Social inferences from physical evidence via Bayesian event reconstruction

Michael Lopez-Brau, Joseph Kwon, Julian Jara-Ettinger

Department of Psychology, Yale University

Abstract

Human Theory of Mind is typically associated with the ability to infer mental states from observed behavior. In many cases, however, people can also infer the mental states of agents whose behavior they cannot see, based on the physical evidence left behind. We hypothesized that this capacity is supported by a form of mental event reconstruction. Under this account, observers derive social inferences by reconstructing the agents' behavior, based on the physical evidence that revealed their presence. We present a computational model of this idea, embedded in a Bayesian framework for action understanding, and show that its predictions match human inferences with high quantitative accuracy. Our results shed light on how people infer others' mental states from indirect physical evidence and on people's ability to extract social information from the physical world.

Key words: Computational modeling, Event reconstruction, Social cognition, Theory of Mind

1. Introduction

As social animals, humans possess a specialized cognitive system to process, understand, and predict each other's behavior, known as a *Theory of Mind* (Gopnik et al., 1997; Wellman, 2014). Theoretical and empirical work suggests that human Theory of Mind is instantiated as a mental model that specifies the causal relation between other people's unobservable mental states and their observable actions. That is, Theory of Mind captures how we expect other people's thoughts, preferences, and feelings to guide what they do. Equipped with this intuitive theory, people can infer the mental states that causally give rise to other people's observed behavior.

A rapidly growing body of work suggests that the causal model within Theory of Mind is structured around an assumption that agents act to maximize their utilities—the difference between the subjective costs they incur and the

subjective rewards they obtain—capturing the idea that we intuitively expect others to act rationally and efficiently (see Jara-Ettinger 2019 for review). Consistent with this view, computational models of mental-state inference via utility maximization reach human-level performance on simple social tasks (Baker et al., 2017; Jern et al., 2017; Jern & Kemp, 2015; Jern et al., 2011; Jara-Ettinger et al., 2020), they capture richer forms of social behavior including pedagogy (Bridgers et al., 2020; Ho et al., 2019) and moral reasoning (Ullman et al., 2009), they explain social reasoning in early childhood and infancy (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; Liu et al., 2017; Lucas et al., 2014), and they have identifiable neural correlates (Collette et al., 2017).

Despite its success, this approach implicitly posits that mental-state inference requires access to someone's observable behavior, as it is these observed actions that enable us to evaluate the plausibility of different mental states. In some cases, however, people can even infer the mental states of agents whose behavior we did not get the opportunity to see (Gosling et al., 2002, 2008). For example, imagine walking into an office building and finding a vacant receptionist desk with a chewed-up pencil, a half-filled crossword puzzle, and a cellphone. From this arrangement of objects, we can immediately infer that the receptionist might have been experiencing anxiety or restlessness (as the pencil was chewed-up), that they were likely procrastinating or had few tasks to complete at the moment (as they were working on a crossword), and that they expected to be gone only momentarily (as they chose to leave their valuable belongings unattended).

As the example above shows, human mental-state inference is not limited to an ability to extract mental states from observable actions—we can also infer mental states from physical scenes with no direct social or temporal information. How do we achieve this and how fine-grained are these inferences? Here we propose that social inferences about unobservable agents are supported by a basic form of *event reconstruction*, where, upon seeing indirect evidence of an agent's presence, we reconstruct what actions they likely took, enabling us to reason about the mental states that best explain the reconstructed behavior.

While it has long been known that the ability to infer mental states from observed actions emerges early in infancy (Gergely & Csibra, 2003; Onishi & Baillargeon, 2005; Woodward, 1998), recent studies suggest that social reasoning from physical events also emerges early in childhood. By preschool, children can estimate the difficulty associated with building different physical arrangements of objects (Gweon et al., 2017); they understand which kinds of actions leave

physical traces in the environment and which kinds of actions do not (Jacobs et al., 2021); they can infer what someone knew based on physical evidence for how they searched an area (Pelz et al., 2020); and they can even detect the transmission of ideas by comparing artifacts created by different agents (Pesowski et al., 2020).

This past research suggests that the capacities needed to perform mental-state inference via event reconstruction might be in place from childhood. However, to our knowledge, no work has formally explored the event reconstruction hypothesis or quantitatively evaluated people's capacity to derive social inferences from indirect physical evidence. Here we present a computational model of social reasoning from agent-less physical scenes. Given indirect evidence that someone was present, our model infers what the agent was doing (i.e., reconstructs their actions) and why (i.e., infers their goals) through a generative model of how mental states produce actions, and how actions leave observable evidence.

In Experiment 1, we first tested whether our model matched human inferences in a task where participants had to infer an agent's entry point into a room and their goal, all from a single pile of cookie crumbs that revealed their presence (see Figure 1). In Experiment 2, we then explicitly tested people's ability to reconstruct the actions they believe different agents took based on indirect physical evidence of their presence, lending further support to the idea that the inferences in Experiment 1 were supported by an ability to reconstruct events. Finally, if social reasoning from physical scenes is supported by event reconstruction, people should be able to also infer how many agents might have been present in a room, based on how many paths they need to reconstruct to explain the scene. We tested this prediction in Experiment 3. Combined, our results suggest that people have a nuanced capacity to infer mental states from indirect evidence, and that these inferences are based on a basic capacity to "enhance" physical scenes by inferring agents' spatiotemporal behavior based on the indirect evidence that they leave behind. All studies were approved by the Yale University Institutional Review Board (protocol: "Online reasoning" #2000020357).

2. Computational Framework

Our model builds on a growing body of work showing that mental-state attribution is instantiated as Bayesian inference over a generative model of utility-

maximizing action plans (Baker et al., 2009, 2017; Jara-Ettinger et al., 2020; Jern et al., 2017; Jern & Kemp, 2015; Jern et al., 2011; Lucas et al., 2014). In our model, however, rather than evaluating unobservable mental states against observable actions, we model how people might use physical evidence to reconstruct the actions that an agent took, and use these reconstructed actions to attribute mental states.

To make our focus concrete, consider a situation like the ones shown in Figure 1a. Each of these displays represents a room with three possible goals (A in blue, B in orange, and C in green), two different doors (1 at the top in both rooms and 2 on the bottom and left, respectively), a set of walls (shown in dark gray), and a small pile of cookie crumbs that reveals that someone was previously in this room. Although we cannot see where this agent came from, what actions they took, or what goal they were pursuing, the cookie crumbs nonetheless contain information that we might be able to extract. In Figure 1a (left), the cookie crumbs intuitively reveal that the agent entered through door 1 and that they were likely pursuing goal A or C, but not goal B. In Figure 1a (right), the cookie crumbs intuitively reveal that the agent was pursuing goal C, but it is unclear whether they entered through door 1 or door 2. Our computational model aims to explain how we performed these inferences.

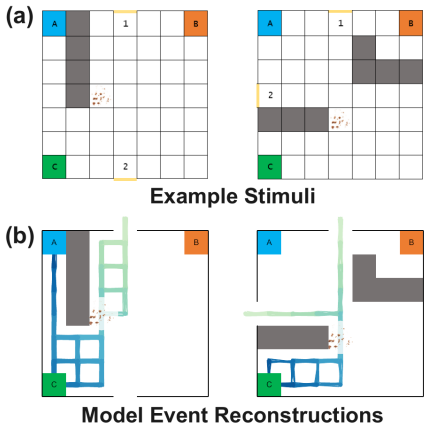

Figure 1: (a) Example stimuli from Experiment 1. Potential goals are positioned in the corners, labeled alphabetically, and color-coded. Doors are shown in yellow and coded numerically. Walls are shown in dark gray. Each trial included a pile of cookie crumbs positioned in a part of the room. (b) Visualizations of the underlying event reconstruction performed by our computational model for the examples above. Each line represents an inferred possible path, color-coded to indicate time, moving from light green to dark blue.

Formally, we model the environment as a gridworld, where the possible states

of the world are given by the different positions in space that agents can occupy. At each time step, we assume that agents can move in any of the four cardinal directions and that these actions successfully move them in their intended direction (except when attempting to cross a wall, in which case the agent remains in the same position as they were before).

Given an observed static scene s (a gridworld with a set of goals, doors, walls, and a pile of cookie crumbs), the objective is to infer where the agent entered the room from (a door d) and which goal they pursued (a goal g), formally expressed as

$$p(d, g|s) \propto \ell(s|d, g)p(d, g), \tag{1}$$

where $\ell(s|d, g)$ is the likelihood of encountering scene s if an agent had indeed pursued goal g after entering through door d, and $p(d, g)$ is the prior over doors and goals.

According to our proposal, the ability to compute the likelihood function is mediated by a capacity to reconstruct the agent's actions. Under this view, if we can reconstruct the actions that the agent took, then judgments about the agent's entry point and goal are immediately revealed, as these are part of the reconstructed behavior (i.e., if we have access to the full reconstructed behavior, we can "see" where the agent entered from and where they were going). Formally, this idea can be implemented by expressing the likelihood function as

$$\ell(s|d, g) = \sum_{t \in \mathbb{T}} \underbrace{p(s|t)}_{\substack{\text{how do actions} \\ \text{leave traces?}}} \times \overbrace{p(t|d, g)}^{\substack{\text{how do agents} \\ \text{pursue goals?}}}. \tag{2}$$

Here $t = (\vec{s}, \vec{a})$ is a trajectory (from the set of all possible trajectories $\mathbb{T}$), which consists of an ordered sequence of pairs of states and actions that the agent took. $p(s|t)$ is the probability that an agent who took trajectory t would produce the observed scene s, and $p(t|g, d)$ is the probability that the agent would take trajectory t if they entered from door d with the intention to pursue goal g. This equation reveals the two components critical to our theory: an expectation of how agents navigate to complete their goals ($p(t|d, g)$), and an expectation of how agents' actions leave observable traces in the environment ($p(s|t)$).

To compute the expectations for how agents complete their goals, we used the standard framework previously developed in computational models of goal

inference (Baker et al., 2009, 2017; Jara-Ettinger et al., 2020) through Markov Decision Processes (MDPs)—a planning framework that makes it possible to compute the action plan or *policy* that maximizes an agent's utility function (Bellman, 1957). Classical MDPs are designed to produce a single trajectory that fulfills the agent's goal as efficiently as possible. In the cases that we consider, however, there are often multiple trajectories that can be equally efficient. As such, using a simple MDP can erroneously treat an efficient trajectory as unlikely if it is not an exact match to the solution that the MDP produced. To solve this problem, we built a probabilistic MDP that creates a probability distribution over all possible action plans, assigning higher probability to trajectories that are more efficient. Formally, we achieved this by softmaxing the MDP's value function when building the probabilistic policy. We used a low temperature parameter to identify all possible action plans that are equally (or approximately equally) efficient, enabling us to implement the expectation that agents navigate efficiently towards their goals. Using a probabilistic MDP, the probability that an agent would take trajectory t, starting from door d with the intention to fulfill goal g is given by

$$p(t|g,d) = \prod_{i=1}^{|t|} p(a_i|s_i,g), \tag{3}$$

where $p(a_i|s_i,g)$ is the probability of taking action a_i in state s_i, and the state sequence is given by trajectory t.

Finally, in our paradigm, we assume that the agent has a uniform probability of dropping the pile of cookie crumbs at any point in their path. The probability of observing scene s if the agent took trajectory t is therefore given by $p(s|t) = 1/|t|$ if the pile of cookie crumbs lies within the trajectory and 0 otherwise.

2.1. Implementation Details

To generate testable predictions, we set a number of parameters in our model prior to data collection. We began by setting a uniform prior distribution over doors and goals, such that agents were equally likely to enter through any of the doors and equally likely to pursue any of the goals. Next, to model the forces that shape agents' actions, we assumed that agents incur a constant cost of 1 for any action that they take, and that goals produced numerical rewards over the range $0 - 100$. Finally, to make our MDP probabilistic, we applied a temperature parameter $\tau = 0.15$ to the value function. This parameter was set

a priori to ensure that the model would give equal probability to all paths that were equally efficient, while only placing a negligible probability on erroneous and inefficient trajectories.

Model inferences were obtained via Monte Carlo methods, sampling 1000 combinations of doors and goals and 1000 trajectories conditioned on the selected door and goal. Figure 1b visualizes our model's inferred trajectories for the examples shown in Figure 1a, with each line corresponding to a sample from the posterior distribution, color-coded to indicate time, moving from light green to dark blue. These visualizations show how our model reconstructs the agents' probable spatiotemporal behavior, which in turn reveal the agent's entry point and goal, matching the intuitive inferences associated with these examples in the introduction.

3. Experiment 1

In Experiment 1, we tested our model in a task where people had to infer which goal an agent was pursuing and where they came from, all from a single piece of indirect evidence about their presence. If people's ability to infer goals from physical evidence is mediated by event reconstruction, then their judgments should show a quantitative fit to our model predictions, including fine-grained patterns of uncertainty.

3.1. Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 37.02$ years, $SD = 11.20$ years).

3.2. Stimuli

Stimuli consisted of 23 gridworld images, like those in Figure 1a. Each gridworld was 7-by-7 squares in size and represented a room that contains three goal squares (A in blue, B in orange, and C in green), up to three doors (labeled 1, 2, and 3), and a pile of cookie crumbs. The goals were always in the same corners, but the position of the doors and the pile of cookie crumbs varied between trials. In addition to these three features, a subset of trials included walls (shown by the dark gray squares in Figure 1a) that agents could not walk through.

Our stimuli set was designed to capture different types of inferences while also controlling for features that simple heuristics could exploit (e.g., ensuring

that the target goal was not always the one closest to the cookie crumbs, and that it could not be determined by projecting a straight line that intersected the entrance and the location of the cookie crumbs). We began by considering four different possible inference patterns: full certainty (assigning probability close to 1 to a hypothesis; D trials), full negative certainty (assigning probability close to 0 to a hypothesis, while also not having full certainty over two remaining hypotheses; N trials), partial certainty (assigning a higher probability to one of the hypotheses; P trials), and no certainty (assigning a uniform distribution to the hypothesis space; U trials).

We first designed seven single-door trials that captured each of these inference patterns in goal inference (two D, N, and P trials, and one U trial; schematic versions shown in Figure 3a). We then designed 16 additional trials with multiple doors by combining every possible inference pattern for the goal the agent was pursing and the entrance that they took (schematic versions shown in Figure 3b).

3.3. Procedure

Participants read a brief tutorial that explained the logic of the task. After learning how to interpret the images, participants were told that agents were equally likely to enter the room from any of the doors with the intention of going directly to one of the three goals (to remove the possibility that agents pursue multiple goals, or wander aimlessly before selecting one). After the introduction, participants completed a questionnaire that ensured they had read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 23 trials in a random order. On each trial, participants answered a multiple-choice attention-check question ("Which corner is farthest from Door 1 (there may be more than one)?") and were asked to infer the agent's goal ("Which corner is the person going for?") using three continuous sliders, one for each goal (each ranging from 0, labeled as "definitely no," to 1, labelled as "definitely"). Trials with at least two doors included a third question that asked participants to infer the agent's entry point ("Which door did they come from?") using one slider per door. Participants were allowed to submit their responses for each trial only when they correctly answered the

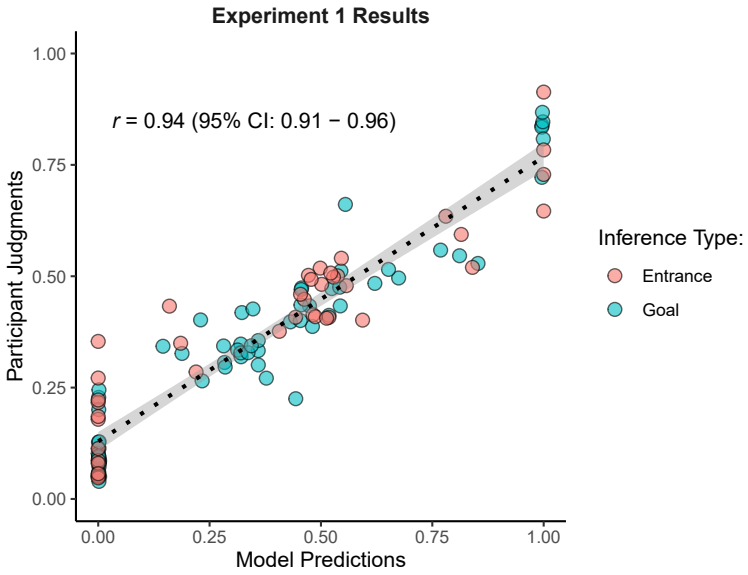

Figure 2: Results from Experiment 1. Each point corresponds to a judgment, with model predictions on the x-axis and mean participant judgments on the y-axis. Color indicates inference type and the dotted line shows the best linear fit with 95% confidence bands (in light gray).

attention-check question. Otherwise, participants were prompted to "please pay attention and try again."

3.4. Results

Participant judgments were first normalized within-trial (such that every distribution over goals or doors added up to 1) and then averaged across participants. Figure 2 shows the results from Experiment 1. Overall, our model showed a correlation of $r = 0.94$ (95% CI: $0.91 - 0.96$) with participant judgments, and the strength of the model fit was similar when looking only at goal inferences ($r = 0.95$; 95% CI: $0.92 - 0.97$) or door inferences ($r = 0.92$; 95% CI: $0.86 - 0.95$).

Figure 3 shows our model's results as a function of trial. In each subplot, the image at the top shows an abstract schematic of the trial, with the pile of cookie crumbs marked as a brown square. This figure reveals how our model not only predicted participant judgments in situations where the agent's entry point and goal were clear, it also matched participant judgments in its expression of uncertainty. Critically, our model's uncertainty reflects how well it was able

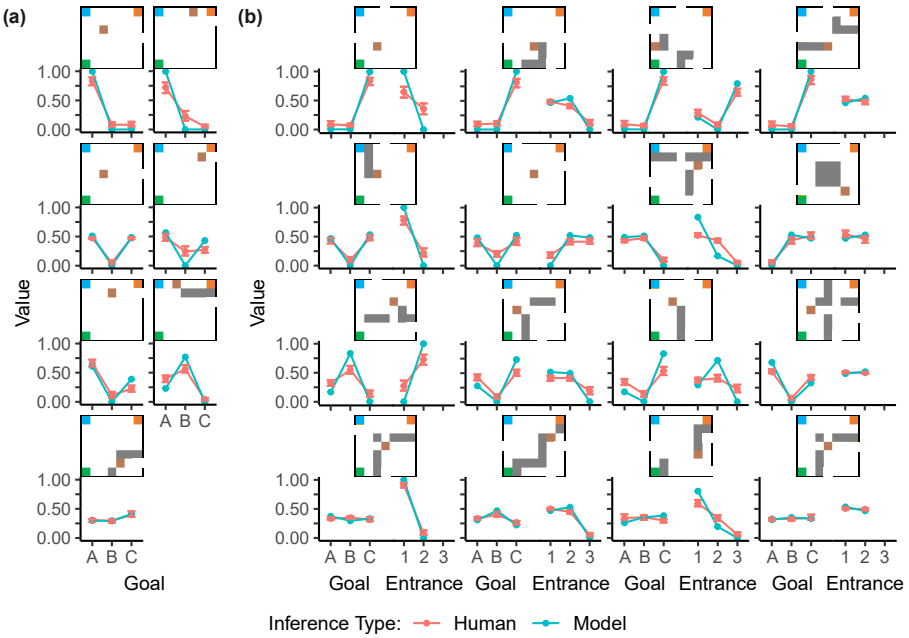

Figure 3: Detailed results from Experiment 1. From top to bottom, each row of subplots corresponds to the *D*, *N*, *P*, and *U* trials for goal inferences, respectively. (a) Results for trials that only had one door. (b) Results for trials that had more than one door. From left to right, each column of subplots corresponds to the *D*, *N*, *P*, and *U* trials for door inferences, respectively. The goals A, B, and C are indicated by the blue, orange, and green squares, respectively. The doors are sequentially numbered in a clockwise fashion, with door 1 starting from the top (or from the right if there is no top door). Walls are marked as dark gray squares and the pile of cookie crumbs are indicated by the brown squares. Red lines represent mean participant judgments and blue lines represent our model's predictions. Error bars on participant judgments represent 95% bootstrapped confidence intervals.

to reconstruct the event, becoming less confident as a function of how much conflict there is in entry points and goals across different hypothetical event reconstructions. The fact that this event-based uncertainty matched participant judgments with quantitative accuracy suggests that participants may have also been performing these inferences via some form of event reconstruction.

One alternative possibility is that participant judgments were driven by superficial features of the stimuli, rather than by event reconstruction. We tested this possibility through a multinomial logistic regression trained to predict participant goal inferences as a function of the distance between the pile of cookie crumbs and each goal, the average distance between the pile of cookie crumbs and each door, the number of doors, and all of their interactions. To train this regression, we transformed participant judgments into a one-hot vector,

marking 1 for the goal with the highest probability and 0 for the rest, and implemented LASSO regularization (Tibshirani, 1996) to avoid overfitting. We generated the alternative model's predictions in a leave-one-out fashion—that is, the predictions for each trial consisted of the output of a regression trained on all remaining trials.

Even though this alternative model was trained on the qualitative structure of participant judgments, it nonetheless only produced a correlation of $r = 0.49$ (95% CI: $0.30 - 0.63$) with participant judgments, which was substantially lower than the one produced by our model ($\Delta r = 0.46$; 95% CI: $0.33 - 0.65$). These results show that, while superficial features can capture the broad structure of participant judgments, they fail to do so at our model's level of granularity, further suggesting that people's inferences were centered on a form of Bayesian event reconstruction.

4. Experiment 2

In Experiment 1 we found that people can infer where an agent was going and where they came from, all from a single piece of indirect evidence about their presence. Participant judgments were quantitatively predicted by a model centered on an ability to reconstruct what happened. If our account is correct, then people should also be able to explicitly reconstruct the actions that an agent took in a way similar to our model. We test this prediction in Experiment 2.

4.1. Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 38.25$ years, $SD = 11.02$ years).

4.2. Stimuli

The stimuli were the same as those from Experiment 1 (see Figure 1a for examples and Figure 3 for schematic versions).

4.3. Procedure

Participants read a brief tutorial that explained the logic of the task. Participants were then instructed on how to draw their paths. After the introduction, participants completed a questionnaire that ensured they had read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to

participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 23 trials in a random order. On each trial, participants were asked to infer the path they thought the agent took, given the pile of cookie crumbs. Participants generated their paths by sequentially clicking on the squares they believed the agent walked through. Participants were only allowed to proceed when they had successfully generated a valid path, which consisted of paths that started at a door, ended at a goal, and passed through the pile of cookie crumbs. Participants were allowed to reset the drawn path as many times as they wished.

4.4. Model Predictions

To evaluate the participant-generated path reconstructions, we used our framework to calculate

$$p(t|s) \propto p(s|t)p(t), \tag{4}$$

where $p(s|t)$ is the likelihood of a trajectory t generating scene s and $p(t)$ is the prior over possible trajectories. Here, $p(s|t) = 1/|t|$ (like in Equation 2) and $p(t)$ is obtained by marginalizing over the agents' potential entry points and goals, as follows:

$$p(t) = \sum_{d,g} p(t|d,g)p(d,g). \tag{5}$$

4.5. Results

Our computational framework enables us to calculate the probability assigned to each path generated by participants. However, directly interpreting these probabilities is difficult, as they are sensitive to the length of the path and to the number of competing paths that fulfill a goal efficiently. To make our results easier to interpret, we compared our model's evaluations of the participant-generated path reconstructions with that of a baseline model. This baseline model used a uniform transition function over all actions, excluding the one that would generate a transition to the previous state (to prevent infinite back-and-forth loops). We then computed the Bayes factor for each reconstructed path by dividing the probability of that path, as predicted by our model (i.e., $p(t|s)$), by the probability predicted by the baseline model. A

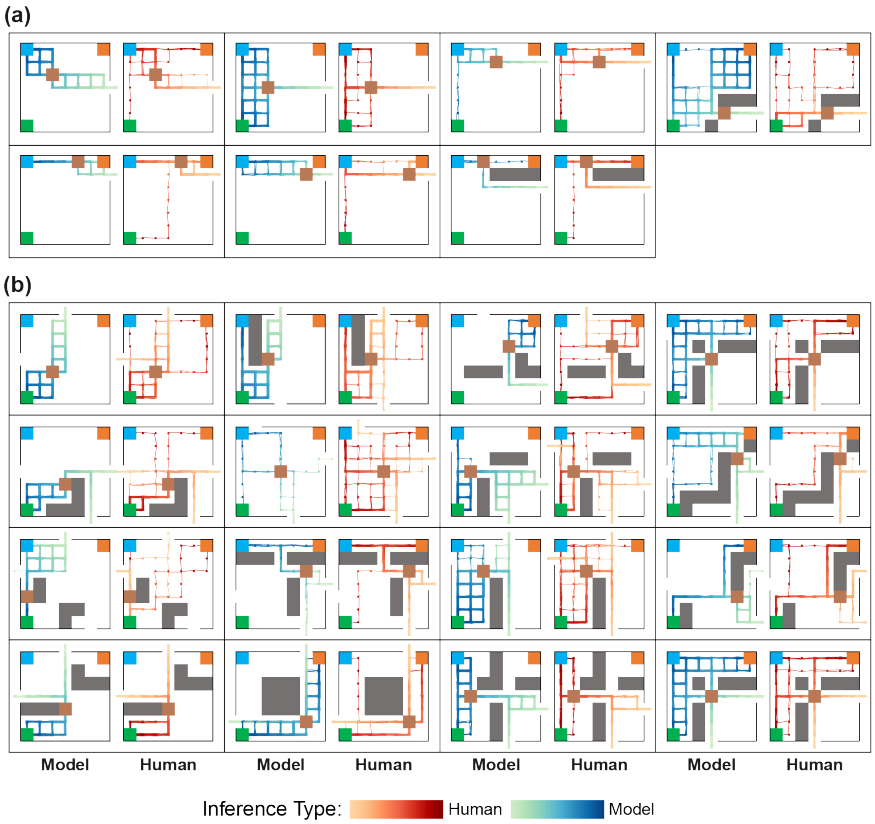

Figure 4: Comparison of reconstructed paths generated by our model and participants in Experiment 2. From left to right, each column of subplots corresponds to the D, N, P, and U trials for goal inferences, respectively. (a) Results for trials that only had one door. (b) Results for trials that had more than one door. From top to bottom, each row of subplots corresponds to the D, N, P, and U trials for door inferences, respectively. The goals A, B, and C are indicated by the blue, orange, and green squares, respectively. The doors are sequentially numbered in a clockwise order, with door 1 starting from the top (or from the right if there is no top door). Walls are marked as dark gray squares and the pile of cookie crumbs are indicated by the brown squares. Each line represents a reconstructed path, color-coded to indicate time, moving from light orange to dark red (for participants) or light green to dark blue (for the model).

Bayes factor greater than one would indicate that our model explains participant judgments better than the baseline model; a Bayes factor less than one would indicate that the baseline model explains participant judgments better than our model.

Our model outperformed the baseline model on all trials. The average Bayes factor in our experiment was 16935.33 (lowest factor = 7933.79; highest factor = 84383.12), meaning that our model was 16,000 times more likely to produce the participant-generated path reconstructions relative to the baseline model ($t(39) = 9.10$, $p < 0.001$ using a Bayes factor of 1 as the reference level).

Figure 4 shows trial-by-trial results from Experiment 2. Each trial is presented twice, with our model's path reconstructions on the left and participant-generated path reconstructions on the right. All paths are color-coded to indicate time (with darker colors occurring later in time). For both our model and participants, the higher path density indicates where the majority inferred the agent to have traveled. As this figure shows, the distribution of participant-generated path reconstructions largely matched those generated by our model (although participants were more likely to generate suboptimal paths).

5. Do explicit event reconstructions in Experiment 2 predict inferences from Experiment 1?

The previous results showed that that people can not only reconstruct agents' probable actions, but do so in a way similar to our model. According to our proposal, this event reconstruction underlies people's capacity to infer mental states from indirect physical evidence. If this is the case, then the information implicitly encoded in the path reconstructions from Experiment 2 should have predictive power over the inferences that participants made in Experiment 1. To test this possibility, we extracted the goals and doors from the participant-generated path reconstructions. To achieve this, we calculated the proportion of paths that originated from each possible entrance, and the proportion of paths that reached each possible goal, and compared these values to the corresponding goal and door inferences from Experiment 1. Figure 5 shows the results from this analysis. Overall, the goals and doors extracted from the participant-generated path reconstructions showed a correlation of $r = 0.89$ (95% CI: $0.83 - 0.92$) with the inferences participants made in Experiment 1, and the strength of this fit was similar when looking only at goals ($r = 0.88$; 95% CI: $0.80 - 0.93$) or doors ($r = 0.90$; 95% CI: $0.82 - 0.95$). Furthermore, when we compared these

extracted goals and doors against our model's predictions in Experiment 1, we found a correlation of $r = 0.86$ (95% CI: $0.79 - 0.91$), and a similar fit when looking only at goals ($r = 0.85$; 95% CI: $0.76 - 0.91$) or doors ($r = 0.88$; 95% CI: $0.78 - 0.93$).

Critically, participants in Experiment 2 could only generate a single path per trial. By combining the paths of multiple participants, we were able to reveal distributions over goals and doors that quantitatively resembled the inferences participants made in Experiment 1. The fact that these distributions predicted inferences from Experiment 1 suggests that generated paths were samples from the posterior distribution (rather than maximum likelihood or maximum *a posteriori* estimates, which would not contain enough information to reconstruct the full probability distribution over inferences). This analysis suggests that people had access to and sampled their paths in accordance to these goal and door distributions.

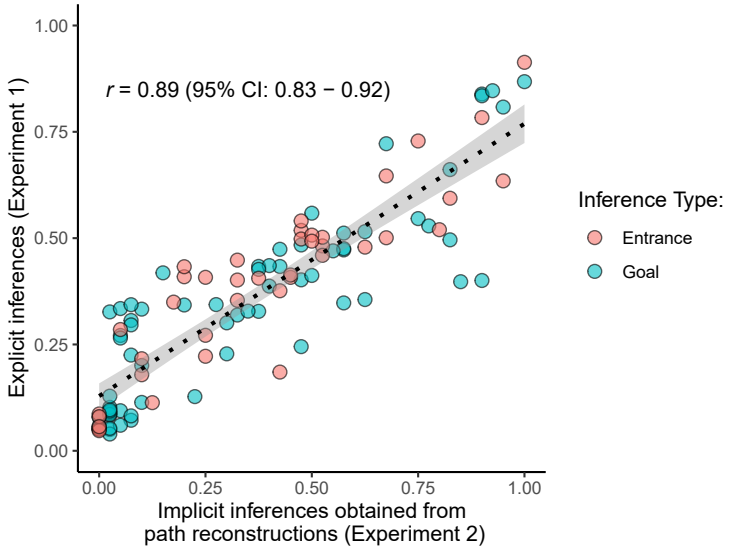

Figure 5: Comparison between the extracted goals and doors from Experiment 2 and the participant inferences from Experiment 1. Color indicates inference type and the dotted line shows the best linear fit with 95% confidence bands (in light gray).

6. Experiment 3

Experiment 1 showed that people can infer an agent's goals and origins, and that these inferences exhibit the quantitative structure predicted by a model of

event reconstruction. Experiment 2 further showed that people could explicitly reconstruct the paths in a way similar to our model. In Experiment 3, we test a further prediction of our account: If our model of event reconstruction is correct, then people should not only be able to infer a *single* agent's probable actions and goals, but also be able to estimate how many agents might have been in a room, based on how many path reconstructions are needed to explain a given scene.

6.1. Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 37.62$ years, $SD = 11.94$ years).

6.2. Stimuli

Our stimuli consisted of 15 gridworld images that were similar to those in Experiment 1 with the difference that each trial now has two piles of cookie crumbs instead of one (see Figure 6 for examples). Our stimuli set was designed to capture different types of inferences that our model supports. Specifically, we designed three different trials for each of the following possible inference patterns: high certainty that one agent was in the room (definitely one, or *D1*, trials), partial certainty that one agent was in the room (probably one, or *P1*, trials), uncertainty whether it was one or two agents in the room (uncertain, or *UN*, trials), partial certainty that two agents were in the room (probably two, or *P2*, trials), and high certainty that two agents were in the room (definitely two, or *D2*, trials).

6.3. Procedure

The procedure was nearly identical to Experiment 1, except that participants were shown two piles of cookie crumbs and were told that their task was to infer if one or two agents had been in the room. After the introduction, participants completed a questionnaire that ensured they had read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 15 trials in a random order. On each trial, participants answered a multiple-choice attention-check question ("Which corner is the farthest walk from Door 1? If there is more than one correct answer,

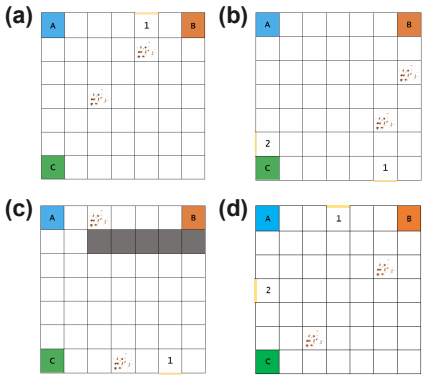

Figure 6: (a-d) Example stimuli from Experiment 3 for *D1*, *P1*, *P2*, and *D2* trials, respectively (see Experiment 3 Stimuli for details). Potential goals are positioned in the corners, labeled alphabetically, and color-coded. Doors are shown in yellow and coded numerically. Walls are shown in dark gray. Each trial included two piles of cookie crumbs positioned in various parts of the room.

just choose one of them.") and were asked to infer how many agents were in the room ("How many people were in the room?") using a continuous slider (ranging from 0, labelled as "definitely one," to 1, labelled as "definitely two"). Participants were allowed to submit their responses for each trial only when they correctly answered the attention-check question. Otherwise, participants were told to "please pay attention and try again."

6.4. Model Predictions

To predict how many agents might have been in a scene we computed the probability that a agents were in scene s, through

$$p(a|s) \propto p(s|a)p(a), \tag{6}$$

where $p(a)$ is a prior over the number of agents that could have been present. In natural contexts, this prior should reflect the statistics of how often different agents might interact in different environments. To model our experiment, however, we used a simple uniform prior over the possibility of having one or two agents. This prior was then weighted by the likelihood of a particular number of agents a generating scene s, given by

$$p(a|s) \propto \begin{cases} \sum_{t \in \mathbb{T}} p(s|t)p(t) & a = 1 \\ \sum_{t_1, t_2 \in \mathbb{T}} p(s|t_1, t_2)p(t_1)p(t_2) & a = 2 \end{cases} \tag{7}$$

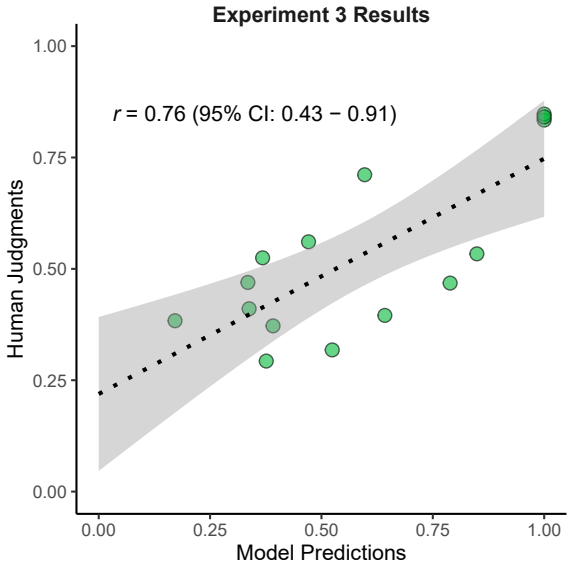

Figure 7: Results from Experiment 3. Each point corresponds to a judgment, with model predictions on the x-axis and mean participant judgments on the y-axis. The dotted line shows the best linear fit with 95% confidence bands (in light gray).

To compute the likelihood that two trajectories explain the scene (i.e., $p(s|t_1, t_2)$), we modified our generative model to sample two sets of entry points, goals, and trajectories at a time instead of one, where the likelihood is defined as $1/(|t_1| + |t_2|)$ if there was a scene match (i.e., both piles of cookie crumbs lie within both trajectories, and each trajectory was responsible for one of the cookie crumbs) and 0 otherwise.

6.5. Results

Participant judgments were averaged across trials and compared against our model's predictions. Figure 7 shows the results from Experiment 3. Participant's relative confidence about the number of agents in the scene was quantitatively similar to our model's predictions, yielding a correlation of $r = 0.76$ (95% CI: $0.43 - 0.91$).

Figure 8 shows our model's results as a function of each trial. In each subplot, the image at the top shows an abstract schematic of the trial, with both piles of cookie crumbs marked as brown squares. From left to right, each column corresponds to the *D1*, *P1*, *UN*, *P2*, and *D2* trials, respectively. This figure reveals how our model quantitatively predicts participant judgments across the

various trials and levels of uncertainty.

Interestingly, the model fit in Experiment 3 was lower relative to Experiment 1. Under our account, this difference arises because Experiment 3 requires reconstructing paths for a single agent, reconstructing paths from multiple agents, and weighting their relative probability of generating the observed scene. Consistent with this, we found higher mismatches between our model and participants in the P trials ($MSE = 0.053$) over the D ($MSE = 0.021$) and U trials ($MSE = 0.019$). That is, participants struggled more in trials that relied on a capacity to make precise comparisons between the number of single-agent reconstructions and two-agent reconstructions.

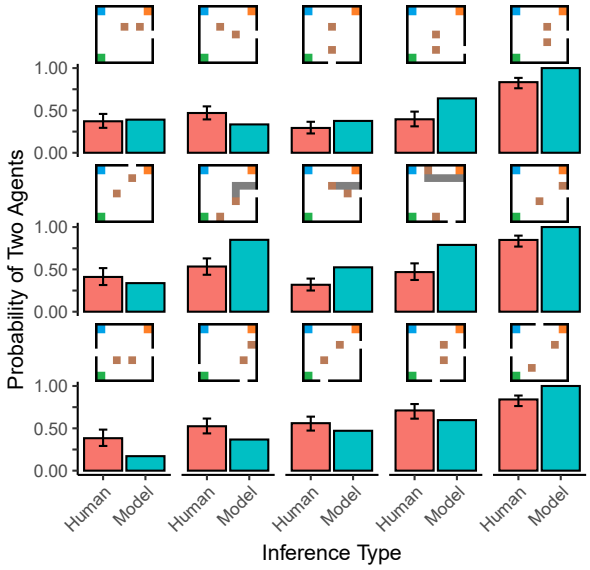

Figure 8: Detailed results from Experiment 3. From left to right, each column corresponds to *D1*, *P1*, *UN*, *P2*, and *D2* trials, respectively. Red bars represent mean participant judgments and blue bars represent our model's predictions. Error bars on participant judgments represent 95% bootstrapped confidence intervals.

Like in Experiment 1, we also evaluated whether participant judgments could be explained by superficial features of the stimuli rather than via event reconstruction. We tested this possibility through a logistic regression trained to predict participants' distribution over the number of agents they thought were in the room as a function of the distance between each goal and each pile of cookie crumbs, the average distance between each pile of cookie crumbs and the doors, the number of doors, and all of their interactions. We trained and tested

this alternative model in the same way as the one described in Experiment 1.

Even though this alternative model had access to the qualitative structure of participant judgments, it nonetheless produced a correlation of $r = 0.19$ (95% CI: $-0.30 - 0.66$) with participant judgments, which was substantially lower than the one produced by our model ($\Delta r = 0.58$; 95% CI: $0.12 - 1.17$). These results extend our findings from Experiments 1 and 2, suggesting that people can not only infer an agent's goals and origins based on indirect evidence of their presence, but also whether multiple agents may have been present in a given scene.

7. Discussion

Research on human action understanding has historically focused on how we infer the goals and mental states of agents whose behavior we are observing. Our results show that our capacity to reason about others goes beyond face-to-face interactions and includes nuanced social inferences from simple physical scenes. In Experiment 1, we showed that people can infer an agent's desires (i.e., where an agent was going) and past actions (i.e., where an agent came from) from just a single piece of indirect evidence about their presence. The tight correspondence between our model's predictions and the fine-grained structure of participant judgments suggested that these inferences were structured around a form of mental event reconstruction, where people infer the actions that an agent took and use this reconstructed behavior to make richer social inferences. Experiment 2 showed further support for our proposal, revealing that people can explicitly reconstruct the actions that someone took in a way similar to our model. Furthermore, these explicit reconstructions predicted the inferences participants made in Experiment 1, showing a direct link between people's social inferences from physical evidence, and people's ability to reconstruct behavior. Finally, in Experiment 3, we showed that people can use this capacity to infer the number of agents that were in a given scene, based on the number of paths they needed to reconstruct to explain the scene.

Our computational model formalized these inferences as the process of reconstructing behaviors that can explain the indirect observable evidence. Our model's quantitative fit with participant judgments, as well the failure of our alternative models (despite being trained on participant judgments), suggests that people were performing similar computations. In particular, the similarity between the paths generated by our model and those drawn by participants (see

Figure 4) further show that people can indeed reconstruct an agent's behavior through an expectation that agents act rationally and efficiently.

The heart of our proposal—expressed in Equation 2 (see Section 2)—posits event reconstruction as the key representation that connects two different cognitive capacities. The first is a model of how agents act as a function of their goals and environmental constraints (i.e., a Theory of Mind). The second is a model of how agents' actions may or may not leave observable traces in the environment. In this paper we focused on testing the general framework that we proposed, using a simple model of how agents leave traces in the environment. However, our computational framework only requires an ability to calculate the likelihood of scenes given behaviors ($p(s|t)$ in Equation 2), and the computations within this component can be arbitrarily complex. Here we consider two richer models that might be employed in future work.

A first way in which our framework could tackle richer inferences is by using a full-fledged model of intuitive physics to evaluate how actions leave traces in the environment. A recent body of work in cognitive science has found that human intuitive physics is instantiated as a *physics engine* that supports rich probabilistic simulations of how objects and forces interact in the environment (Fischer et al., 2016; Battaglia et al., 2013), and that physical simulations might underlie how we reason about the interaction between agents and objects (Yildirim et al., 2019). Thus, using a physics engine to simulate how the forces that agents apply to the world leave observable changes might enable our computational framework to handle more complex physical events that contain social information.

A second possible extension lies in changing what we consider to be an observable scene. Our focus here was on inference from physical information, but recent studies have found that people can also infer other people's actions from social evaluations, such as inferring what someone might have done by learning that they were blamed by others about a failure (Davis et al., 2021). These inferences might be understood as an extended form of this framework, where the second term is replaced with an expectation of how people's behavior causes social outcomes, rather than physical ones.

Similarly, our computational framework also allows for more complex models of agents' behavior, as long as they can express the likelihood of different actions under different goals and environmental constraints. In our work, we used a model structured on an assumption that agents act rationally and efficiently under perfect knowledge. In future work, we hope to extend this work to use

models where agents can have partial or incomplete knowledge of their environment (e.g., Baker et al., 2017). This would enable our framework to consider situations where indirect evidence reveals an agent's intention to explore and understand their surroundings rather than to complete a known goal.

While our work focused on adults, some recent research suggests these capacities might emerge in early childhood. In particular, preschoolers can judge what types of physical constructions (such as different types of block towers) require more physical effort (Gweon et al., 2017), suggesting an early understanding between actions and physical outcomes. At the same age, children can also determine what actions are more likely to leave physical traces. For example, lifting an upside-down cup filled with rice will likely leave visible rice grains after the cup has been repositioned. But it is possible to lift and reposition an upside-down cup filled with a few large rocks without leaving any evidence behind (Jacobs et al., 2021). Moreover, children can also associate physical outcomes with the corresponding mental states of the agent who generated them (Pelz et al., 2020). Finally, and most strikingly, young children can infer the transfer of ideas by seeing how different agents create artifacts (Pesowski et al., 2020), a capacity known as "intuitive archaeology" (Hurwitz et al., 2019; Schachner et al., 2018). While these results point towards an early understanding of the relation between the social and physical world, to our knowledge, it is an open question whether these inferences are also linked to some form of explicit or implicit event reconstruction.

At first sight, our computational framework appears to suggest that any creature with some form of naïve psychology and naïve physics ought to be able to perform social inferences from physical evidence (i.e., access to the two key components of Equation 2). This may not be the case, however, because our model also requires an ability to transfer information across these intuitive theories (reconstructing behavior via naïve psychology and evaluating how they compare to the environment via naïve physics). While this is an open empirical question, research suggest that intuitive physics and intuitive psychology rely on separate neural circuitry (Fischer et al., 2016; Saxe & Powell, 2006), leaving open the question of how these two intuitive theories might work in tandem to reconstruct other people's behavior from physical evidence.

One interesting case that suggests such a feat might not be simple comes from research with vervet monkeys. Vervet monkeys have an astonishing degree of social intelligence, including a nuanced repertoire of vocal calls to signal different types of predators, each associated with different escape responses (Seyfarth

et al., 1980a,b). Yet, vervet monkeys routinely fail to identify predators from indirect physical evidence. For instance, vervet monkeys fail to infer that a python is hiding in a nearby bush when they encounter the distinct tracks that they leave behind. Similarly, vervet monkeys also fail to infer the presence of a leopard upon encountering a gazelle carcass on a tree (where leopards usually drag their prey so they can feed in solitude; Cheney & Seyfarth, 1985). Critically, this failure appears to persist even after vervet monkeys have, in past events, seen the direct association between the physical evidence and the predator (Cheney & Seyfarth, 1985, 2008). These results might point to the possibility that the form of event reconstruction that we present here might require capacities that go beyond simple physical and social reasoning, as they involve an ability to combine the two capacities to derive richer inferences than would be otherwise possible.

Our work also leaves a critical question open. In our experiments, we focused on situations where people already knew that an agent was previously present. Our work therefore does not speak to how people recognize that a scene contains traces of someone's behavior in the first place. One possibility is that people engage in a pervasive and constant social analysis of all physical scenes. Doing so, however, might be prohibitively costly and unnecessary. As such, it is likely that people are attuned to the physical signatures that reveal the presence of an agent, which then trigger social reasoning from physical evidence. Consistent with this second view, research suggests that people can infer the presence of an agent based on apparent order (Newman et al., 2010; Keil & Newman, 2015) and on a sensitivity to human-like errors that people leave behind when interacting with the world (Lopez-Brau et al., 2021). An open question is how the ability to detect the presence of an agent interacts with the ability to reconstruct their behavior and infer their mental states.

Overall, our results illustrate the sophistication of human social intelligence. Beyond being able to read the mental states of agents that we are personally interacting with, we can also infer the mental states of agents we have never encountered, just from minimal indirect evidence that reveals their presence. Researchers have long argued that humans are unique in their ability to reason about and navigate the social world (Herrmann et al., 2007). Our work shows that this ability is not confined to social interactions, but can fundamentally affect how we reason about the physical world, allowing us to see social meaning embedded in physical structures, like a pile of rocks, where other animals may see merely just that: a pile of rocks.

8. Acknowledgments

This work was supported by NSF award BSC-2045778 awarded to JJE. All analysis and materials can be found at `https://osf.io/q3ct5/?view_only=f2fa5332eb4545bda9fb5353eb73daab` (Lopez-Brau, 2021).

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.

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

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327–18332.

Bellman, R. (1957). A markovian decision process. *Journal of Mathematics and Mechanics*, (pp. 679–684).

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*, 144–152.

Cheney, D. L., & Seyfarth, R. M. (1985). Social and non-social knowledge in vervet monkeys. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*, 187–201.

Cheney, D. L., & Seyfarth, R. M. (2008). *Baboon metaphysics: The evolution of a social mind*. University of Chicago Press.

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, *6*, e29718.

Davis, Z., Allen, K., & Gerstenberg, T. (2021). Who went fishing? inferences from social evaluations. *CogSci proceedings*, .

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*, E5072–E5081.

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

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). Words, thoughts, and theories.

Gosling, S. D., Gaddis, S., & Vazire, S. (2008). First impressions based on the environments we create and inhabit. *First Impressions*, (pp. 334–356).

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*, 379.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. In *CogSci*.

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*, 1360–1366. doi:10.1126/science.1146282.

Ho, M. K., Cushman, F. A., Littman, M., & Austerweil, J. L. (2019). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, .

Hurwitz, E., Brady, T., & Schachner, A. (2019). Detecting social transmission in the design of artifacts via inverse planning.

Jacobs, C., Lopez-Brau, M., & Jara-Ettinger, J. (2021). What happened here? children integrate physical reasoning to infer actions from indirect evidence. *CogSci*, .

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

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

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

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38.

Jern, A., Lucas, C., & Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. *Advances in Neural Information Processing Systems*, *24*, 2276–2284.

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

Keil, F. C., & Newman, G. E. (2015). Order, order everywhere, and only an agent to think: The cognitive compulsion to infer intentional agents. *Mind & Language*, *30*, 117–139.

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.

Lopez-Brau, M. (2021). Social inferences from physical evidence. URL: `https://osf.io/q3ct5/?view_only=f2fa5332eb4545bda9fb5353eb73daab`.

Lopez-Brau, M., Colombatto, C., Jara-Ettinger, J., & Scholl, B. (2021). Attentional prioritization for historical traces of agency. *Vision*, .

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.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., & Wynn, K. (2010). Early understandings of the link between agents and order. *Proceedings of the National Academy of Sciences*, *107*, 17140–17145.

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

Pelz, M., Schulz, L., & Jara-Ettinger, J. (2020). The signature of all things: Children infer knowledge states from static images. *PsyArXiv*, .

Pesowski, M., Quy, A., Lee, M., & Schachner, A. (2020). Children use inverse planning to detect social transmission in design of artifacts. *PsyArXiv*, .

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*, 692–699.

Schachner, A., Brady, T., Oro, K., & Lee, M. (2018). Intuitive archeology: Detecting social transmission in the design of artifacts.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980a). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, *210*, 801–803.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980b). Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Animal Behaviour*, *28*, 1070–1094.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288.

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

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

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

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk.