Minimal covariation data support future one-shot inferences about unobservable properties of novel agents

Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)[1], Hyowon Gweon (hyo@stanford.edu)[2],
[1] Department of Psychology, Yale University, New Haven, CT 06520
[2]Department of Psychology, Stanford University, Stanford CA, USA

Abstract

When we reason about others' behavior, there are often many equally-plausible explanations. If Bob climbs a tree to get an apple, we may be unsure if Bob found climbing difficult but really wanted an apple; if he found climbing easy and was not particularly excited about the apple; or if he found climbing intrinsically fun and just got the apple because it was convenient. Past research suggests that we solve this problem by obtaining repeated observations about the agent and about the world. Here we argue that, beyond allowing us to sharpen our inferences about agents and the world, covariation data also enables us to do one-shot inferences about novel agents. We show that given minimal covariation data, people can infer objective and subjective properties of a new agent from a single event. We show that a model that assumes that agents maximize utilities matches participant judgments with quantitative precision.

Keywords: theory of mind; social cognition; computational modeling

Introduction

In our everyday social interactions, we easily learn aspects of people that are directly observable. We hear people's names, see what they look like, and recognize their jobs. But getting to know someone means much more: what they like, what they're good at, and even what they think of themselves. We invest much of our social interactions gathering observable evidence about these unobservable qualities of others, and even plan opportunities that serve specifically this purpose such as interviews with applicants or dates with potential partners.

A growing set of studies suggest that when we reason about others we assume that they act to maximize the rewards that they obtain relative to the costs that they incur (see Lucas et al. 2014 and Jara-Ettinger et al. 2016 for review). If, for instance, we watch an agent walk straight to a coffee shop, we can infer that getting coffee is rewarding (explaining why the agent went there) and that walking is costly (explaining why she took the shortest path). Despite its simplicity, this ability to reason about behavior in terms of costs and rewards, called a *Naïve Utility Calculus*, supports rich explanations, enabling observers to distinguish between highly motivated agents (high rewards) and poorly motivated agents (low rewards), and supporting reasoning about agents who ignore goals because of a lack of competence (costs are too high) and because of a lack of motivation (rewards are not high enough).

Decomposing behavior into costs and rewards, however, means that action-understanding is usually confounded, even in the simplest scenarios. If, for example, an agent jumps over an obstacle to reach an object on the other side, her behavior can be explained equally well by appealing to different combinations of costs and rewards. The agent may have found jumping very costly, but the outcome even more rewarding. Alternatively, she may have found jumping relatively easy, and the outcome not particularly rewarding. Or she may have even found jumping rewarding, and not cared about the object. To complicate matters further, agents not only incur costs and obtain rewards, but they also have beliefs about their own costs and rewards, and these beliefs guide their behavior. Imagine, for instance, watching a girl pull out a sword from a stone. While it is trivial to see that her goal was to get the sword (and that it was therefore rewarding), it is difficult to determine how much she wanted the sword (was the reward high or low?), how strong she is (is the cost low for her?), how difficult it is to pull the sword out (is the cost high in general?), or what she thought about her own strength before trying (what did she believe about her own costs?).

The problem of confounded explanations is most obvious when we only have access to a single event. But in more realistic situations, we often watch different people pursue the same goal, and we watch the same person pursue different goals (see, e.g., Figure 1). This *covariation* most directly allows us to learn about the agent we are observing (Kelley & Michela, 1980). However, it may also enable us to make stronger inferences about new agents. Returning to the example above, what if you knew that several other people had already tried to pull the sword out and failed, and that the girl decided to try anyway? Even though the information about the girl is the same, you might be more confident about your inferences in this second case: the girl probably really wanted the sword (she probably believes that the cost is in general high), she thought she'd be strong enough to succeed (she believes that the cost may be lower for her specifically), and she was right (our observation of her success suggests the cost was indeed lower for her)!

Here we propose that minimal covariation data about the outcomes of agents' goal-directed actions, combined with our commonsense psychology, enable us to make richer inferences about novel agents. We show that even from a brief history of actions, people can make powerful joint inferences about a new agent's desire, competence, and even beliefs about their own competence, all from a single action. Below we briefly review research that motivates our proposal, we present our theory instantiated as a computational model in a Bayesian framework, and we then present two experiments that test our model predictions.

Agent-dependent and agent-invariant dimensions of costs and rewards

Costs and rewards are partially objective and agent-invariant (e.g., a high hill is more costly to climb than a low hill, and three cookies are more rewarding than one), and partially subjective and agent-dependent (e.g., some are better than others at hill-climbing, and some like cookies more than others). Thus, to effectively explain an event, we not only need to infer the underlying costs and rewards, but we must also uncover what aspects of the costs and rewards are specific to the agent and what aspects of the costs and rewards are properties of the world and apply to all agents. Decomposing costs and rewards into agent-dependent and agent-independent dimensions not only helps us understand the event better. It also helps us understand new events more easily. If we know what costs and rewards are specific to an agent, then we can use this knowledge to explain the agent's behavior in new events (e.g. if learn that someone is strong, this helps us interpret their successes and failures in new events). If we know what aspects of costs and rewards are properties of the environment, then we can use this knowledge to make sense of new agents acting in this familiar situation (e.g. if we learn that a box is heavy, this helps us interpret the success or failure of new agents when trying to lift the box).

One-shot learning from covariation information with a Naïve Utility Calculus

Based on these intuitions, we propose that people rely on covariation information to break down costs and rewards into their agent-dependent and agent-independent components, and that, with this decomposition at hand, people rely on their Naïve Utility Calculus to make rich inferences from single events. Returning to the example above, if we already understand that getting the sword is difficult to pull out because many have failed, then, if we a new agent succeed, we can be sure that it was not because the sword was easy to pull out, but because the person was strong; an inference that would have been impossible to make the first time we saw this. Similarly, if the successful agent had already watched others try and fail, we can assume that she also knew the sword as difficult to lift, and so she probably thought she has strong enough to succeed; otherwise, she would not have bothered trying. If she succeeds, then we can also be certain that she really was strong.

Recent work suggests that even infants can use covariation information to infer properties of the world and properties of agents. When one agent successfully activated a toy twice but the other failed twice (suggesting one is more competent than the other), infants attributed their own failure with the toy to their incompetence and sought help from others; conversely, when each agent succeeded twice and failed twice on the toy, infants attributed their failure to the toy, seeking a different one instead (Gweon & Schulz, 2011). Furthermore, older children (4- and 6-year-olds) use covariation information between characters and activities to generate different causal explanations for their behaviors (Seiver, Gopnik, & Goodman, 2012). For instance, if Sally and Anne both tried activity A but not B, children were more likely to appeal to the properties of the activities to explain their actions (e.g., A is more fun than B); but when Sally tried both A and B but Anne tried neither, children appealed more to the characters' attributes (e.g., Sally is older). Furthermore, children generalized these explanations to predict whether the characters would try a new activity, or what another character would do on the same activities. These results suggest that humans, even early in life, are sensitive to the covariation information embedded in others' actions: they infer both the relevant properties of people and the physical world (e.g., toys, activities) and readily use it to explain their actions.

Similarly, children have a Naïve Utility Calculus by age five, with some form of it tracing back to infancy. Even infants have some expectation that agents navigate efficiently (Csibra, 2003) and that this expectation reflects some understanding of cost minimization (Liu & Spelke, 2016). Also before their second birthday, children understand that both competence and rewards vary across agents (Repacholi & Gopnik, 1997; Jara-Ettinger, Tenenbaum, & Schulz, 2015). And by age five, children can explicitly explain behavior by inferring the unobservable costs or rewards, given partial information (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016).

As reviewed above, the two main accounts that our proposal relies on -understanding covariation, and having a Naïve Utility Calculus- are both available early in life. Although our goal here is to explore this possibility with adults, the developmental research suggests that the abilities our account requires are likely central to social reasoning as they can be traced to our first years of life. The next section describes our computational model that formalizes these intuitions. We use the model to obtain quantitative predictions and compare them against empirical data across two experiments. In Experiment 1 we test if our account explains how we jointly infer properties about agents and the world using covariation information, and how this past information, in turn, supports one-shot learning of objective and subjective properties of novel agents. Because participants in Experiment 1 explicitly make judgments about the covariation in formation, in Experiment 2 we test if this step is critical for people to integrate this information when reasoning about new agents.

Computational modeling

In order to test our predictions more formally, we implemented a computational model of our account and a simple alternative model that ignores the covariation data when inferring properties of the new agent. The principles of our model apply to any situation in which the outcome of events depend jointly on properties of agents and properties of the world; here, we describe it in the context of our experimental paradigm (see Procedure section in Experiment 1 and Figure 1), where agents with different levels of strength attempt to

lift boxes of different weighs in order to obtain rewards.

Naive Utility Calculus model

Our implementation is a simplified variation of the Naïve Utility Calculus model (Jara-Ettinger, Schulz, & Tenenbaum, 2015). Whereas the past models were designed to reason about agents navigating in two-dimensional environments, this model is adapted for reasoning about agents making choices without any spatial information (along the lines of (Lucas et al., 2014)). For a single event, the agent's strength and the box's weight are inferred using Bayesian inference:

$$p(W,S|O) \propto p(O|W,S)p(W)p(S) \tag{1}$$

where W is the weight of the box, S is the strength of the actor, and O is the observed outcome (success or failure). For simplicity, we use a deterministic likelihood function where agents can successfully lift a box only when their strength is higher than the box's weight. As such, we represent strength and weight using a common scale, using real values ranging from 0 to 1.

By providing covariation information, where agents interact with different boxes, Equation 1 enables observers to break down events into agent-dependent (strenght) and agent-invariant (weight) components. With this information at hand, when we watch a single event from a new agent (henceforth the *one-shot agent*), we compute her preference by relying on the assumption that she is attempting to maximize her subjective utilities (see Introduction). An agent's expected utility for any given box is given by the reward associated with the box times the probability that the agent will be able to retrieve it. As such, an agent's choice reflects a trade-off between the magnitude of the reward, and the probability that the agent will be able to get it if she tried. Given a choice C, the posterior probability of the agent's underlying preferences is given by

$$p(P|C) \propto p(C|P)p(P) \tag{2}$$

where P represents the rewards associated with each option. For simplicity, we assume that the observer has a uniform prior over the agent's preferences (p(P)), and we compute $p(C|P)$ by integrating the observer's prior belief over the actor's strength:

$$p(C|P) = \int_S p(C|P,S)p(S), \tag{3}$$

where S is the agent's strength, and C is the agent's choice. This intermediate term, $p(C|P,S)$, integrates the assumption that the agent is attempting to make choices that maximize her utilities. Finally, the one-shot agent's objective strength is also computed using equation 1.

Alternative model

To test the role of the past covariation information in the final (one-shot) trial, we implemented a simple alternative model. In this baseline model we assume that participants ignore the covariation information and make all judgments about the one-shot agent using that event alone. As such, this model is computationally equivalent to the main model, as it relies on Equations 1-3 to reason about the agent, but it does not use the covariation data to sharpen its estimates.

Experiment 1

To test our hypothesis that people can use past observations of multiple agents to make one-shot inferences about a novel agent, we designed a behavioral experiment where participants received covariation data about three agents, each of whom attempted to lift four different boxes (see Figure 1). Next, participants watched a single agent choose one of the boxes and either succeed or fail to lift it. After this single event, participants were asked to infer three properties of the agent: her preference, her beliefs about her own strength, and her true strength.

Methods

Participants 100 adults participants (mean age = 35.95; range: 19-70) from the US (as determined by their IP address) were recruited using Amazon's Mechanical Turk framework. Participants were randomly assigned to one of 10 conditions (see Procedure).

Procedure Participants read a brief story that consisted of two parts. In the first part (Part 1 in Fig.1), participants learned about a game where, if players could successfully lift a box, they were allowed to keep its contents. Next, participants learned about three players (Circle, Rhombus, Triangle) who played with different boxes. There were five boxes, but the agents only had four coins and interacted with just the first four (Candy, Teddy Bear, Rubber Duck, and Baseball boxes); no one interacted with the fifth box (Yoyo box) and no mention was made about it other than stating that it was an option. For each action of each agent, participants learned whether the agent succeeded or failed; the first agent (Circle) sequentially tried the four boxes (in fixed order as shown in Figure 1), followed by the second (Rhombus), and then the third (Triangle). After each attempt, the cumulative outcomes were summarized visually as in Figure 1. After observing this covariation data, participants were asked to determine how heavy each box was and how strong each agent was. Both types of questions were answered on a numerical scale ranging from 0 to 9. In the weight questions, 0 indicated very light, 5 indicated average, and 9 indicated very heavy. In the strength questions, 0 indicated very weak, 5 indicated average, and 9 indicated very strong.

In the second part of the task (Part 2), participants learned about a fourth agent (Square) who had also watched the other three agents. Participants learned that this final agent only had enough money to play the game just once. Participants were then shown which box (of the five) the agent selected, and whether she succeeded or failed in lifting it. Crossing agent's choice (5 boxes) and outcome (success or failure) produced 10 conditions, to which participants were randomly as-

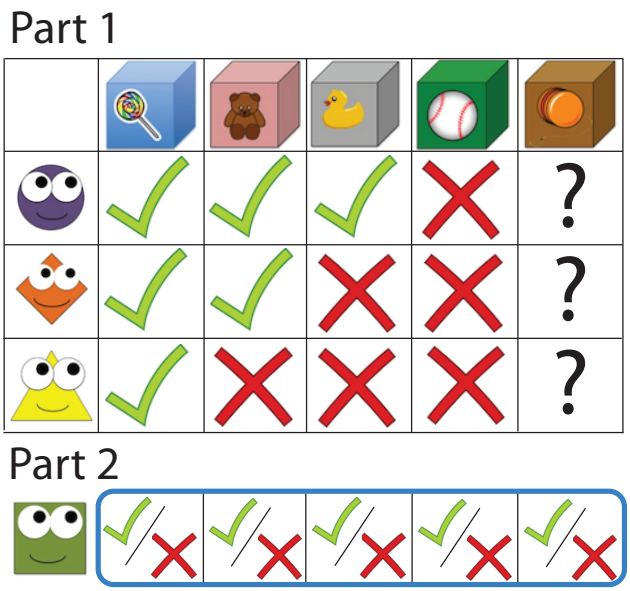

Figure 1: Visual summary of the experiment. Participants were introduced to four agents and five boxes. The first three agents (the square, rhombus, and triangle) interacted with the first four boxes (but not the fifth box). In Experiment 1, after observing these trials (and before seeing the final agent), participants were asked to rate the relative strength of these three agents, and the relative weight of the four boxes. In the second part of the experiment, the final agent (the square) chose one of the five boxes and either succeeded or failed to lift it (producing a total of 10 conditions that we tested across participants). Participants were then asked to determine this agent's preference, strength, and beliefs about her own strength when she made her choice.

signed (see Part 2 in Figure 1). Participants were then asked three questions in the following order. First, participants were asked to rate how much the agent wanted the object in the box using a scale from 0 ("not at all") to 9 ("very much"); Preference. Second, they were asked to rate the agents strength on a scale from 0 ("very weak") to 9 ("very strong"); True Strength. Third, participants were asked to rate the agent's beliefs about their own strength on a scale using an identical scale to the one used in the second question (Perceived Strength).

Results

Participants' responses from the experiment were z-scored within response type (preference inferences, weight inferences, and strength inferences) and then averaged across participants.

First, we looked at people's use of covariation data by looking at their inferences about agents' strength and boxes' weights from Part 1. The model provided very high quantitative fits (Figure 2). On the joint inference over strength

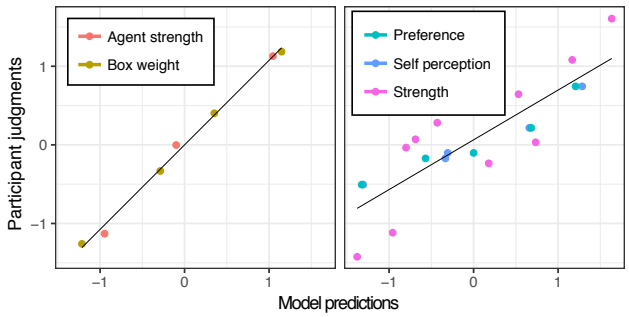

Figure 2: Overall results from Experiment 1. The x-axis shows the model predictions and the y-axis shows participant judgments. The left plot shows inferences obtained from the covariation data (see Figure 1). The right plot shows inferences made from the one shot event.

and preference for the first set of agents (see Part 1 in Figure 1), the model showed a correlation of r=0.99 with participant data (95% CI: 0.99-1.00).

Having verified that participants attended to the covariation data in Part 1 and accurately inferred the boxes' weight and the agents' strength, we then looked at whether participants made used this information when interpreting the event from the one-shot agent in Part 2.

Qualitatively, the results from the one-shot learning trial were as expected (see Figure 3). Participants judgments about the agents true strength varied both as a function of the box that she chose, and the outcome. Similarly, inferences about the agent's beliefs about her own strength also varied as a function of the box that she chose to lift.

On the joint inferences about the final agents preference, true strength, and perceived strength (Part 2), participant judgments showed a correlation of r=0.86 with participant data (95% CI: 0.67-0.94).

By contrast, our alternative model, which used the same computations but did not learn from the covariation data, failed to predict the one-shot inferences participants made about the novel agent. Because the model ignores the covariation data, it does not make any predictions about the first set of agents; thus we only report the fit between the alternative model and people's responses in Part 2, about the target agent. The model showed a correlation of r=0.40 (95% CI: -0.06,0.71) against participant judgments.

Experiment 2

Experiment 1 established that, when given covariation data, people can infer a novel agent's preference, strength, and perceived strength from a single event. In this experiment participants were explicitly asked to think about the covariation data and judge the strength of each agent and the weight of each box. It is possible that people do not naturally decompose preferences and competence into agent-dependent and agent-independent features, and this only happens when par-

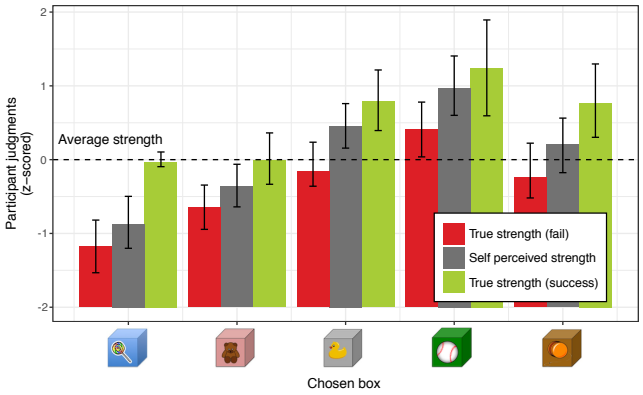

Figure 3: Results from Part 2 of Experiment 1. The x-axis shows the box that the protagonist shows and the y-axis shows participant's ratings for the agent's perceived strength and the agent's true strength. Red bars show the conditions where the agent failed to lift the box, green bars show the conditions where the agent successfully lifted the box, and the grey bars show the agent's self-perceived strength. Judgments are z-scored within participants and averaged and the vertical bars represent 95% confidence intervals. People inferred lower strength when the agent failed relative to when the agent succeeded, and these inferences depended on the box that the agent chose.

ticipant's attention is drawn to the information they can use. We test this possibility in Experiment 2. Experiment 2 was identical to Experiment 1 with the exception that participants were not asked about the covariation data and were just asked to rate the one-shot agents preference, true strength, and perceived strength.

Methods

Participants 100 adult participants (mean age = 35.51; range: 20-70) from the US (as determined by their IP address) were recruited using Amazons Mechanical Turk framework.

Procedure The procedure was identical to Experiment 1 with the exception that people were not asked to judge the weight of each box or the strength of any of the agents in the first part of the story (shown in Figure 1).

Results

As in Experiment 1, results from the experiment were z-scored within response type (preference inferences, weight inferences, and strength inferences) and then averaged across participants.

Figure 4 shows the results from the experiment. As in Experiment 1, the model fit participant judgments with high accuracy (Figure 4a). On the joint inferences about the one-shot agent's preference, true strength, and perceived strength, participant judgments showed a correlation of 0.88 with participant data (95% CI: 0.72-0.95). Consistent with this, participant responses in Experiment 2 resembled the responses

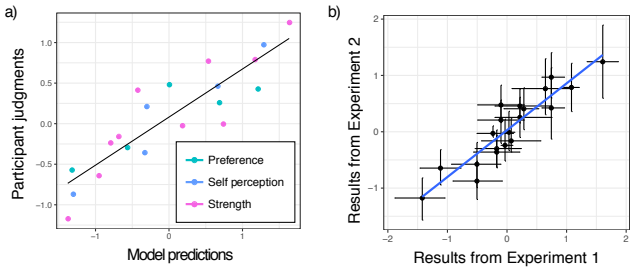

Figure 4: Results from Experiment 2. (a) Model predictions plotted against participant judgments. (b) Comparison of results from Experiment 1 and Experiment 2.

from Experiment 1. Figure 4b shows the comparison between the results in Experiment 1 and the results in Experiment 2. The two sets of data showed a correlation of r=0.92 (95% CI: 0.80-0.97).

General Discussion

Inferences about unobservable qualities of others from single observations are often ambiguous. Across two experiments, we showed that people can rely on past knowledge to make strong inferences about new agents from a single action. Consistent with previous research, Part 1 of Experiment 1 showed that people can decompose ambiguous events into properties of agents and properties of the world by relying on the covariation structure in the data (Kelley & Michela, 1980; Gweon & Schulz, 2011; Seiver et al., 2012). We also showed that these representations about the agents and the physical world support powerful one-shot inferences in future events. People accurately inferred an agent's preferences, their true strength (competence), and the agent's beliefs about her own strength, all from a single event. In Experiment 2, we replicated these results and showed that people spontaneously make use of covariation data in new events. Even when people were not asked to explicitly reason about the covariation in events, they made the same inferences about the novel agent as the participants in Experiment 1.

To test our proposal, we presented a computational model that jointly infers properties of agents through Bayesian inference over a model of utility maximization. This model enabled us to generate quantitative predictions and test participants' relative judgments holistically. Overall, we found that our formalization predicted participant judgments with high accuracy. In our experiment, inferences about the final agent were tested across participants. As such, each participant only watched a single event. Thus, the graded inferences about the properties of the novel agent (see Figure 3) are not judgments that are relative to each other, but rather absolute estimates relative to past experiences.

In our experiments we clarified that the one-shot agent - the square (see Figure 1)- had seen all other agents. This assumption is critical for our model, as its inferences about the agent's mental states -her preference and her perceived

strength- rely on the assumption that the agent herself had some rational estimate of the weight of the boxes.

Intuitively, if the one-shot agent had not seen the covariation information (and was therefore ignorant about the possible weight of the boxes or the strength of the other agents), then her choice would not be as revealing with respect to the strenght of her preference or her beliefs about her own strength. Consistent with this intuition, our model predicts that if the agent did not see the other agents interact with the boxes, participants should continue to infer the agent's true strength as a function of the selected box and the outcome, but they should now infer the same preference independent of the agent's choice, and they should be unable to infer her beliefs about her own strength. Future work may explore this.

Here we focused on cases where participants bring their knowledge about the world (e.g., weight of boxes) to infer properties of a new agent. As discussed in the introduction, people may also bring knowledge about agents they know to infer new properties of the world. Imagine in our paradigm, for example, if people saw the covariation data in Part 1 first, and then in Part 2, one of the agents from Part 1 interacted with a new box. In this case, our account predicts that people should be able to infer the agent's belief about the weight of the box as well as the true weight of the box from that event. Our paradigm can be flexibly adapted to explore this possibility, and future work might test this prediction.

In our experiment, some participants observed the one-shot agent interact with a new box that no one had tried lifting before (the yoyo box). Participants' inferences suggest that they did not have any prior expectations about the weight of this box (see Figure 3). In our experiment, we were clear that all the covariation agents selected the boxes in a fixed order and they only had four coins, explaining why they never tried to lift the yoyo box. If the agents from the covariation stage had freely chosen which box to play with, then their choices would suggest that the yoyo box has a low reward, or that they thought it was too heavy. In future work we may integrate choice reasoning into the covariation stage to test if people can also integrate this information when reasoning about agents.

One open question is whether the type of account that we proposed here is specific to the social domain. Although our model relies on the assumption that agents maximize utilities, much of the model relies on general principles of Bayesian inferences and inductive generalization. The logic behind these inferences -finding the causes of confounded events, and then using this knowledge to infer hidden causes of new events- is likely to be common in non-social tasks as well (Kemp & Tenenbaum, 2009).

In sum, our current work provides a window into the richness and the complexity of how people reason about others. Developmental work on Theory of Mind (Wellman & Cross, 2001), and even tests of Theory of Mind used with adults (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), often rely on inferences about a single, isolated event. However, it is important to keep in mind that we are constantly observing others' actions and their outcomes in the physical world, and reason about other people who act on the same (or similar) physical world. Exploring the social-cognitive mechanisms that underlie our ability to *learn from others* to *learn better about others* is an exciting direction for future research.

Acknowledgments

This work was supported by the Simons Center for the Social Brain, and Varieties of Understanding grant from the John Templeton Foundation.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001, February). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry*, *42*(2), 241–251.

Csibra, G. (2003, March). Teleological and referential understanding of action in infancy. *Philos Trans R Soc Lond B Biol Sci*, *358*(1431), 447–458.

Gweon, H., & Schulz, L. (2011, June). 16-Month-Olds Rationally Infer Causes of Failed Actions. *Science*, *332*(6037), 1524–1524.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016, August). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. B. (2015). The naïve utility calculus: Joint inferences about the costs and rewards of actions. In *Cogsci.*

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

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual review of psychology*, *31*(1), 457–501.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Liu, S., & Spelke, E. S. (2016, December). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014, March). The Child as Econometrician: A Rational Model of Preference Understanding in Children. *PLoS ONE*, *9*(3), e92160.

Repacholi, B., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12–20.

Seiver, E., Gopnik, A., & Goodman, N. D. (2012, September). Did She Jump Because She Was the Big Sister or Because the Trampoline Was Safe? Causal Inference and the Development of Social Attribution. *Child Development*.

Wellman, H., & Cross, D. (2001). Theory of mind and conceptual change. *Child Development*.