which can, in turn, provide indirect evidence about how much knowledge they possess. Thus, we propose that adults can infer how much an agent knows based on the subjective costs that they appear to act under (and we explain this in detail in our computational framework). In the example above, for instance, the fact that your friend bypassed a potential low-cost option (going to the bodega), and chose to immediately incur a seemingly higher cost (walking to a place that was farther away) for the same reward (getting flour), suggests that she possessed privileged information—leading her to conclude that the large supermarket was a better option than you'd originally assumed.

In this paper we present a computational model of epistemic quantification through an expectation that agents maximize utilities, and we test its performance on tasks where participants must infer how much someone knows or thinks they can learn based on their behavior. Our work shows that people can seamlessly make graded quantitative estimates of how much someone knows or expects to learn, and that these inferences can be explained through an expectation that agents maximize utilities (the difference between the costs they incur and the rewards they obtain), and an understanding that the costs agents incur depend on the knowledge they possess.

Computational Framework

Our computational framework builds on a recent family of computational models of mental-state inference structured around an expectation that agents act rationally—formalized as a generative model of utility maximization, combined with a mechanism for inverting this causal model via Bayesian inference (Lucas et al., 2014; Jern et al., 2017; Baker et al., 2017; Jara-Ettinger et al., 2020). We extend this framework by proposing that adults often expect agents' costs to be mediated by their knowledge—and can thus infer others' epistemic states from observing the apparent costs they choose to incur.

For simplicity, we will explain our framework within the context of our Experiment 1 paradigm. In these scenarios, an agent must choose one of two different fields for an Easter egg hunt. Each field contains a different spatial and numerical configuration of eggs (see Fig 1), and exactly one egg in each field contains a prize, while all other eggs are empty. Suppose that the agent arrived while the fields were being set up, and was able to see the contents of some of the eggs in each field (either empty or full). Let k_1 denote the subset of eggs in field 1 that the agent observed, and similarly for k_2.

Given this knowledge, the agent can compute the expected cost of finding the prize in each field, which we assume is equal to the expected distance traveled before finding the prize, plus a small fixed cost C of opening each egg to check its contents. If the agent's knowledge for a field includes the egg e_i that contains the prize, then the cost of finding the prize in that field is simply the distance $\text{dist}(e_0, e_i)$ from the entrance e_0 to the target egg e_i, plus the cost C of opening the egg. Now suppose that the agent's knowledge specifies that eggs $k = \{e_1, \ldots, e_k\}$ are empty, and that the prize must be in one of the remaining eggs $k^c = \{e_{k+1}, \ldots, e_n\}$. Let π be a path that starts at the entrance and passes through each egg in k^c, and let π_i denote the ith stop of π, so that π_0 is the entrance to the field, and π_i is the ith egg on the path. The cost of traversing the entire path, stopping to check each egg, is

$$\text{cost}(\pi) = \sum_{i=1}^{|k^c|} \text{dist}(\pi_i, \pi_{i+1}) + C \tag{1}$$

where $\text{dist}(a, b)$ is the distance from point a to point b. Most of the time, however, the agent will not have to traverse the full path, as they can stop once they find the egg containing the prize. Assuming that each egg has equal probability of containing the prize, such that $P(\text{prize in egg } i) = 1/|k^c|$ for all i, then the expected cost of finding the prize along path π is equal to

$$E[\text{cost}(\pi)] = \sum_{i=1}^{|k^c|} \frac{1}{|k^c|} * \text{cost}(\pi|_i) \tag{2}$$

Here, $\pi|_i$ is the sub-path obtained by following π until the ith egg, then stopping. Thus, the expected cost of finding the prize along path π is equal to the sum of the costs of traversing each sub-path $\pi|_i$, weighted by the probability that the prize is in the ith egg along path π.

Given that people expect each other to act rationally and efficiently, we assume that the agent will compute the search path that minimizes the expected cost of finding the prize in field X, which we refer to as $E[\text{cost}(X)|k]$. If the reward of getting the prize is equal to R, then the total expected utility of an agent with knowledge state k choosing field X is equal to $U_X = R - E[\text{cost}(X)|k]$.

Now suppose that the agent computes the expected utility for each field, U_1 and U_2. We assume that agents will generally try to maximize their expected utilities, but are not deterministic and may be prone to errors (e.g.: due to distraction or errors while computing expected costs). Thus, rather than assuming the agent will always choose the field with higher expected utility, we make a standard assumption that the agent will choose a field with probability

$$P(\text{choice} = \text{field}_i|k) \propto e^{U_i/\tau} \tag{3}$$

This is the standard softMax function, which takes a vector of real numbers (in this case, the expected utilities) and converts it into a probability vector. The "temperature" parameter τ controls the agent's level of rationality: a very high value of τ entails nearly uniform behavior (i.e.: choosing each option with equal probability), while very low values entail nearly deterministic behavior (i.e.: choosing the highest utility option with probability near 1). Thus, equation 3 specifies the probability that an agent with knowledge states k_1, k_2 (about fields 1 and 2, respectively) will choose to enter each field.

Given this generative model of the agent's behavior, a Bayesian observer can infer the agent's knowledge of each field k_1, k_2 based on the field configurations and the agents choice according to Bayes' rule:

$$P(k_1, k_2|\text{choice}, \text{field}_1, \text{field}_2) \propto P(\text{choice}|k_1, k_2, \text{field}_1, \text{field}_2)P(k_1, k_2) \tag{4}$$

Here, $P(k_1, k_2|\text{choice}, \text{field}_1, \text{field}_2)$ is the posterior probability of the agent's knowledge states, $P(\text{choice}|k_1, k_2, \text{field}_1, \text{field}_2)$ is the likelihood of the agent's choice given these knowledge states (given by equation 3), and $P(k_1, k_2)$ is the prior probability of the agent having these knowledge states.

In our scenarios, however, the richness of the agent's possible knowledge states (all possible subsets of eggs in each field) and the coarseness of the agent's behavior (a binary choice between two fields) make the exact contents of the agent's knowledge highly underdetermined by the observed behavior. That is, there will always be a large number of possible knowledge states compatible with the agent's choice. But even when we can't infer the precise contents of others' knowledge representations, we may still be able to infer approximately how much they know (getting a rough sense of how knowledgeable they are). Thus, given a posterior distribution over what the agent might know (equation 4), we formalize the quantity of amorphous knowledge Q as the expected quantity of knowledge encoded in the probable epistemic states that the agent has, given by Equation (5) below.

$$Q = \sum_{k \in K} |k| p(k|\text{choice}) \tag{5}$$

where K is the set of all possible epistemic states, $|k|$ is a quantification of how much the agent knows in that state, and $p(k|\text{choice})$ is the posterior probability of that knowledge state (Eq. 4). Naturally, precisely defining the measure $|k|$ may be highly context-sensitive. Here we focus on its application in a particular experimental context but return to the idea of how this might generalize in the discussion.

We evaluate this framework in two experimental paradigms. The first paradigm tests people's capacity to infer how much someone knows about two related environments based on which one they choose to seek a reward in. The second paradigm tests people's capacity to jointly infer how much someone knows and how much they expect to learn based on whether they seek additional information before trying to attain a reward. Additional details about the inference procedure can be found in each experiment.

Experiment 1

To test our model, we designed a task where an agent's behavior (and its costs) could reveal approximately how much they knew—but was too impoverished to reveal precisely what they knew. Specifically, participants watched an agent choose which of two fields to search for a prize hidden in an easter egg, knowing that each field had only one egg with a prize inside (and that the reward was always the same in every field).

The cost of locating the prize in any given field was determined by the number of eggs, their spatial distribution, and the true location of the prize. By manipulating all three variables, we test if participants infer how much others know by quantifying and comparing their expected costs—or whether participants rely on a simpler heuristic that does not require them to track or reason about others' costs when inferring epistemic states. Our procedure, stimuli, sample size, and analysis plan for our main model were preregistered (see OSF project page).

Model Parameters

Our main model has four parameters: the reward of obtaining the prize, the cost of checking an egg's contents upon reaching it, a prior over the agent's knowledge, and the softmax parameter (τ). All parameter values and model predictions were preregistered prior to data collection.

The reward function for the prize is the same across fields, and we set it as a constant $R(a_i) = 100$. Because the reward is constant across action plans, the difference in utilities between the two plans would be unchanged by different reward functions. We simply selected (and preregistered) a reward function large enough to ensure that no action plan could have a negative utility.

For each knowledge state sample, the cost of stopping to check an egg's contents was modeled as a continuous uniform distribution $[1, 3]$. This range was chosen to capture the expectation that stopping to open an egg does incur some cost, but that this cost is relatively minimal but its precise value unknown.

We specified a prior over the agent's knowledge: the agent had a 50% chance of knowing each egg's contents. We also explicitly communicated this to participants in our task (see Procedure) to ensure that participants and the model both relied on similar epistemic priors. Finally, we selected a softmax τ value that produced graded action predictions in proportion to each plan's expected utility ($\tau = 3$).

We implemented our inference procedure via Monte Carlo sampling, drawing 10,000 knowledge states from each field. We then compute equation 5, by quantifying amount of knowledge in an epistemic state as $1 -$ the proportion of eggs the agent is still uncertain about (if the agent knows where the prize is, they know the rest of the eggs are empty, and thus the proportion known is 1; if the agent is unsure about half of the eggs, the proportion known is .5; and so on).

Alternate Model

Our main model assumes that people quantify the cost of obtaining the prize in each field under different degrees of knowledge, and then reason about the knowledge states under which the agent's actions would have been utility-maximizing. However, it is possible that adults generally do not apply such complex computations when inferring others' knowledge states, and instead rely on simpler rules or heuristics. Such heuristics could get things right most of the time, while requiring less effort to apply.

To address this possibility, our alternate model encoded the simple heuristic that agents tend to choose options they know more pieces of information about. Critically, this alternate model did not consider agents' knowledge states in a full mentalistic way: it did not compute the utility of each field based on the agent's knowledge state, and did not expect agents to navigate directly to an egg if they knew it contained the prize. It simply considered the proportion of eggs with known contents in each field, and expected the agent to always choose the field where this proportion was larger (or choose randomly when this proportion was equal across fields). We then generate predictions from this alternate model using the same sampling procedure as in the main model.

Our alternate model was not preregistered, but uses only one parameter: the same knowledge prior as in our main model. Because our alternate model encodes an expectation that agents will always choose fields they know more pieces of information about, we do not compute the utility of each field, and thus we do not need to specify agents' costs, rewards, or a softmax parameter.

Participants

40 adult participants with U.S.-based IP addresses were recruited via Amazon Mechanical Turk (M = 35.05 years, SD = 9.23). 7 additional participants were recruited but excluded from the study for failing a preregistered inclusion trial.

Stimuli

Stimuli consisted of 19 test trials, plus one inclusion trial. The test trials were presented in a randomized order, and the inclusion trial was always presented last. Each trial showed an agent, and two fields. The fields each had easter eggs placed inside, and one egg in each field contained a hidden prize. This egg was circled for participants. An arrow indicated the agent's path to their chosen field, thus showing which field the agent chose to visit on each trial (see Figure 1).

Stimuli were based on three scenarios (pairs of fields) we thought could elicit a range of model ratings. To manipulate the cost of searching each field, eggs in the first field (field A) were always wide-spread. The second field (field B) contained the same number of eggs, but these eggs were instead clustered near the middle of the field. The first scenario is shown in Figure 1a. The second scenario was based on the first: we selected a subset of 6 eggs from each field, thus varying the number of eggs but not their position. The third scenario was in turn based on the second, but here we instead varied the position of the eggs in field A (capturing a case where most of the eggs in field A were extremely costly; see Figure 1b).

To select the final locations of the prize in field A, we provided each scenario as input to the model, but systematically varied which egg in field A contained the prize, yielding 42 trials (21 unique scenarios x 2 choices per scenario).[1] We selected 24 trials (12 unique scenarios x 2 choices per scenario) that both produced a range of model responses, and were not too similar to each other. In preparation to present stimuli to participants, some trials were mirrored, and we slightly varied the position of the prize in field B amongst similar scenarios (to prevent participants from noticing similarities between trials).[2] We then obtained final model predictions, and excluded any trials where the model's knowledge predictions were based on less than 500 samples (that is, where the predicted choice of field was consistent with the observed choice in less than 5% of cases). This yielded 19 final trials; this criterion and our final set of stimuli was preregistered.

Procedure

Participants were introduced to an agent going on easter-egg hunts in a two-dimensional grid-world. Participants learned that a farmer had placed easter eggs in his fields, hiding a prize inside one egg in every field. This prize (one silver token) was always the same in every field, and the prize egg was always circled for participants.

Participants learned that because the grass in the fields was quite short, the agent could always see where the eggs were located in a field before entering it. But while the

[1] We did not expect the location of the prize in field B to strongly affect the model's predictions; to test if this was the case, we did also replicate one scenario given a different prize location in field B, yielding an additional 18 additional trials. The location of the prize in field B indeed had little effect (as all of these eggs are so close to each other), and thus we selected our final stimuli by considering primarily the location of the prize in field A.

[2] Despite slightly varying the prize's location in field B across similar trials in our preregistered stimuli, our model predictions were accidentally not updated accordingly prior to preregistration. Because we collected our data using the preregistered stimuli, we obtained new model predictions for any trials where the location of the prize in the stimuli did not match the coordinates originally used in the preregistered model predictions. No aspect of the model itself was modified and we used the same preregistered parameters.

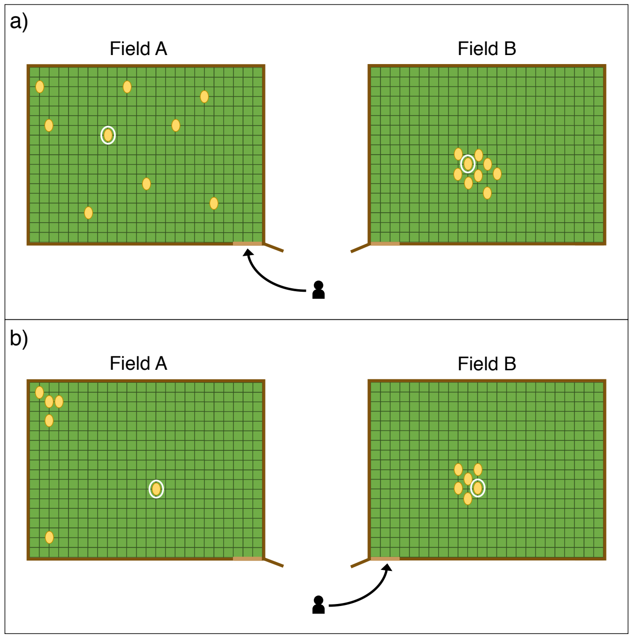

Figure 1

Example of the experimental stimuli. The arrow indicates the agent's chosen field; eggs containing a prize are circled. Panel A depicts a strong epistemic contrast: here, you might infer that the agent knows approximately where the prize is located in their chosen field, and very little about the other field. Panel B depicts a more graded contrast: here, you might suspect that the agent knows more about the prize's location in their chosen field, but may be less certain they know a lot (because their chosen field is also much less costly to search).

prize egg was circled for participants, the agent didn't necessarily know which egg contained the prize. Participants learned that the agent had seen the farmer set up some of the eggs; it was unclear what prior over knowledge participants would bring to the task, so we specified that the agent had a 50/50 chance of knowing the contents of any given egg, and that these probabilities were independent (i.e.: knowing the contents of one egg does not affect the probability of knowing the contents of any other egg). Additionally, participants were explicitly instructed that the agent did not always know the same amount about every field; the amount she knew about the location of the prize in each field could differ.

Participants learned that the agent always had to choose between two fields, and could only search the field she chose. An arrow indicated which field the agent had chosen to search (see Figure 1). Participants were oriented to factors that might affect the agent's search decision: they were told that the agent always wanted to find the prize as quickly and easily as possible, and that the difficulty of finding the prize was determined by the number of eggs in a field, their distance from the entrance, and the amount the agent already knew about the location of the prize. Note that while this tutorial ensured participants were attentive to the main features of our task, we are interested in how participants combine these different pieces of information and reason over them to infer what others know. The tutorial did not specify how participants should weight or use any of these features in their judgments.

To access the task, participants then completed a preregistered inclusion quiz that assessed their understanding of the task instructions. Participants were given two chances to pass the inclusion quiz; those who failed on their first attempt were required to review the task introduction before trying again. Participants who failed both attempts were not given access to the task. Upon passing the inclusion quiz, participants then completed the 19 test trials (presented in a randomized order), plus one inclusion trial at the end. For each trial, participants were asked to rate, on a sliding scale from 0 - 100, how much the agent knew about the location of the prize in each field. Critically, participants rated how much the agent knew about both fields, not just the field she had chosen. The preregistered inclusion trial always came last. It was similar to the test trials, but presented an extreme contrast where we could make a strong prediction about the pattern of judgments an attentive participant should make. Participants whose judgments differed from our preregistered criteria were excluded. Finally, participants were asked what they thought the point of the task had been, and were given an opportunity to provide feedback or note any technical difficulties.

Results

Participants rated the agent's knowledge about both fields in 19 test trials, yielding 38 ratings. As preregistered, participant responses were averaged by question, and then z-scored; the corresponding model predictions were also z-scored.

Figure 2 shows the overall results, revealing that our model was highly correlated

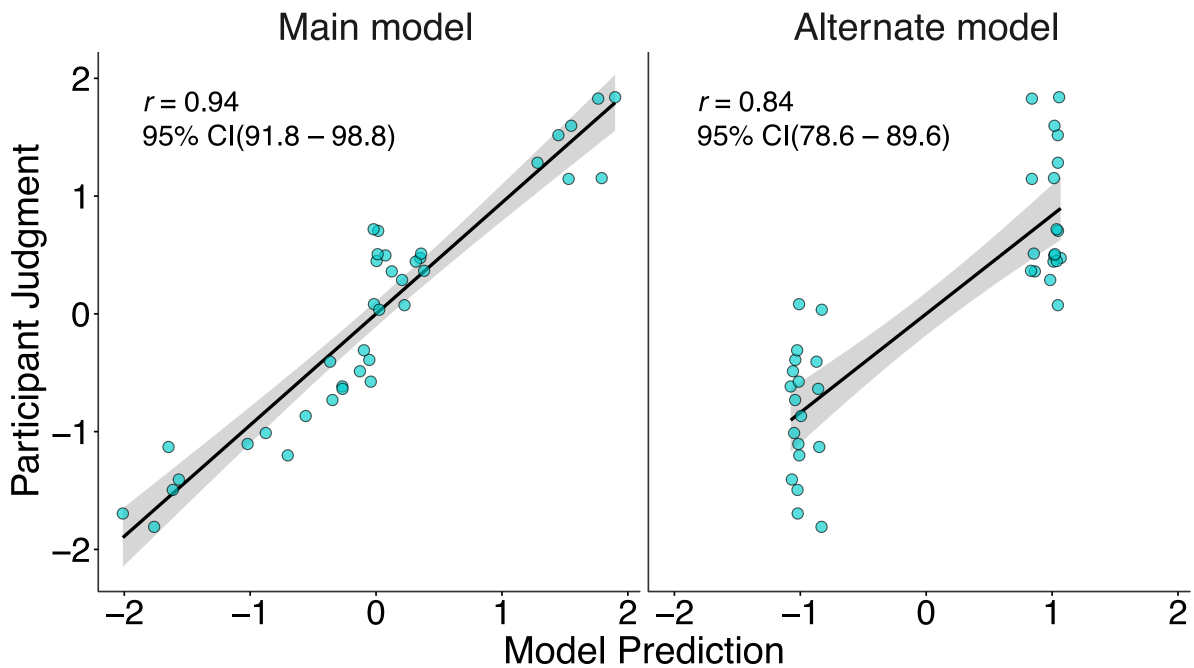

Figure 2

Comparison between our model and the alternate model, with linear regressions fit to each dataset. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression.

with participant judgments, $r = 0.94$ (95% CI: $91.8, 98.8$). Critically, this correlation did not only reflect extreme cases where both the model and participants inferred a lot of knowledge or very little knowledge: it also included cases where both the model and participants were equally uncertain, in a graded manner, about how much the agent knew. Figure 3 plots the trial-by-trial correspondence between model and participant ratings, showing that participants' judgments were not bi-modal, but rather graded in a way that closely tracked our model's predictions.

To ensure that these results could not be the product of a simple heuristic, we implemented an alternate model. Rather than performing full mental-state inference, our alternate model simply assumed that agents always choose fields where they know about a greater proportion of eggs. Note that we only preregistered an analysis plan for our main model, but test the performance of the alternate model in the same way. The alternate model showed a weaker correlation with participant judgments, $r = 0.84$ (95% CI: $78.5, 89.5$), demonstrating that the amount of locations an agent knows about in each field does matter, but that predictions made on the basis of this one factor (without considering costs) do not capture the graded structure of participant judgments. A bootstrap over the correlation difference revealed that the main model was reliably better correlated with participants judgments than the alternate model (correlation difference, alternate model − main model = −0.11, 95% CI: $-17.4a, -4.3$; not preregistered). As Figure 2 reveals, although the correlation between the alternate model and participant judgments was still high, this is only because the alternate model categorized every judgment into two rough bins. These predictions were approximately correct, but lack the nuance that participants' epistemic inferences showed, and that our model was able to capture.

Experiment 2

Experiment 1 shows that adults are able to make precise epistemic inferences even in underdetermined scenarios—and that these inferences are well-captured by our main model. Experiment 2 both conceptually replicates and extends these findings. Specifically, in Experiment 2 we test whether our framework can capture not just adults' inferences about how much someone knows, but also about how much they believed they could learn. To do so, we designed a task where an agent's information-seeking choice (and its cost) could reveal approximately how much they knew and believed they could learn (but again, could not reveal these states with any precision). Specifically, participants watched agents search islands for hidden treasure (Figure 4). Agents had the option to obtain a treasure map first, or to skip the map and go straight to the island. Importantly, the map was not always informative: sometimes it might contain a lot of information about the treasure's location, sometimes it might contain a little, and sometimes it might contain no information at all. To elicit graded inferences, we manipulated the distance of the map (varying information's cost), the size of the island (varying the potential difficulty of finding the treasure), and agents' information-seeking choices (varying whether or not they

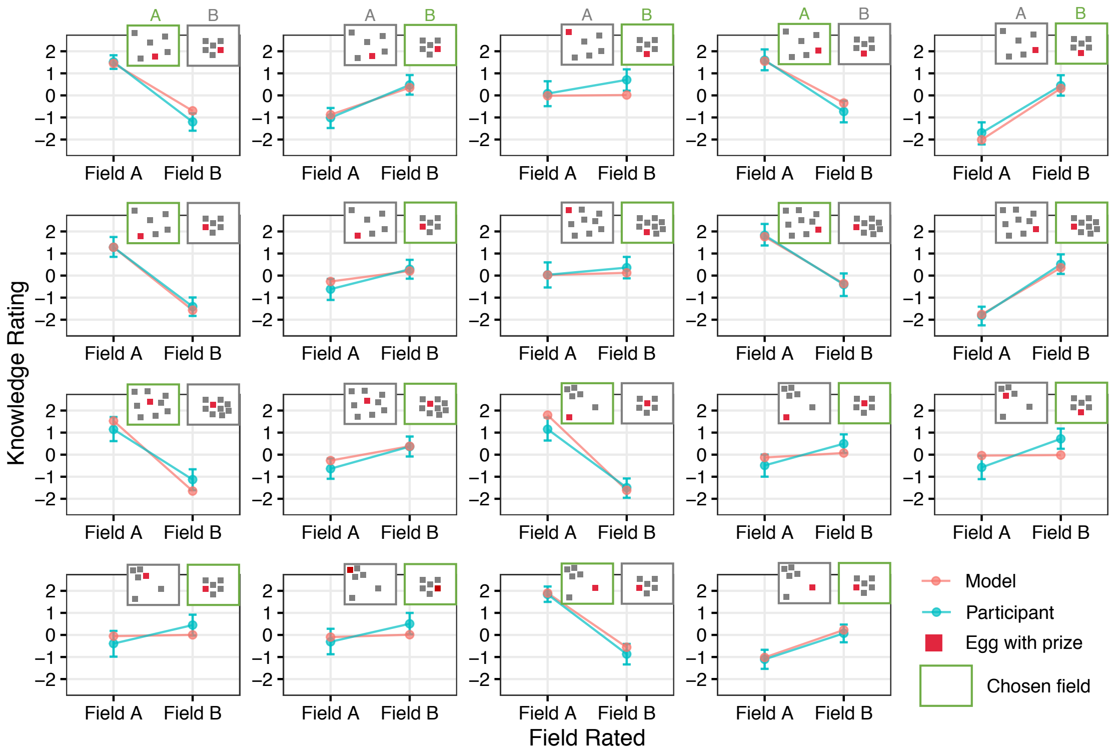

Figure 3

Detailed results for Experiment 1. Each panel presents one trial, with results split by the field rated (Field A or Field B, indicated on the x axis). The y axis indicates standardized knowledge ratings. Participant judgments are plotted in blue; model predictions are plotted in red. Vertical bars show 95% confidence intervals over participant judgments. The schematics show the position and number of eggs in each field, the egg with the prize, and the field the agent ultimately chose in each trial.

pursued the map). Our procedure and sample size were pre-registered.

Model Structure and Parameters

The computational model followed the same conceptual structure as Experiment 1. The key difference was that the two competing utilities no longer referred to two possible search areas. Instead, the first utility represented going directly to the island to search for the treasure (same logic as searching a field in Experiment 1). The second utility represented obtaining the map first. Thus, this second utility integrated the additional deterministic cost of obtaining the map and going to the island, and then computed the revised search cost after obtaining the map. Using Bayesian inference, we then applied joint inference to recover the agents' (1) amount of knowledge about the island and (2) amount of information expected to be contained in the map.

Our main model for Experiment 2 has six parameters: the reward of obtaining the prize (set to a constant $R(a_i) = 100$), the cost of sailing across one grid square, the cost of searching one island square, the softmax parameter (τ), and two priors: one over how much agents know in general, and one over how much information maps generally hold.

We pre-registered the first four parameters prior to data collection, basing the relative cost of sailing vs. searching upon empirical estimates from a pilot sample. Our pilot sample judged that searching one island square was, on average, 2.25x more difficult than sailing across one ocean square, and thus we pre-registered a sailing cost of 1, a searching cost of 2.25, and $\tau = 4$ (based upon the range of utilities these costs produced). However, we explicitly pre-registered that we would re-estimate these based on our final sample, and re-adjust our softmax parameter if needed. In our final sample, most participants judged that searching was more difficult than sailing, judging that it was on average 3.9x harder. Thus to generate our final predictions, we set the cost of searching to 3.9. Because this affected the range of possible utilities, as preregistered we adjusted our softmax parameter, setting $\tau = 6.5$.[3]

We also defined a uniform prior over the probability that the map might contain each degree of knowledge, and defined a non-uniform prior over the probability that the pirates might have each degree of knowledge (not preregistered). This was intended to capture the possibility that adults might generally expect agents to be knowledgeable (and unlike in Experiment 1, we did not specify precisely how likely agents were to know the contents of each island square). We defined this prior using the binomial distribution ($p = 0.8$).

[3] Note that Experiment 2 was conducted before Experiment 1. The pre-registered procedure for Experiment 1 was simpler due to the realization that the tau parameter did not particularly matter for our predictions.

Alternate Model

Our preregistered alternate model is a linear regression, trained on participants' z-scored average ratings in our task. It predicts knowledge based on an interaction between agents' information-seeking choice (to retrieve the map / skip the map), and the type of knowledge (what agents know / what information they believe the map contains). The formula for this regression in R is: `lm(mean participant rating ~ choice*knowledge category)`.

Participants

40 adult participants with U.S.-based IP addresses were recruited via Amazon Mechanical Turk (M = 38.73 years, SD = 12.23). 9 additional participants were recruited but excluded from the study for failing a preregistered inclusion trial.

Stimuli

Stimuli consisted of 18 test trials, plus two inclusion trials. The test trials were presented in a randomized order, and the inclusion trials were always presented last. Each trial showed a pirate ship (represented by a yellow star), a treasure map (represented by a green square), and an island (represented by brown squares); see Figure 4. Each island had a beach (represented by a lighter brown square), which was the only point on the island pirates could land their ship. An arrow indicated agents' path, showing whether they chose to pursue added knowledge (obtaining the treasure map first), or whether they chose to search the island without obtaining the map (see Figure 4a).

To construct our stimuli space, we varied the size of the island pirates needed to search (12, 24, or 36 grid-squares), the detour required to obtain the treasure map (adding approximately 10, 20, or 40 grid-squares to the journey), and agents' choices to obtain or skip the map. This yielded 18 test trials which systematically varied information's cost (as well as agents' information-seeking choices).

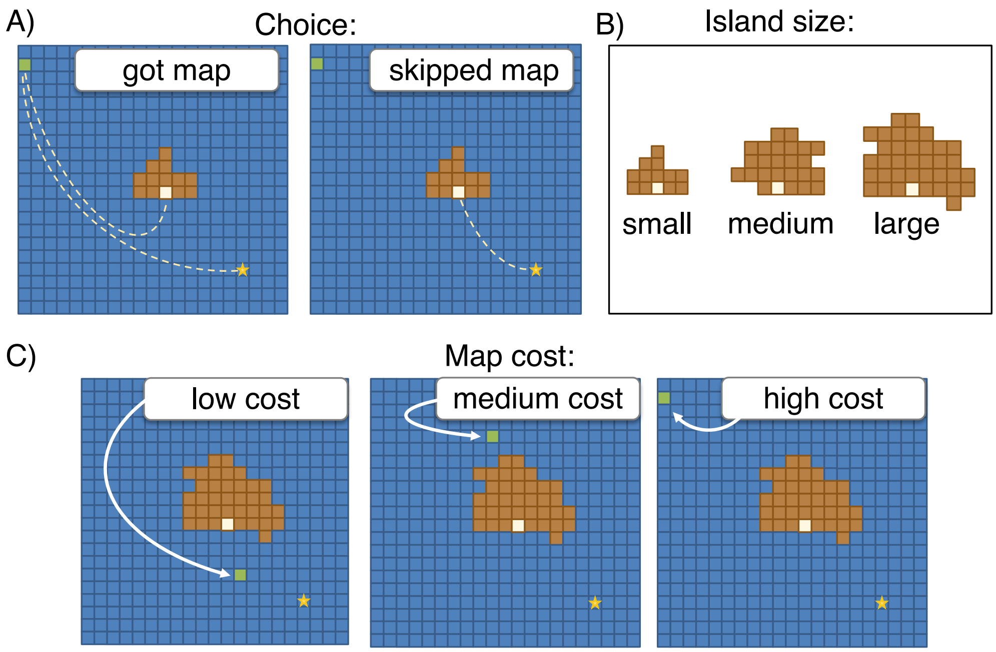

Figure 4

Space of all possible experimental stimuli. We varied A) agents' choices (to pursue information or ignore it), B) the size of the island to be searched (small, medium, large), and C) the cost of pursuing information (small, medium, large). This yielded 18 test trials. The first choice (panel A, left) depicts a strong epistemic contrast: here, you might infer the agents knew relatively little, and believed they stood to gain a lot of information (because they chose to incur a high cost to obtain the map, even though the island was small and thus relatively easy to search). The second choice (panel A, right) depicts a more graded contrast: while the agents clearly did not think the map was worth it, it may not be entirely clear why (did they know a lot, or did they simply believe the island would be easy to search even given ignorance?)

Procedure

Participants were introduced to pirates searching for treasure in a two-dimensional grid-world. Participants were shown how to identify the pirate ship (marked by a star), and learned that pirates could only land on the island at the beach (this was intended to explain why the pirates sometimes took circuitous, high-cost paths to the island; e.g., see Figure 4a). Participants learned that pirates sometimes knew a lot about the treasure's location, sometimes knew a little, and often knew something in between.

Participants learned that islands could be all different sizes, and that there was always a map somewhere in the ocean, marked by a green square. However, this map was not always helpful: sometimes it contained a lot of information about the location of the treasure, sometimes it contained only a little, and often it contained something in between. To obtain the map, pirates needed to sail to the green square first, before going to the island. An arrow indicated pirates' final choice (showing their chosen path).

Participants were oriented to factors that might affect agents' information-seeking decisions: they were told that the less pirates knew, the more work it might take to locate the treasure; the bigger the island, the more work it might be to search for treasure; and the farther the map, the more time and effort might be required to obtain it. Participants were explicitly told that, in each case, the pirates needed to decide whether it was worthwhile to pursue the map. As before, note that while this tutorial ensured participants were attentive to the main features of our task, we are interested in how participants combine these different pieces of information and reason over them to infer what others know and believe they can learn. The tutorial did not specify how participants should weight or use any of these features in their judgments.

Before the task, participants completed three simple attention check questions that assessed their understanding of the task instructions. Participants were asked to identify how the pirate ship was marked (by a star), to recall the pirates' goal (find treasure), and finally were asked to identify both that the map was always on the green square, and that pirates could only get on an island via the beach (distinguishing these from three other incorrect statements). Participants were able to select as many answers as they chose to each question; however, attentive participants should have noticed that the first two questions could only have one correct answer. Any participants who selected more than one answer in response to these two questions was excluded (preregistered). Participants who answered any question incorrectly were corrected.

Finally, participants were again reminded that both the pirates' knowledge and the informativeness of the map might vary, and that in each case, pirates needed to decide whether it was worthwhile to pursue the map. For each trial, after observing pirates' information-seeking choices (and their expected costs), participants were asked to rate, on a sliding scale from 0 - 100, how much the pirates knew about the location of the treasure, and how much information the pirates thought the map had about the location of the treasure.

Two inclusion trials always came last. These were similar to the test trials, but presented an extreme contrast where we could make a strong prediction about the pattern of judgments an attentive participant should make. Participants whose judgments differed from this pattern were excluded, as preregistered.

Participants were also asked to judge which was more difficult: to sail across one ocean square, or search one island square for treasure. After identifying which was harder, participants were asked to judge how much more difficult their chosen option was, in relation to the other. This choice was preregistered, with the idea that the cost our model assigned to each action (sailing vs. searching) would be scaled based upon participants' judgments. Finally, participants were asked what they thought the point of the task had been, and were given an opportunity to provide feedback or note any technical difficulties.

Results

Participants rated how much the pirates knew, and how much they believed they could learn from the map, in 18 test trials. This yielded 36 final ratings. As in Experiment 1, participant responses were averaged by question, and then z-scored; the corresponding model predictions were also z-scored. [4]

Figure 5 shows the overall results, revealing that our model was highly correlated with participant judgments, $r = 0.86$ (95% CI: $81, 92.9$). And this correlation did not reflect only cases where both the model and participants inferred a lot of knowledge or very little knowledge. Critically, it included cases where both the model and participants were equally uncertain, in a graded manner, about how much the agent knew.

[4] We mistakenly preregistered a slightly different z-scoring procedure—z-scoring participant ratings and then averaging by trial and prediction type. For consistency, we follow the process outlined in Experiment 1.

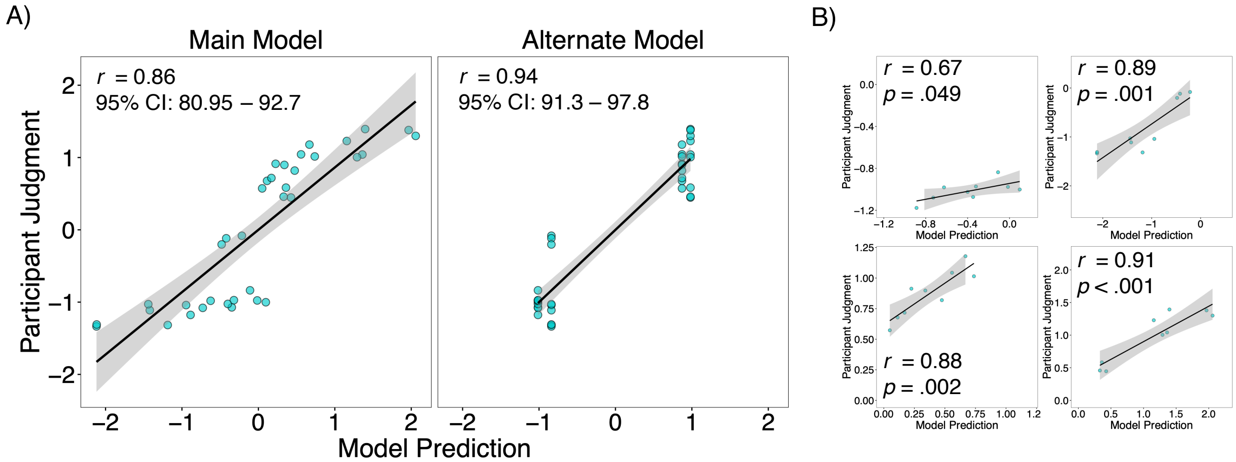

Figure 5

A) Comparison between our model and the alternate model, with linear regressions fit to each dataset. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression. B) Correlation between participant judgments and main model, binning participant judgments according to the alternate model's predictions. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression. This reveals meaningful variation our alternate model was not able to capture.

To ensure that these results could not be the product of a simple heuristic, we implemented an alternate model. Rather than performing full mental-state inference, our alternate model simply assumed that an agent who skipped the map didn't need information, and vice versa. Because this model was insensitive to cost, it did not consider more graded cases we expected humans might (e.g., that if the map is right on the way you might check even if you're not sure how much you'll learn; whereas if the map is far away, you may choose not to obtain it even if you lack some knowledge). This alternate model showed a stronger correlation with participant judgments, $r = 0.94$ (95% CI: $91.4, 97.7$); a bootstrap over the correlation difference revealed that the alternate model was reliably better correlated with participants judgments than the main model (correlation difference, alternate model $-$ main model $= 0.079$, 95% CI: $0.9, 14.6$; not preregistered).

Although the alternate model was better correlated with participant judgments (perhaps not unexpectedly, as it was trained on participant judgments in the first place), it did not capture any of their gradedness. While it is generally true that in our task, agents sought out information when they needed it and skipped it when they did not, both participants and our main model were able to make much more nuanced epistemic inferences. Thus, following our preregistered analysis plan, we test whether there is actually meaningful variation in participant judgments that the alternate model fails to capture (despite well-capturing the overall trajectory of participants' responses).

Specifically, because the alternate model binned all predictions into four categories, we tested whether participant judgments *within* each of these categories were still well-correlated with those of our main model. If this is the case, this would suggest that the alternate model fails to account for meaningful variation. In other words, obtaining meaningful correlations within each bin suggests that there is still structure in each category that only our main model is able to capture. Consistent with this possibility, even when separating participant judgments according to the predictions of our alternate model, participants' judgments were significantly correlated with the corresponding judgments from our main model (all r's between $[0.67, 0.91]$, all p's $< .05$; see Figure 5). This demonstrates that our alternate model fails to capture meaningful variation in participant judgments, despite the high overall correlation between participant judgments and the predictions of our alternate model.

General Discussion

Here we presented two experiments and a computational model designed to test people's capacity to make amorphous epistemic inferences: quantitative estimates about how much someone knows or expects to learn, but without internal representations of the contents of this knowledge. We found that people can make quantitative inferences about how much someone knows (Experiment 1), and joint inferences about how much someone knows and how much they expect to learn (Experiment 2), all from minimal observable choices. These inferences were predicted by a normative model that estimates amount of knowledge via Bayesian inference, but could not be explained by alternate models that did not consider how knowledge would affect agents' expected costs (and thus their behavior); these alternate models failed to capture the graded structure of participants' judgments.

Our computational model followed the same principles that shape related models of Theory of Mind, where mental-state inference is structured around an assumption that agents act to maximize utilities—the difference between the costs that agents incur and the rewards they obtain Jara-Ettinger et al. (2016); Gergely & Csibra (2003); Lucas et al. (2014); Jern et al. (2017). Our model builds on these ideas, and extends them by explicitly modeling the idea that, by observing the apparent costs agents incur, we can recover the amount of knowledge they possess. The quantitative fit between our model and participants suggests that the mechanisms supporting inferences about specific epistemic states follow the same principles as the mechanisms supporting inferences about broad epistemic states.

Related work has developed computational models that explain how people infer each other's beliefs about the world (Baker et al., 2017). These inferences, however, depend on access to a highly constrained set of epistemic hypotheses, and to observable behavior that is diagnostic of the agent's epistemic state. While these inferences are undoubtedly critical for social interaction, many everyday social behaviors lack the information needed to make such precise and targeted epistemic inferences. We show that, in such situations, people can nonetheless derive quantitative estimates of how much knowledge someone might possess (or believe they can come to possess). This capacity might be particularly important in informal pedagogy, as it might help us identify agents who are knowledgeable, who we could subsequently seek out to learn from. These inferences, given that they require fewer observations, might also serve as a powerful attention cue. Imagine, for instance, being a competitor in a setting like Experiment 1. Quickly detecting that an agent is knowledgeable might prompt us to attend to them carefully as they take additional actions, so that we can further uncover what specific knowledge they have.

Following a large tradition in computational cognitive science, our model was designed to explain human behavior at a computational level of analysis (Marr, 1982).

Models at the computational level typically remain agnostic about the underlying algorithmic implementation in the human mind. In our case, however, we believe there are strong reasons to suspect our model is not a plausible candidate for an algorithmic implementation. This is because our model makes two critical assumptions: First, observers must have access to a range of epistemic hypotheses that they can evaluate; second, they must have a way to quantify the amount of knowledge contained within each epistemic hypothesis.

While the first assumption may seem plausible in some situations, there are many cases where we cannot represent the internal structure of epistemic hypotheses, or have access to the hypothesis space. For instance, while we know that pilots can fly planes, most of us do not know how to represent what a pilot knows (unlike in our experiments, where we knew how to represent different possible knowledge states the agents might have). This suggests that some amorphous inferences cannot be supported by an algorithm that requires people to integrate many specific hypotheses about an agent's knowledge. Similarly, the second assumption (that it is possible to quantify the amount of knowledge in each hypothesis) was easy to formalize in our experimental contexts. But this is not always the case. In the same example about pilots, even when we build specific representations of knowledge, such as "the pilot knows how turn the autopilot on and off", it is difficult to gauge the amount of knowledge involved without having the knowledge ourselves. For instance, if it just a simple button press, little knowledge is needed. But if using autopilot requires managing a wide range of other parameters, then a lot of knowledge is needed.

The fact that our model is an unlikely candidate for an algorithmic implementation makes people's results, in some sense, even more interesting. Somehow, participants in our task were able to generate estimates of knowledge that quantitatively resembled our normative model. This suggests that people have access to some approximations that manage to produce inferences that approximate normative inferences. Thus, our results are best thought as establishing that people have a capacity to make quantitative and graded amorphous inferences, and opens questions for future research about how exactly people accomplish this.

Our results also leave an empirical question open: although our focus was on amorphous knowledge inferences, we do not know if people also spontaneously attempted to make specific epistemic inferences too. Although it is impossible to infer exactly what the agent knew, some context might reveal partial information. For instance, in Experiment 2, if a ship bypasses the island port and travels far away to collect a map, people might think that the pirates were confident that the treasure would not be close to the port. This suggests that people's inferences also be studied as a hierarchical two-tiered inference where we use observable action to simultaneously make broad epistemic inferences and specific targeted inferences when possible.

Overall, our work sheds light on a common everyday epistemic inference: the ability to infer how much others know or believe they can learn, even when there is insufficient information to infer the exact contents of their knowledge. This work highlights a space of inferences that have been historically understudied in Theory of Mind, but that might be equally important. The capacity to build quick, high-level snapshots of what's in other minds might be one of the most important representations that direct our decisions over whom to attend to, seek information from, and trust.

Acknowledgments

We thank the members of the Yale Computational Social Cognition Lab for helpful conversations and advice, and we thank Emily Gerdin for her assistance with rubber duck debugging. This work was supported by NSF award BSC-2045778, and also by Yale's Franke Program in Science and the Humanities, via a Franke Interdisciplinary Graduate Award to RA.

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*, 1–10.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*, 447–458.

Davis, I., Carlson, R., Dunham, Y., & Jara-Ettinger, J. (2023). Identifying social partners through indirect prosociality: a computational account. *Cognition*, *240*, 105580.

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

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

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

Jern, A., & Kemp, C. (2014). Reasoning about social choices and social relationships. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, .

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

Landrum, A. R., & Mills, C. M. (2015). Developing expectations regarding the boundaries of expertise. *Cognition*, .

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

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

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*, 1073–1084.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.*. MIT Press.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*.

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