Decision
2019, Vol. 6, No. 1, 17–60

2325-9965/19/$12.00 http://dx.doi.org/10.1037/dec0000088

Reasoning About Social Sources to Learn From Actions and Outcomes

Daniel Hawthorne-Madell and Noah D. Goodman
Stanford University

Learning what others know, especially experts, is a crucial shortcut to understanding the world. Other people's actions and utterances are thus a powerful source of evidence. However, people do not simply copy others' choices or stated beliefs; rather, they infer what others believe and integrate these beliefs with their own. In this paper, we present a computational account of the inference and integration process that underpins learning from a combination of social and direct evidence. This account formalizes the learner's intuitive understanding of psychology—or *theory of mind* (ToM)—including attributes such as confidence, reliability, and knowledgeability. It then shows how ToM is the lens used to interpret another person's choices, weighing them against the learner's own direct evidence. To test this account, we develop an experimental paradigm that allows for graded manipulation of social and direct evidence, and for quantitative measurement of the learner's resulting beliefs. Four experiments test the predictions of the model, manipulating knowledgeability, confidence, and reliability of the social source. Learners' behavior is consistent with our quantitative and qualitative model predictions across all 4 experiments, demonstrating subtle interactions between evidence and the attributes of those learned from.

Keywords: social cognition, theory of mind, information integration, Bayesian computational models

Which is the best stock, job candidate, or race horse? When confronted with decisions like these, people are information omnivores, flexibly learning from multiple information sources. Stock investors, for example, look for direct information about a company's strength, scouring the stock's fundamentals. They also look for more indirect, socially-mediated information. What is Warren Buffet's position on the stock? How confident is he? What about a presumably knowledgeable company insider? The choices these actors make represent a rich source of social information, even when they are not intentionally trying to advise or influence others.[1] At its core, social learning involves interpreting intentional actions in the context of the learner's understanding of the situation, her prior beliefs, and direct evidence. A theory of social learning must both explain how a learner interprets an actor's actions and how the results are integrated with nonsocial evidence and prior beliefs. This suggests two core questions of social learning. First, what is the relative impact of social sources? How do people integrate social information with other evidence like direct observations? Do they heavily discount advice as has been previously found (Harvey & Fischer, 1997; Yaniv, 1997; Yaniv & Kleinberger, 2000)? Second, how do learners evaluate and understand a social source? What is the role of each trait attributed to the actor, how are they represented, and how do they interact to determine an action's informativity?

To understand the actions of an actor, people must have a framework for relating people's

This article was published Online First September 14, 2017.

Daniel Hawthorne-Madell and Noah D. Goodman, Department of Psychology, Stanford University.

Correspondence concerning this article should be addressed to Daniel Hawthorne-Madell. E-mail: d.j.hawthorne@alumni.stanford.edu

[1] Although the computational framework we propose is general to multiple kinds of social cues, we focus our discussion on reasoning about intentional actions and refer to the social source with the generic actor. For clarity, we subsequently refer to the *actor* with a male pronoun and the *learner* with a female pronoun.

knowledge, traits, thoughts and behavior—a theory of mind (ToM; Wellman, 1990; Perner, 1991; Premack & Woodruff, 1978). A learner can use her ToM to reason about the informativeness of an actor's intentional actions, inferring the actor's beliefs and the traits that will affect learning. Previous research shows that learners prefer to learn from people who they think acted intentionally (e.g., Carpenter et al., 1998; Olineck & Poulin-Dubois, 2005), confidently (e.g., Birch et al., 2010; Moore et al., 1989; Sabbagh & Baldwin, 2001), and are knowledgeable (e.g., Pratt & Bryant, 1990) and reliable (e.g., Koenig et al., 2004; Birch et al., 2008). We refer collectively to these latent properties as *actor attributes*. The ToM required to account for social learning thus relates beliefs, actions, and actor attributes.

In this article, we address the two core questions of social learning with a computational model that formalizes how learners interpret social information and integrate it with direct evidence, which we tested with a series of empirical studies. First, we review the factors that learners use to identify the informativeness of social information. We then discuss the existing empirical work in which learners have been shown to integrate social information with other sources. Building off these experimental paradigms, we introduce one capable of both quantitatively manipulating an actor's attributes and quantifying the information content of a learner's direct and social evidence. We then introduce our model, which provides a unified account of how people interpret and learn from social information and integrate it with other sources of knowledge; crucially, this model relies on inferences about the actor—his reliability, knowledge, and other attributes. We perform four experiments and show their close fit to the predictions of our model. In the first experiment, we testeded how learners' integration of an actor's choices with their own evidence is affected by the actor's knowledgeability and perceived reliability. The second experiment tested the impact of an actor who also reports his degree of confidence in his action, providing metacognitive information about his choice. The third experiment looks for signatures of spontaneously attributing actors' confidence when it is not explicitly provided. In the final experiment, we manipulated the apparent reliability of the actor directly. Following the experiments, we discuss alternative models and then model two previous experiments to show how our model captures how people learn from and about others. We conclude with a discussion of the model's role in understanding social cognition, learning, and the future empirical and theoretical directions it inspires.

Actor Attributes Guide Social Learning

Learners make use of information from a variety of social cues like a person's testimony, pedagogical actions, advice, and goal-directed actions. Goal-directed actions are the simplest of these cues and therefore provide the clearest domain to study how people infer an actor's attributes from social cues and how these attributes are used to determine from whom to learn fand how to weigh their advice.[2] Exploration of these complementary questions has been undertaken in two separate intellectual traditions. The epistemic trust literature has focused on the early developmental emergence of sensitivity to actor attributes when selecting an actor to learn from (for a review see Harris, 2007; Koenig & Harris, 2005). The judge–adviser system (JAS) literature has explored how social information from actors with different attributes are integrated with a person's own opinions (for a review see Bonaccio & Dalal, 2006).[3] This section provides an overview of the key social factors these independent lines of inquiry have converged upon—knowledgeability, reliability, and intentionality.

Knowledgeability

An actor's knowledgeability is the accuracy of his beliefs about the world. His knowledge is a valuable source of information to the extent that it is independent from, or more extensive than, what the learner knows. His value as an information source could be the result of superior conceptual understanding of the domain, as is the case with Warren Buffet, or he could simply have more information (or unique infor-

[2] Communicative actions, for example, require recursive reasoning to interpret Shafto et al., 2012; Goodman & Stuhlmüller, 2013.

[3] The classic adviser in the JAS literature simply states what they would do akin to the actors in the developmental literature.

mation).[4] Learners therefore need to know the actor's knowledgeability to appropriately learn from him. Previous research has shown that learners are sensitive to a range of cues of an actor's knowledgeability, such as what evidence he has, his past actions, or his claimed knowledge.

An actor's perceptual access is a very clear cue explored extensively in the developmental literature. A particularly clear case of an independently knowledgeable actor is when he has seen something that the learner has not; that is, when the actor has privileged perceptual access. Sensitivity to this kind of knowledge requires understanding that other people have their own perceptual access, which is evident in infants as early as 14 months (Brooks & Meltzoff, 2002, 2005). Understanding that this independent perceptual access can allow other people to form their own beliefs emerges later with the development of a full-fledged ToM (Flavell, 1999; Wellman et al., 2001). Children with a ToM understand that "looking leads to knowing" and correctly identify actors who have had perceptual access to the item in question (Pillow, 1989; Pillow & Weed, 1997; Pratt & Bryant, 1990). Children are also sensitive to the relevance of the perceptual access to the question. For example, when asked about the color of a hidden toy, children preferentially learn from actors who saw the toy over those who felt it (Nurmsoo & Robinson, 2009; Robinson et al., 2011; Robinson & Whitcombe, 2003; Whitcombe & Robinson, 2000).[5]

Claimed knowledge or expertise is a more indirect source of evidence of an actor's knowledgeability. Children prefer to heed actors who claim to be knowledgeable (Birch et al., 2010; Jaswal & Malone, 2007; Koenig & Harris, 2005; Moore et al., 1989; Moore & Davidge, 2009; Sabbagh & Baldwin, 2001), and adults have been shown to be more influenced by actors who claim to be knowledgeable experts (Jungermann & Fischer, 2005). Another indirect source of evidence of an actor's knowledgeability is their past actions. Young children are naturally trusting of older actors, but once an actor makes a mistake they will preferentially learn from others (Birch et al., 2008; Harris & Corriveau, 2011; Jaswal & Neely, 2006; Koenig et al., 2004; Koenig & Harris, 2005). Similarly, adult learners use an actor's past performances to infer their knowledgeability and are subsequently less influenced by poorly performing actors (Yaniv, 2004).

In the absence of information about perceptual access, past actions, or claimed knowledge, learners fall back on more indirect cues. Infants prefer actors that exhibit nonverbal confidence cues (Birch et al., 2010) or accompany their advice with words like "know" instead of "think" (Jaswal & Malone, 2007; Matsui et al., 2006; Sabbagh & Baldwin, 2001; Stock et al., 2009). Adults are also more influenced by confident actors (Phillips, 1999; Sniezek & Buckley, 1995; Sniezek & Van Swol, 2001; Swol et al., 2005; Yaniv, 1997). Children use additional heuristics like familiarity (Corriveau & Harris, 2009), cultural similarity (Hu et al., 2013; Kinzler et al., 2010), expertise (Lutz & Keil, 2002; Sobel & Corriveau, 2010), and age (Jaswal & Neely, 2006), whereas adults defer to actors of greater age, education, life experience, and wisdom (Feng & MacGeorge, 2006).

Reliability

Although reliability is often used interchangeably with knowledgeability in the epistemic trust literature, we argue for a distinct meaning—a reliable actor chooses reasonable actions to achieve their goals, informed by their beliefs. Reliable actors' actions are diagnostic of their knowledge, unlike actors that are noisy decision makers or purposefully deceitful. As discussed previously, preschoolers track actors' past mistakes and prefer accurate, "reliable" actors. Although these actors' previous mistakes could be interpreted as stemming from their lack of knowledge, the true cause is underspecified in many experiments (e.g., Clément et al., 2004; Jaswal & Neely, 2006; Koenig et al., 2004).

There is a fundamental distinction between the possession of knowledge and the reliable use of it, either of which could be the cause of an informant's past mistakes. Although difficult, some experimental paradigms have begun

[4] If there is independent noise in the belief formation process, then the actor's beliefs are informative even is he has observed the exact same information.

[5] Children also appear to be sensitive to the independence of actor's knowledge, which is consonant with Whalen, Buchsbaum, and Griffiths' (2013) recently finding that adults weigh the evidence of actor's with independent information sources.

to dissociate these two potential causes by creating situations where the learner knows the actor is knowledgeable and yet does not act reliably on their understanding (Chow et al., 2008; Poulin-Dubois et al., 2011). They find that learners are sensitive to an actor's reliability, preferentially learning from actors who have reliably used their knowledge in the past.

Intentionality

A final necessary condition for learning from the actions of a knowledgeable and reliable actor is that the action be intentional. Unintended actions are effectively independent of the actor's knowledge. Eighteen-month-olds already display sensitivity to intentionality, selectively learning from purposeful actions over accidental actions (Carpenter et al., 1998; Olineck & Poulin-Dubois, 2005). Infants preferentially reproduced the actions of adults playing with a novel toy when the action was intentional (e.g., adult said "There!" while completing the action) than when it was unintentional (e.g., the adult said "Whoops!" while completing the action). The sensitivity to the intentionality of an action is also observed in adults who make strong inferences about the causal structure of a device when they are presented with evidence from an intentional action but not from identical physical evidence or accidental action (Goodman et al., 2009).

Toward a Cohesive Account of Actor Attributes

We have seen that learners spontaneously reason about an actor's attributes and preferentially learn from the actions of knowledgeable and reliable actors who act intentionally and with confidence. However, we have neither a coherent account of how learners combine these attributes to interpret social evidence nor of how this social evidence is integrated with their other evidence, like direct observations. In the next section, we propose such an account. Our account formally defines knowledgeability, reliability, and confidence in terms of a learner's ToM. It makes specific predictions about the relative importance of these factors and their interactions. We embed this ToM into an integrative framework to describe how learners combine social and direct evidence. The resulting model predicts how a Bayesian learner would (a) use available evidence to characterize an actor, that is, infer how knowledgeable, confident, and reliable he is and (b) integrate social evidence from the actor given the learner's characterization of him. That is to say that the model provides a standard to evaluate whether learners learn from social evidence as much as they "should," given their model of the actor and their characterization of his attributes.

To test our model's predictions requires an experimental paradigm in which the influence of social evidence can be quantitatively measured relative to a well understood reference point (be it other social sources, direct evidence, or prior knowledge). Additionally, the learner's characterization of the actor's attributes must be measured or manipulated. However, previous experimental paradigms have not been clear on one or both points, limiting their ability to rigorously test our models predictions. We briefly review some key experiments in the advice integration literature and describe the predictions of our integrative framework and how carefully measuring the learner's characterization of the actor and measuring his influence relative to a clear reference point could test them.

A number of experiments in the JAS literature explored advice utilization by comparing a learner's initial estimate to her estimate after receiving advice. The learner's initial estimate is effectively a second source of social evidence (a decision made by a person) and is used to calculate how influential the adviser was. A robust finding using this paradigm is that learners exhibit an "egocentric bias," weighing their initial opinions more than they "should" (e.g., Gardner & Berry, 1995; Harvey & Fischer, 1997; Yaniv, 1997; Yaniv & Kleinberger, 2000). However, from the perspective of a learner in these experiments, there is insufficient evidence to decide the extent to which they should rely on the advice. For example, there is considerable uncertainty about both the actor's knowledgeability and their own. Our framework can formally describe the learner's uncertainty about the adviser and herself and represents how a Bayesian learner would integrate the information provided in these experiments (this is further explored in the applied section *Learning from and about advisers in JAS experiments*). It suggests that the seeming underutilization could be due to the substantial

uncertainty the learner has about the adviser, which is compounded by the uncertainty she has about herself. Testing this prediction requires measuring the learner's characterization of both the adviser's and her own attributes. Additionally, using a less complicated reference point would allow for a clear test of whether people truly exhibit an egocentric bias.

Direct evidence, like observations or base rate information, can provide a clearer reference point for social evidence. For example, Birnbaum and Mellers (1983) used base rate information to measure the impact of social evidence while manipulating the adviser's accuracy and bias. They found that learners were sensitive to both. Learners listened to accurate advisers more and discounted advice from biased sources. Birnbaum and Meller's paradigm enabled a clean test of how participants integrated social and direct evidence by parametrically manipulating the adviser's accuracy and the learner's direct evidence. Do they weigh advice as much as an ideal Bayesian integrator would? Although the magnitude of advice integration was well-captured by the Bayesian model, a clear violation of Bayesian reasoning led Birnbaum and Mellers to reject his formulation of an ideal Bayesian integrator. When the adviser endorsed an option that was highly likely given the base-rate, his advice paradoxically reduced the learner's estimate of its likelihood.

An alternative explanation to this surprising result lies in learners' characterization of the adviser in Birnbaum and Mellers' experiment. Although they carefully controlled the adviser's attributes in his experiment, they did not manipulate or measure the adviser's confidence. In the subsequent experiments, we find evidence that learners spontaneously attribute confidence to advisers, and we describe how spontaneously attributing low confidence would result in paradoxical integration of advice. Again, our "learnercentric" perspective that focuses on the learner's characterization highlights a new interpretation of advice integration findings and points to the importance of pinning down the learner's characterization of advisers.

Taking stock of the strengths and limitations of these previous investigations, we can describe the desiderata for an empirical paradigm to test our framework. Social evidence should have a clear reference point to gauge its relative influence—experimentally controlled direct observations. Having both the learner and the adviser's knowledge clearly quantified in terms of direct observations ensures that they are both commensurable and amenable to parametric manipulation. Similarly, other actor attributes should be defined in a graded way and either explicitly manipulated or the learner's unmanipulated estimate should be measured. We now introduce a betting scenario built on these insights, capable of rigorously testing our framework and formalization of ToM.

Social Learning at the Racetrack

Testing how learners combine social and direct information requires an experimental design that quantifies both knowledge sources in the same way. To do this, we use a horse racing scenario in which both the learner and the actor independently observe some number of races. Hence, the knowledgeability of each is simply the number of races seen, which can be individually manipulated. Instead of receiving a direct report of the race results from the actor, the learner only observes the bet placed by the actor. This requires the learner to infer the actor's beliefs from his actions, and combine those beliefs appropriately with the learner's own evidence.

We consider a restricted class of races in which the actor, who we refer to as *Zach*, has followed two racehorses, x and y. Zach observed some number, n_z, of their previous head-to-head races. The learner has seen some number, n_l, of other races from this same rivalry in which horse x has won some number of times, k_l. Although the learner knows the outcomes from his n_l observations, she doesn't know the outcomes of the n_z races that Zach saw. The learner just sees Zach say which horse he bet on, b_z, and how many races he saw, n_z.

Suppose that the learner sees horse x win two out of five races and is told that Zach saw 10 races and bet on horse x in the next race—given the available information (b_z, k_l, n_z, n_l), how should they bet on subsequent races? This is precisely the question of how learners integrate social information with their own observations. Given this experimental formulation of the question, we can construct a computational model that explicitly formalizes how learners'

ToM is used to interpret social evidence and integrate it with other sources of evidence. The model makes quantitative predictions for manipulations of b_z, k_l, and n_z that suggest systematic tests of the model.

A Computational Account of Learning From Others

This section introduces a model of how learners reason about social information and integrate it with other sources of evidence. Specifically, we outline a model of an *integrative learner* who infers a horse's chance of winning future races by integrating observations of its previous performance with the way Zach bets given his observations. We describe the learner's mental model of both sources in terms of probabilistic generative models (Griffiths et al., 2008; Tenenbaum et al., 2011). These models describe the learner's theory about the causal processes in the world giving rise to her observations. For direct observations, the generative model represents the learner's *theory of the domain*, for example, what latent properties dictate a horse's performance (Figure 1a.). The learner's model of an actor's actions represents her *theory of mind*, that is, the learner's prediction of how the actor uses his observations to form beliefs and decide which actions will further his goals (Figure 1d.). Combining these two mental models results in the unified representation used by the integrative learner to learn from social and direct evidence (Figure 1c). Using Bayesian inference to learn about the common latent properties of these two information sources, the integrative learner model represents a coherent reference point for social learning. The integrative

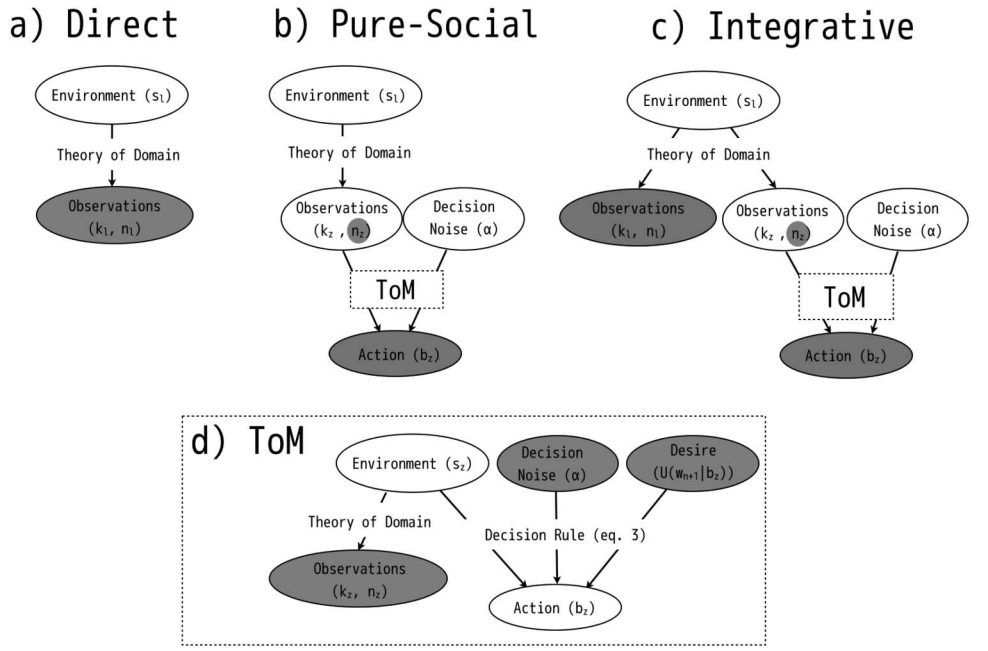

Figure 1. Modeling learning from direct and social evidence in causal graph notation. Random variables are represented as oval nodes, and dotted boxes specify the causal relation between variables. Solid nodes are observed variables; empty nodes represent latent variables whose values must be inferred. Panel a: The "direct learner" who learns about the environment from observations. Panel b: The "pure-social learner" who learns about the environment from an agent's actions. Panel c: The "integrative learner" who combines information from both observations and an agent's actions to learn about the shared environmental cause. Panel d: The learner's theory of mind (ToM). It represents the learner's model of how agents form beliefs and make decisions given observations, beliefs, and degree of rationality.

learner described subsequently is an instance of a general framework for formalizing how social information is interpreted and integrated with other evidence. The framework can be used to describe learning in complicated situations requiring elaborate theories of domain and mind, for example, why your car whines only when turning slightly left and the implication of your mechanic's knowing smile. The simple racetrack scenario elides very little of the general structure and allows for a concise treatment of the integrative learner with clear formalizations of the actor's knowledgeability and reliability. The full integrative learner model can be built incrementally. We first describe the *pure-direct learner*, who updates her beliefs about the world from direct observations. We next introduce the learner's ToM, which combines a belief update model with a decision rule to model how Zach forms beliefs from evidence and then acts on those beliefs. We then show how a *pure-social learner* uses this ToM to learn from Zach's actions. Combining these components, we introduce the full integrative model. Finally, we conclude with a discussion of the predictions of the integrative learner model.

Direct Learner

In the racetrack scenario, two horses, x and y, repeatedly race against one another. By observing a number of races, the learner can gain information about the probability horse x wins the next race. The learning process is guided by the learner's theory of the domain—what causes a horse to win. Learners likely entertain a number of complicated theories about the factors contributing to a horse's performance (jockey, pedigree, betting odds, etc.). However, the restricted information the learner and Zach receive admits of a simple domain theory in which the probability that a horse wins a race w_n is determined solely by their relative "skill," which remains constant. That is to say that the learner's estimate of horse x's skill (s_l) is their estimate of the probability that horse x wins the next race w_{n+1}. Because the evidence is restricted to the results of repeated races between the same two horses with no ordering information provided, s_l is sufficient to describe the learner's beliefs about the domain (with $1 - s_l$ representing horse y's relative skill).[6] Given this framing, races between the two horses are independent and their outcomes can be predicted by flipping a coin weighted by their skill for each of the n_l races:

$$p(k_l \mid s_l, n_l) \sim \text{binomial}(s_l, n_l). \quad (1)$$

Assuming learners use this simple domain theory and Bayesian inference to update their beliefs, their posterior beliefs after seeing some number n_l of observations where horse x wins k_l times is

$$p(s_l \mid k_l, n_l) \propto p(k_l \mid s_l, n_l)p(s_l). \quad (2)$$

where $p(s_l)$ is the learner's prior belief in a horse's chance of winning. Equation 2 represents the direct learner (Figure 1a) who relies solely on direct observations to update her beliefs.

ToM

To interpret Zach's actions, a learner must have a theory of the relationship between Zach's observations, beliefs, goals, and choices—a ToM (see Figure 1d). The first component of the learner's ToM is how Zach learns from evidence. We assume that the learner thinks Zach shares the learner's domain theory ($p(w_n = x|s_z)$) and forms beliefs similarly, by Bayesian inference (Equation 2).

The second component of the learner's ToM captures how Zach makes decisions given his beliefs. We understand Zach as an (approximately) rational decision maker, who (softly) maximizes his utility (in the tradition of Goodman et al. (2009) and Baker et al. (2009)). In the horserace scenario there are no "odds" or differential payoffs for the two horses, making Zach's utility function simply "1" if he bets on the winning horse and "0" otherwise. Which horse Zach thinks will win is specified by his beliefs $p(w_{n+1}|s_z)$. Given this belief distribu-

[6] Indeed, under the assumption that the order of races is unimportant (i.e., that the chance horse x will win m of n races is the same independent of the order of the wins and loses), Finetti's theorem on exchangeable distributions Finetti (1974) implies that the complex theory of the horses' skill can be simplified to a single probability, without losing any information.

tion, an optimal decision theoretic Zach would choose the action, b_z, that maximizes his expected utility $\mathbb{E}_{p(w_{n+1}|s_z)}U(b_z; w_{n+1})$, always betting on the horse he thinks is most likely to win. If the learner thinks that Zach may not be entirely optimal, she may assume that Zach sometimes chooses other actions but does so in a way that higher expected-utility actions will be more likely; we model this as soft maximization, so that the probability of a bet is

$$p(b_z \mid s_z, \alpha) \propto e^{\alpha \cdot \mathbb{E}_{p(w_{n+1}|s_z)}U(b_z;w_{n+1})}. \quad (3)$$

Or, given the simple structure of the racetrack scenario, ($p(w_{n+1} = x|s_z) = s_z$), it is simply

$$p(b_z = x \mid s_z, \alpha) \propto e^{\alpha \cdot s_z}, \quad (4)$$

where alpha is a "rationality" parameter that corresponds to the amount of noise in Zach's decision making. As alpha goes to infinity, the decision rule approaches a strict max-rule where the action with the highest expected utility is always taken. High values correspond to "reasonable" agents whose actions are predictable given their beliefs.[7] When alpha approaches zero, actions are chosen randomly, without regard to their expected utility. Agents with low alpha values are "unreasonable": Their actions are erratic, permitting only weak inferences about the latent causes of the choice. The alpha parameter therefore corresponds to the agent's reliability. We simplify our model by positing that learners either think Zach is "reasonable" with a high alpha or "unreasonable" with a low alpha.

Combining these two components—belief formation (Equation 2) and action selection (Equation 3)—we have the learner's model of how Zach bets given his observations:[8]

$$p(b_z \mid k_z, n_z, \alpha) = \int_{s_z} p(b_z \mid s_z, \alpha) p(s_z \mid n_z, k_z)\, ds_z. \quad (5)$$

Pure-Social Learner

The pure-social learner does not have any direct observations and must therefore rely solely on social information. In our racetrack scenario, the pure-social learner is trying to infer the horse skill most likely to make Zach bet as he did. She uses her ToM (Figure 1d) to reason about Zach's bet. By Bayes' rule,

$$p(s_l \mid b_z, n_z, \alpha) \propto p(b_z \mid s_l, n_z, \alpha) p(s_l). \quad (6)$$

We make the further simplification that the learner reasons using a fixed assumption about Zach's alpha (rather than jointly updating her beliefs about alpha and other unknown quantities), unless there is direct and salient information about Zach's rationality. In our experimental setting this means that alpha is only updated when Zach's observations and actions are both observed. Different learners may have different beliefs about Zach's a priori reasonability, resulting in a population (i.e., mixture) of learners who think Zach is reasonable with $p(\alpha)$ and unreasonable with $1 - p(\alpha)$. Each learner has incomplete knowledge of other latent factors governing Zach's bet. She knows how many races Zach saw (n_z), but she doesn't know the content of his observations (k_z). The learner must therefore consider both what Zach is likely to have seen and his rationality in order to evaluate how likely he is to have made this bet b_z. For a learner attributing Zach rationality α, the probability Zach bets b_z is

$$p(b_z \mid s_l, n_z, \alpha) = \sum_{k_z} p(b_z \mid s_l, n_z, k_z, \alpha) p(k_z \mid s_l, n_z, \alpha) \quad (7)$$

$$= \sum_{k_z} p(b_z \mid n_z, k_z, \alpha) p(k_z \mid s_l, n_z), \quad (8)$$

where the second line follows by simplifying using the independencies evident in Figure 1b (namely, b_z does not depend on s_l once k_z is known). We see by combining Equations 6 and 8 that the pure-social learner can be defined in terms of the learner's theory of mind and domain (Equations 5 and 2) and prior expectations about Zach's rationality and the horse's skill.

The pure-social learner model formalizes how a learner can use Zach's bet to infer information about the world. It includes a nested model of how Zach updates his beliefs and makes decisions (Equation 5), which is inverted

[7] We use the terms *reliable*, *reasonable*, and *rational* interchangeably in our discussion of the model and our experiments. However, in the General Discussion we introduce distinctions between reliable and reasonable or rational.

[8] We assume joint update of beliefs and action plans.

to infer what Zach saw from how he bet. The outer level of the model must again use Bayes rule to infer horse x's skill from what Zach was likely to have seen. Equation 6 therefore requires two normalizations to compute—one for the application of Bayes rule in the inner model of Zach and one for the application of Bayes rule in the outer level. Nested models with inner normalizations emerge from social situations in which people use their ToM to reason about the actions of others (Stuhlmüller & Goodman, 2013).

Integrative Learner

When a social learner who observes Zach's bet also has her own direct observations, how should she combine these two information sources? It depends on her understanding of the causal relation between the sources. As seen in Figure 1c, the integrative learner models both Zach's and her own observations as being generated from horse x's skill (s_l). This common (unobserved) cause allows information to flow between the learner's model of Zach and her own observations. The learner's belief about horse x's skill constrains what she infers Zach saw. Zach's bet, in turn, provides evidence about s_l. This is to say that an integrative social learner forms her belief (s_l) by jointly considering her observations (k_l) and what she infers about Zach's observations (k_z). This requires elaborating the model of the pure-social learner (Equation 6) to jointly consider Zach's bet and the learner's observations. Using Bayes rule and the independence between the learner's direct and social evidence, given s_l,

$$p(s_l \mid b_z, n_z, k_l, n_l, \alpha) \propto p(b_z, k_l \mid s_l, n_l, n_z \alpha) p(s_l) \tag{9}$$

$$\propto p(b_z \mid s_l, n_z, \alpha) p(k_l \mid s_l, n_l) p(s_l). \tag{10}$$

Equation 10 combines elements from both the direct and pure-social learners (Equations 2 and 8). This integrative learner model formalizes how a learner should combine her direct evidence with Zach's bet to estimate horse x's true skill. Like the pure-social learner, the integrative learner model includes a nested model of how Zach makes decisions (Equation 5). However, the top level of the integrative model is more complex, using Bayes rule to infer s_l from both b_z and k_l.

Predictions of the integrative learner. The integrative learner model makes quantitative predictions about the relationship between Zach's attributes and the learner's estimate of horse x's skill. These predictions are made clear by the concise formal definitions of Zach's attributes within the horserace scenario. The number of races observed represents both Zach's and the learner's knowledgeability (i.e., horse x's record in the races) and the content of these observations. The learner's estimate of Zach's rationality is simply the amount of decision noise (alpha) she thinks he has. Finally, the learner understands Zach's confidence (discussed subsequently) simply as his estimate of the likelihood that his chosen action will result in the desired outcome. To explore the predicted impact of each of these actor attributes, we conducted a series of simulations leading to four core testable predictions. All simulations were conducted in the probabilistic programming language, Church (Goodman et al., 2008).[9]

The first three predictions concern the effect of Zach's knowledgeability and reliability on the learner's estimate of horse x's skill. We simulated the learner's estimate of horse x's skill when Zach's knowledgeability and the directly observed horse records were manipulated (see Figure 2).[10] We simulated this for learners who thought that Zach was reliable and for those who thought he was unreliable. The model predicts that all three variables—Zach's knowledgeability, the learner's estimate of Zach's reliability, and the content of the learner's observations—should have an impact on her estimate of horse x's skill. First, the simulation makes the straightforward prediction that, all else being equal, as horse x becomes more dominant in the learner's observations, she should increase her estimate of the horse's skill. Second, the simulation shows that the influence of Zach's bet is determined by his knowledge-

[9] All code is freely available and executable in a modern browser https://probmods.org/v1/lfz.html

[10] In all simulations discussed, we used the same reliability and horse skill parameters. When Zach was reliable, he maximized his utility ($\alpha = \infty$); when he was unreliable, he bet randomly ($\alpha = 0$). The learner had a uniform prior over horse strengths and she imagined Zach had the same uniform prior.

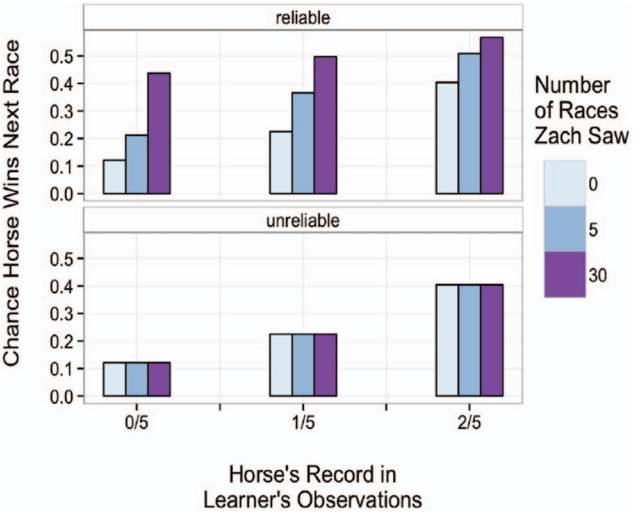

Figure 2. Simulation of the integrative learner model's estimate of horse x's skill (Equation 10), varying the content of her observations, her estimate of Zach's reliability, and Zach's knowledgeability. The learner saw horse x win zero, one, or two of five races, and Zach saw zero, five, or 30 races. Zach always bet on horse x and had an $\alpha = \infty$ when reliable and an $\alpha = 0$ when unreliable. As the figure shows, the more reliable and knowledgeable Zach is, the more influential his bet is predicted to be. See the online article for the color version of this figure.

ability and perceived reliability. When Zach hasn't seen any races, and therefore isn't knowledgeable about horse x's skill, the learner should ignore Zach and rely on her own evidence. The following is the first major prediction of the integrative learner model:

> Prediction 1: When Zach hasn't seen any races, his bet should be ignored. As Zach sees more races (becomes knowledgeable) his bet should increasingly influence the learner's estimate of x's chances.

Zach's reliability should have a similar effect; thus we predict the following:

> Prediction 2: Learners who think Zach does not bet reasonably, that is, has a low rationality parameter, should be less influenced by his bet than those who think he bets reasonably. In the extreme case when Zach is completely unreliable, his bet should be ignored.

As Figure 2 shows, there is an interaction between Zach's reliability and knowledgeability on his influence. Zach must be both reliable and knowledgeable.

> Prediction 3: There should be an interaction between knowledgeability and reliability such that knowledgeability is only relevant to the extent that Zach's bets are diagnostic of his knowledge, that is, he has a high rationality parameter. Conversely, Zach's reliability is only relevant when Zach has some relevant knowledge.

These model predictions are consistent with the finding that children and adult advice-takers learn more from reliable and knowledgeable actors (Harris, 2007; Bonaccio & Dalal, 2006). Thus, we testeded the model's quantitative predictions for these effects. Beyond these factors identified in the epistemic trust and JAS literatures, the model also predicts that a less explored variable should also modulate how knowledgeable learners incorporate information from other's actions: the learner's estimate of the actor's observations.

Learners infer Zach beliefs about horse x's skill from his bet, and understand this belief as a reflection of his observations through Equation 2. Zach's observations, in turn, provide the learner with information about horse x's skill. Indeed, Zach's observations screen off his other

properties from the learner's beliefs. Zach's observations therefore form the link between his actions and horse x's skill. This leads to the following prediction:

> Prediction 4: The learner's estimate of Zach's observations should fully mediate the effect of his knowledgeability.

We tested all four predictions within the horse betting experimental paradigm in Experiment 1. By manipulating the number of observations Zach sees we can test Prediction 1. By asking learners whether they think Zach was betting reliably we can test Prediction 2 and the interaction, Prediction 3. By asking learners to estimate what Zach saw while manipulating his bet and the proportion of horse x wins in their sample, we can test Prediction 4. The models predict that Zach's bet will have the largest impact in inconsistent trials, where the evidence learners receive is inconsistent with Zach's bet. In particular, these situations maximize the difference between the direct, pure-social, and integrative models (assuming for the moment that Zach is reliable). Thus, in the following experiments we focus on inconsistent trials.

Elaborating the Integrative Learner

We next extend the integrative learner model to account for additional sources of information that have been shown to affect social learning. We first elaborate the integrative learner's ToM to allow her to interpret an actor's confidence, a cue shown to have a strong effect on learning (e.g., Birch et al., 2010; Sniezek & Van Swol, 2001). We then turn to describing how the integrative learner can use her ToM to infer Zach's reliability directly from Zach's bet in observed races. More realistically, the learner's mental model allows her to simultaneously learn from *and* about an actor (as is evident early in social reasoner's development; see Kushnir, 2013), across partially observed situations.

Learning from confidence cues. The degree of confidence an actor has in his action is a highly influential cue for social learners. We model the reliable actor's confidence, c_z, as his metacognitive judgment about the likelihood that his action will result in the desired outcome (where $0 \leq c_z \leq 1$). In the horserace scenario, Zach's confidence in his bet is his estimate of the chances the horse he bets on will win, which is equivalent to the probability he bets on the horse:

$$p(c_z = \theta \mid b_z, s_l, n_z, \alpha) = \delta_{\theta = p(b_z \mid s_l, n_z, \alpha)}. \quad (11)$$

Confidence is often expressed with natural language as a gradable adjective, for example, "I am [not/somewhat/very] confident horse x will win." Similarly, we bin the continuous confidence variable c_z into a discrete confidence variable with low, medium, and high levels. A reliable person sends confidence signal σ when his confidence, c_z, falls within the interval $c_{\min}(\sigma) \leq c_z \leq c_{\max}(\sigma)$. We understand Zach's confidence signal as a function of his belief c_z and his reliability alpha vis-à-vis the decision rule described by Equation 4. The probability that Zach with reliability α would send confidence signal σ in their bet b_z is then

$$p(\sigma \mid b_z, s_l, n_z, \alpha) \propto e^{\alpha \cdot \int_{c_{\min}(\sigma)}^{c_{\max}(\sigma)} p(\sigma, c_z \mid b_z, s_l, n_z)\, dc_z} \quad (12)$$

$$\propto e^{\alpha \cdot \int_{c_{\min}(\sigma)}^{c_{\max}(\sigma)} p(\sigma \mid c_z) p(c_z \mid b_z, s_l, n_z)\, dc_z} \quad (13)$$

The second line relies upon the fact that σ is conditionally independent of b_z, s_l, and n_z when c_z is known. Elaborating our integrative social learner with this understanding of how Zach determines his confidence level allows us to model learning from variably confident actors:

$$p(s_l \mid u, b_z, n_z, k_l, n_l, \alpha) \propto p(\sigma, b_z, k_l \mid s_l, n_l, n_z, \alpha) p(s_l) \quad (14)$$

$$\propto p(\sigma \mid b_z, s_l, n_z, \alpha) p(b_z \mid s_l, n_z, \alpha) p(k_l \mid s_l, n_l) p(s_l). \quad (15)$$

Given this formulation, learners know that a reliable and highly confident Zach is more likely to have seen a sample strongly dominated by his chosen horse. Learners are therefore more influenced by Zach's bet when he is highly confident in it.

In Figure 3 we examined the model's predictions for how Zach's confidence in his bet affects his influence on learners. We considered the situation where Zach bet on horse x and has

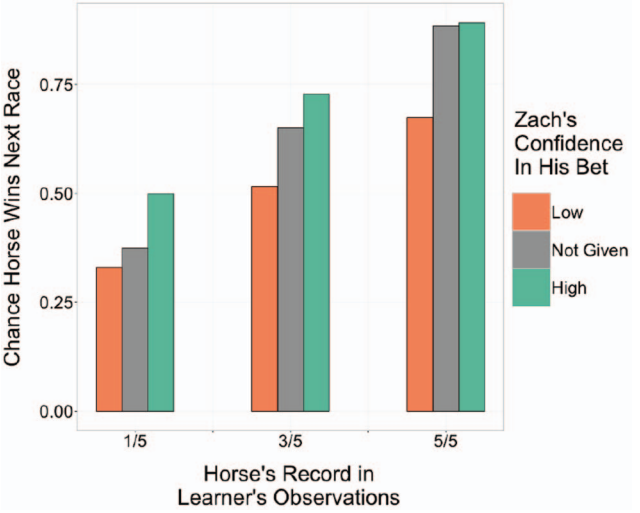

Figure 3. Simulation of the integrated learner model's estimate of horse *x*'s skill when Zach's confidence is varied (Equation 14). The simulation was conducted when Zach saw 10 races and said he had high confidence, low confidence, or did not provide his confidence in his bet. When Zach was unreliable, his confidence did not impact the learner's assessment of horse *x*'s skill in the simulation. However, as the figure shows, when Zach was reliable, the more confident he was, the more influential his bet is predicted to be. See the online article for the color version of this figure.

seen 10 races, varying the amount of direct evidence the learner has and the confidence Zach expresses in his bet. Consistent with previous empirical results (e.g., Sniezek & Van Swol, 2001) the model predicts the following:

> Prediction 5: When Zach is highly confident his bet should be more influential than when he has low confidence or doesn't provide his confidence.

Looking at the left half of Figure 3 where Zach's bet is inconsistent with the learner's evidence, we see that when Zach has high confidence in his bet, the learner is highly influenced by him. Even when the leaner saw horse *x* win only one of five races, she thinks there is a moderate chance horse *x* will win the next one if Zach bet with high confidence. When Zach bet but did not convey confidence, the learner thinks that horse *x* is likely to lose the next race. This seemingly muted influence of Zach stems from the nature of the information he provides, a binary bet. A binary bet only licenses a weak inference about Zach's knowledge; even if Zach were perfectly rational, he could have seen the horse he bet on win every time, half the time, or anything in between.

Looking at the right half of Figure 3 where Zach's bet is consistent with the learner's evidence, we see that Zach's low confidence bets become more influential with the strength of the learner's evidence. When Zach says he has low confidence in his bet, the model thinks it is likely that his bet isn't strongly supported by his evidence, that is, he didn't see horse *x* dominate. In these circumstances, Zach's bet on horse *x* paradoxically *decreases* the learner's estimate of *x*'s strength, thus we predict the following:

> Prediction 6: If Zach states or is interpreted as having low confidence, the learner has strong evidence for an outcome, and Zach bets on that outcome his bet can *decrease* her belief in that outcome. That is to say that when these conditions hold and he bets consistently with her evidence, his bet has a paradoxical effect.

If learners fill in Zach's confidence when he doesn't state it and some non-negligible proportion attributing Zach's low confidence, then Prediction 6 predicts that "paradoxical integration" should be evident in Experiment 1 because confidence cues were not provided. Birnbaum and Mellers (1983) similarly observed this par-

adoxical integration in their 1983 experiment, and used this effect to reject Bayesian models of their data. However, if some of his participants similarly attributed to their advisers low confidence, their findings would be predicted by our Bayesian integrative learner model.

We testeded Prediction 5 in Experiment 2 by having Zach provide either high or low confidence in his bets. In addition to testing Prediction 5, having a condition in which Zach should be highly influential helps distinguish between possible explanations for the predicted modest effect of Zach's bet in Experiment 1. For example, although both our computational account and accounts that find "under-utilization" of advice predict the muted impact of Zach in Experiment 1, our account more naturally predicts the large relative impact of Zach's confidence in Experiment 2.

We testeded Prediction 6 in Experiment 3 where we provide the learner with strong direct evidence that horse x is stronger and then see how Zach betting on horse x influences her estimate. We testeded Zach's influence when he has high or low confidence and compare these results to the corresponding trials in Experiment 1 where Zach doesn't provide confidence information and baseline trials where there is only direct evidence.

Inferring reliability. The generative model underlying the integrative social learner allows reasoning about arbitrary latent properties of the actor. For example, in previous experiments, learners inferred the reliability of an actor from the way they behaved toward known items (e.g., Poulin-Dubois et al., 2011) or named known words (e.g., Clément et al., 2004; Jaswal & Neely, 2006; Pasquini et al., 2007). Did they call a *cow* a *duck* or seem pleased with the contents of an empty box? The equivalent situation in the horserace scenario is seeing how Zach bets when the races he saw are known. Did Zach bet on horse x when she only saw him win one out of 10 races? To infer Zach's reliability, learners consider which alpha value is most consistent with his behavior. The probability that Zach has reliability level alpha is

$$p(\alpha \mid b_z, k_z, n_z) \propto p(b_z \mid k_z, n_z, \alpha)p(\alpha). \quad (16)$$

Although learners could consider Zach having any $\alpha \in \mathbb{R}^+$, we simplify by considering only whether Zach is *reasonable* or *unreasonable*, each corresponding to a particular alpha value.

The model indicates that learners should infer Zach's reliability from the consistency of his action and the information available to him, thus we predict the following:

> Prediction 7: When Zach's observations are more consistent with his bet, learners are more likely to infer he is reliable.

The strength of this reliability inference depends on the learner's model of how Zach bets when he is reliable (a reliable Zach's alpha). If Zach maximizes his expected utility ($\alpha = \infty$), his consistency is highly influential on the learner's estimate of his reliability. However if Zach probability matches ($\alpha = 1$), his consistency is less diagnostic. There is still some chance he could be reliable and bet on horse x when x's record is poor.

We tested Prediction 7 in Experiment 3 by extending the horserace scenario with a *complete information trial* in which the learner sees Zach's observations and his bet. By manipulating the consistency of Zach's bet with his observations we can test the model's quantitative prediction about his reliability inference. We can further test whether learners' updated estimate of Zach's reliability affects Zach's influence in subsequent trials.

Applying the Integrative Learner Model

The use of the integrative model outlined in the preceding text to make quantitative predictions of human behavior requires the calculation of unspecified parameters. These parameters must be filled in by additional hypotheses, left "free" and allowed to take a value that maximizes the fit of the model to the behavioral data it is trying to account for, or on the basis of independent empirical data. Of the three options, using independent data as an empirical prior distribution of the unspecified variable provides the most stringent test of a theoretical framework. We therefore strove to constrain unspecified variables with empirical data wherever possible. We used data from prior experimentation when available and otherwise leveraged the ease of online experimentation to perform additional "prior elicitation experiments" to determine empirical priors. Following this method, no model predictions were the re-

sult of "fitting" to the data they are trying to account for.

In each of the four experimental sections is a model specification subsection that describes the methods used to define the unspecified variables required to make quantitative predictions. The resulting completely specified models were implemented in the probabilistic programming language, Church.[11,12] We used exact inference methods to calculate the posterior estimates. The variability in the resulting model predictions is therefore a reflection of the variability of the empirical priors used, for example, $p(s_l)$.[13]

Experiment 1: Effect of Knowledgeability and Reliability

We set out to test the first four predictions of the integrative learner model by manipulating the direct and social evidence available to participants in our simple racetrack scenario.

Method

For an overview of the flow of the experiment, see Figure 4, and for the details about how the evidence was provided, see Figure 5. The primary stimuli consisted of a table summarizing the results of a series of races between two horses. For each trial, the horses' names were randomly selected from a list of fictional horse names. The questions in the trial were phrased in terms of one of the two horses, henceforth horse x. Horse x's position on the table was randomized, and participants had no indication of which horse was going to be the subject of the eventual query.

The experiment began with an introduction to the betting game. Participants watched an animation of a series of head-to-head races between two horses. As the winning horse passed the finish line, its results were added to a results table summarizing the series. Participants were then told that they were going to see the results of a series of different head-to-head match-ups summarized like the results table they had just seen. They were then given six *baseline trials*. Each baseline trial consisted of a single results table from a different match-up (as pictured in Figure 5c). In the baseline trials, participants saw horse x win each of the possible number of times—zero through five. The order of the baseline trials were randomized. The baseline trials measured the participants' estimate of horse x's skill based solely on their own observations, corresponding to the direct learner model. It was used to calculate their prior beliefs about horse skills.

Following the baseline trials, participants were introduced to the (fictional) last participant to complete the experiment. The last participant was given a random male name (i.e., Zach). Participants were told that Zach "saw different races from the same match-ups," giving him independent information about them. After Zach was introduced, the *test trials* were presented. The six test trials consisted of a result table (Figure 5c), the number of races Zach saw (Figure 5b), and which horse he bet on (Figure 5a). Each test trial was specified by four variables—the number of races the participant saw (n_l), the proportion of times horse x won in those races (k_l), the number of races Zach saw (n_z), and which horse Zach bet on (b_z). The independent variables k_l and b_z were manipulated within participants, whereas n_z was manipulated between participants, and n_l was held constant at 5.

Participants always saw the results of five races (each with five matches). Zach's knowledgeability (n_z) was manipulated by randomly placing participants in one of five conditions in which Zach saw zero, five, 10, 20, or 30 races. In the test trials, participants again saw result tables where horse x won each of the possible number of times—zero through five. In half of these combinations Zach's bet was consistent with the participants' observations, for example, "I saw [horse x] win four out of five races, and Zach bet on him," whereas in the other half, his bet was inconsistent, for example, "I saw [horse x] win two out of five races, and Zach bet on him." Ignoring ordering, there are three unique horse records (zero of five, one of four, and two of three). Participants saw each of the three unique horse records twice for a total of six trials: one trial in which Zach bets consistently with the record and one in which he bets inconsistently.

[11] See https://probmods.org/v1/lfz.html.

[12] The model can be edited interactively to see its sensitivity to the unspecified parameters.

[13] However, the bootstrapped confidence intervals calculated were sufficiently small to be invisible on our plots. This holds true for all subsequent model results.

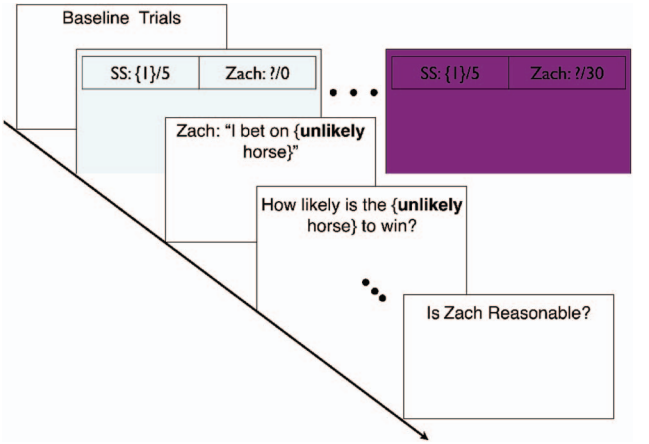

Figure 4. Outline of Experiment 1. The number of races Zach saw was manipulated between participants, from zero to 30. One trial sequence is pictured. Each participant saw six test trials varying horse *x*'s record and Zach's bet. See the online article for the color version of this figure.

The target horse (whose skill was queried) was counterbalanced such that it was not always the weaker horse, and it was not always the horse that Zach bet on. To maintain the plausibility of Zach's inconsistent bets, these trials were not allowed to occur in a row. Three pseudorandom trial orders maintaining this constraint were used.

Following the presentation of all the evidence (see Figure 5), participants were asked to fill in Zach's empty results table with what they thought Zach saw. They clicked on the question marks and freely typed their response (Figure 5b). Then the participants were asked "What do you think is the chance that [horse *x*] wins the next race?" They indicated their response by clicking on a continuum, ranging from *certainly will* to *certainly won't*, which was coded on a scale from 0 to 100, respectively. The evidence—participant s' observed races table and Zach's bet—were presented when making this assessment to ensure that participants' judgments were not limited by memory.

After completing all the test trials, the participants answered a debriefing question ("Which best described Zach?") on a four-point scale, ranging from *very unreliable* to *very reliable.*[14] For an overview of the parameters used in each trial of all the experiments, see Table 1.

Participants

Two hundred participants (mean age = 29 years, 57% female) were recruited through Amazon's Mechanical Turk crowd-sourcing service. They completed the experiment in 6 min to 9 min for a small payment. Forty participants were randomly assigned to each of five conditions in which Zach saw zero, five, 10, 20, or 30 races.

Data Processing and Counterbalance Testing

The data were put into a standard form for analysis. Counterbalanced trials were transformed such that Zach always bet on horse *x* (the horse the participant was asked about). This made all Zach-inconsistent trials occur when horse *x* was dominated in the participants' observed races. Consider a Zach-inconsistent trial in which the participant saw horse *x* win four out of five races, saw Zach bet on horse *y*, and then estimated that horse *x* had an 80% chance of winning the next race. To transform this trial into the standard form, all quantities were inverted such that the participant now observed *x* win one of five, saw Zach bet on horse *x*, and estimated that *x* had a 20% chance of winning.

[14] Participants in the condition in which Zach saw zero races did not answer this question.

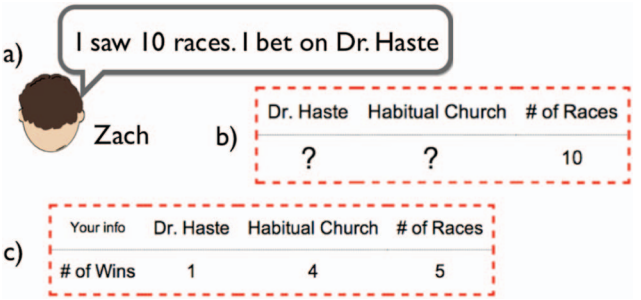

Figure 5. Panel a: Zach telling participant their bet and how many races he saw. Panel b: Zach's obscured result table where participants would click to fill in their estimates. Panel c: Participants' results table. See the online article for the color version of this figure.

The transformed trials were not distinguishable from the untransformed counterbalanced trails ($\chi^2_{df=1} = 7.4 \times 10^{-2}$, $p > .05$).[15] Therefore the results neither significantly depended on which horse was the subject of the query nor on whether the horse that participants were asked about was the same as the one that Zach had bet on. Additionally, the different trial orderings did not significantly influence participants' responses ($\chi^2_{df=3} = 1.8$, $p > .05$, comparing the best-fitting linear model with and without the categorically coded ordering variable).

Model Specification

The integrative model outlined previously (Equation 10) has two unspecified components: The learner's prior over the relative skill of the horses, $p(s_l)$, and the probability that a learner will think that Zach is reasonable, $p(\alpha)$. We discuss the empirical bases for specifying each of these pieces in turn.

We used the nonsocial baseline trials to estimate $p(s_l)$. We bootstrapped an estimate of the participant's mean baseline scores with 1,000 samples. We discretized $p(s_l)$ into a multinomial distribution on 10 evenly spaced points in [0, 1], and for each sampled mean we assigned the points probability such that the mean absolute error between the direct learner model and the sampled mean was minimized (see Figure 6 for fit prior).[16] To lower the dimensionality of this optimization, we added the constraint that the resulting distribution be symmetrical around $p(s_l) = .5$ (a simplification supported by the symmetry of responses seen in Figure 7). We fit our model to the resulting distribution of fitted $p(s_l)$, enabling our model predictions to reflect the variability in this empirically derived prior.

We set the probability of a learner thinking Zach is reasonable, $p(\alpha)$, to the proportion measured in the debriefing question (.74). As described in the model section, we are assuming that participants' estimate of Zach's reliability is constant throughout their trials, not updated within or between trials. This is consistent with the ordering of the trials having no effect (discussed in the previous section—*Data Processing and Counterbalance Testing*). This is not to say that participants are unable to update their beliefs about Zach's reliability, only that they fail to do so from the weak evidence provided in this experiment (i.e., relatively few observed races and no information about what Zach saw). In Experiment 4, we focus on participants' inferences

[15] Chi-square statistic calculated from the difference in deviance of the full model and the (nested) model without the predictor (Pinheiro & Bates, 2009; Barr et al., 2013).

[16] It would have been natural to assume a beta distribution prior on s_l, resulting in a beta-binomial model for the baseline (nonsocial) trials. However, participants' baseline scores deviated from a beta-binomial model, motivating our use of the baseline data to directly estimate a less parametric prior. Participants' baseline scores also deviated from canonical probability weighting functions (Tversky & Kahneman, 1992). The empirically estimated prior exhibits the classic overweighting of extreme probabilities but has high fidelity for moderate probabilities, contradicting both the central tendency predicted by the beta-binomial model and the shallower slope of standard probability weighting functions.

Table 1
Overview of the Manipulated Parameters in the Four Experiments

	Parameters						
	Participant		Zach				
Experiment	Number participant observations	Observed proportion of horse *x* wins	Number of Zach's observations	Observed proportion of horse *x* wins	Chosen horse	Confidence	Total number of trials
Baseline	5	[a]	—	–	—	—	6
1	5	[a]	(5, 10, 15, 20, 30)	—	[a]	—	6
2	5	[a]	10	—	[a]	(low, high)	6
3	5	5/5	10	—	*x*	(low, high)	6
4	5	4/5 → 1/5	10	8/10 → —	(*x*, *y*) → —	—	2

Note. Identical baseline trials were conducted before the test trials of each condition. $x \rightarrow y$ indicates that a trial with parameter *x* is Followed by parameter *y*. (*x*, *y*) indicates that the parameter was manipulated between participants with one half seeing *x* and the other half seeing *y*. — Indicates that participants did not know the value of this parameter.
[a] Indicates that each participant saw Zach betting consistently and inconsistently for each possible horse record.

about Zach's alpha by providing stronger and more salient evidence.

Finally, we fixed the alpha values to describe a "reasonable" and "unreasonable" Zach to empirical values measured in a separate experiment. We elicited 40 participants' intuitive theories regarding how Zach would bet, given his observations and reliability. Participants were randomly split into two conditions. In one, they were told Zach was "reasonable," and in the other, they were told he was "unreasonable." Zach always saw 10 races, and participants were asked how Zach would bet when he saw horse *x* win five, six, seven, eight, nine, or 10 of them. When Zach was reliable, all 20 participants said he would bet on horse *x* as soon as he won a majority of the races (six to ten). This corresponds to imagining that Zach used a strict maximizing rule (a very high α). When Zach was unreliable, participants said Zach would bet on horse *x* exactly half of the time. A logistic regression found no effect of horse *x*'s record on how participants' thought Zach would bet ($\chi^2_{df=1} = .82$, $p > .05$). This corresponds to imagining that an unreliable person just flips a coin to decide who to bet on ($\alpha = 0$), effectively ignoring their observations.

All unspecified parameters of the integrative model are therefore determined by experimental measurements that were independent of the target judgments in integrative trials—thus all parameters were measured rather than fit in the usual sense. We use the same parameters where they are required to model later experiments.

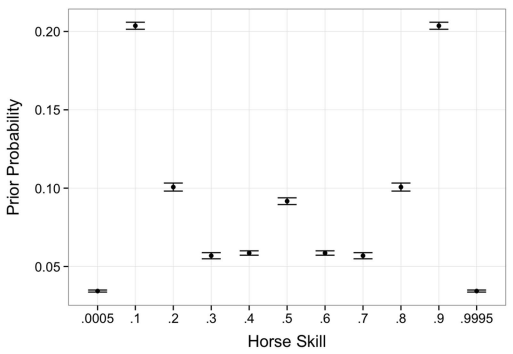

Figure 6. Empirically derived prior distribution over a horse's skill.

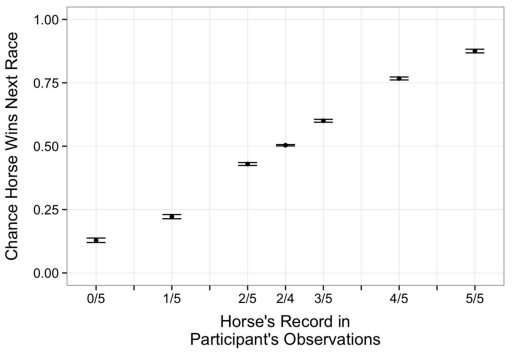

Figure 7. Mean of participants' estimates of horse's chances on the basis of a summary table (i.e., the nonsocial "baseline trials").

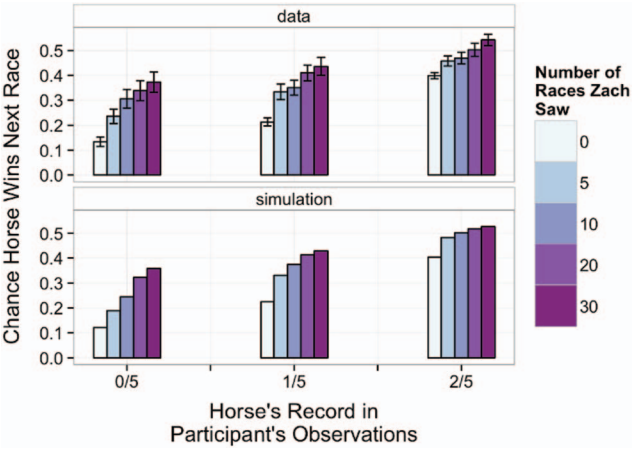

Figure 8. Model's and participants' mean estimate of horse x winning the next race when Zach bet on him, for each directly observed record and level of Zach's knowledgeability. Participants' responses show a clear sensitivity to both Zach's knowledgeability and their direct evidence captured by the integrative model ($R^2 = .96$, $MSE = 6.3 \times 10^{-4}$). See the online article for the color version of this figure.

Results and Discussion

The four predictions from the integrative learner model guide our analysis. All predictions were tested with linear mixed-effects regression models to account for subject-level effects and avoid issues of multiple independent comparisons (Bates et al., 2014; Pinheiro & Bates, 2009). The effect of Zach's bet was predicted to get larger as his bet became more inconsistent with the participants' observations, so we focus our analysis on these Zach-inconsistent trials.[17]

We first verified that the influence of Zach's bet was dependent on the races he observed. We used a planned comparison of when Zach saw no races, when Zach saw 10 races, and the baseline scores. Looking at the inconsistent trials, we would expect participants who were learning from Zach's bet to move their estimate of horse x's chances toward Zach's bet, relative to the baseline scores, which are solely based on their own observed races. If Zach's influence is based on the fact that he saw some races the participant didn't, then we expect him to be influential when he saw 10 races but not when he saw none. We used a mixed-effects model to test the planned comparison. The model also included the participants' observed race results and a (random) term for each participant. In the resulting model, participants' estimates of horse x's chances were more consistent with Zach's bet when he was knowledgeable (saw races) than when he was not (saw no races; $\beta = .13$, $t = 5.0$, $p < .001$).[18] In fact, when Zach saw no races, participants ignored him, responding indistinguishably from those in the baseline trials where Zach was not yet introduced ($\beta = 6.7 \times 10^{-4}$, $t = 2.6 \times 10^{-2}$, $p > .05$). Consistent with the first prediction, participants listened to Zach when he had independent information and ignored him when he didn't.[19]

Figure 8 shows how the integrative learner model and participants respond to the full quantitative manipulation of Zach's knowledgeability. The integrative learner model captures how learners combine their own observations with Zach's bet to infer horse x's chances at specific levels of Zach knowledgeability ($R^2 = .96$, $MSE = 6.3 \times 10^{-4}$). It infers horse x's chances

[17] This is with the exception of the paradoxical integration effect described in Prediction 6 and tested in Experiment 3, which focuses on the maximally Zach-consistent trial.

[18] Denominator degrees of freedom used to calculate p values were approximated using the Satterthwaite method (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2014).

[19] Subsequent analyses focused on exploring the factors modulating Zach's influence will therefore omit the condition in which Zach saw no races, as it was indistinguishable from the baseline trials.

from the number of races Zach observed, the number of times horse x won in his direct observations, its estimate of how many times Zach saw horse x win, and its assessment of Zach's reliability. The best mixed-effects model of the data had precisely these predictors with the addition of a (random) participant term.

Analyzing the mixed-effects model we first see that, like the integrative model, participants used both Zach's bet and their observed races to infer horse x's chances. When x was more dominant in the participants' sample, participants rated him more likely to win ($\beta = .42$, $\chi^2_{df=1} = 7.5$, $p < .001$).[20] The number of races Zach saw was also predictive; as Zach saw more races, participants estimated horse x's chances more in line with his bet ($\beta = 4.9 \times 10^{-3}$, $\chi^2_{df=1} = 19$, $p < .001$). In line with the first prediction, Zach's influence increases with his knowledgeability (see Prediction 1 in Table 2). Whether Zach was thought to be reliable was also predictive (see Prediction 2 in Table 2). The 74% of participants who thought Zach was reliable gave responses more consistent with his bet than did the 26% who did not think he was reliable ($\beta = .12$, $\chi^2_{df=1} = 25$, $p < .001$).[21] Finally, participants' estimate of the percentage of the races horse x won in Zach's observations was also predictive of their estimate of horse x's chances ($\beta = 9.8 \times 10^{-2}$, $\chi^2_{df=1} = 7.5$, $p < .001$). The better they thought horse x's chances were, the more they thought x dominated Zach's observations.

As the model predicted, participants' estimate of Zach's observations mediated the effect of his knowledgeability (see Prediction 4 in Table 2). Including both the number of races Zach saw and the participants' estimate of Zach's observations resulted in a significant effect for his observations ($\beta = 7 \times 10^{-3}$, $t_{477} = 3.5$, $p < .05$), and the previous effect of Zach's knowledgeability is absent ($\beta = 1 \times 10^{-3}$, $t_{477} = .20$, $p > .05$). The Sobel test showed a significant reduction in the effect of Zach's knowledgeability on participants' estimate of horse x's chances ($z = 3.5$, $p < .05$).

Finally, Figure 9 shows that the effect of Zach's knowledgeability is dependent on him betting reasonably in both the data and the integrative learner model (see Prediction 3 in Table 2). The model captures this interaction ($R^2 = .88$, $MSE = 2.5 \times 10^{-3}$). Adding the interaction of Zach's reliability with his knowledgeability to our previous mixed-effects model yields a significant predictor and a better model fit ($\beta = 5.0 \times 10^{-3}$, $\chi^2_{df=1} = 4.6$, $p < .05$). As is evident in Figure 9, participants who said Zach bet unreasonably showed no effect of Zach seeing more races ($\beta = 1.5 \times 10^{-3}$, $\chi^2_{df=1} = .84$, $p > .05$). For participants who said Zach bet reasonably, the more he saw, the more they listened to him ($\beta = 6.4 \times 10^{-3}$, $\chi^2_{df=1} = 21$, $p < .001$).

In the integrative learner model, Zach's observations are the common currency between his bet and what the learner already knows about the world. They form the bridge between the learner's observations and her final estimate of horse x's chances. To test the model's predictions about the crucial role played by Zach's unseen observations, participants were asked to predict what Zach saw. Explicitly posing this question could have induced participants to consider Zach's observations, which they may not have done otherwise. Investigations into this possibility indicate that asking participants to estimate what Zach saw does not influence their estimate of horse x's chances.[22] This implies that learners spontaneously appreciate the importance of what Zach saw and form an estimate of this quantity in the course of reasoning about horse x's chances.

[20] All β values reported are taken from full model, and the chi-square statistic is calculated from the difference in deviance of the full model and the (nested) model without the predictor (Pinheiro & Bates, 2009; Barr et al., 2013).

[21] The proportion of participants that thought Zach was reliable was independent of the number of races he saw ($\chi^2_{df=3} = 1.2$, $p > .05$).

[22] Thirty participants were recruited through Amazon's Mechanical Turk crowd-sourcing service and completed a modified version of Experiment 1. The experiment was identical to the Zach sees 10 races condition, except that participants did not report what they thought Zach saw. The 30 participants were not distinguishable from those in the corresponding Zach sees 10 races condition of Experiment 1. An exact Wilcoxon's Mann–Whitney rank sum test was performed that found that the estimate of horse x's chances was not affected by estimating what Zach saw ($z = .95$, $p > .05$). An additional analysis was conducted in which the best model with the addition of a dummy coded condition variable was applied to this control and the corresponding Experiment 1 data. In the resulting data model, the dummy coded condition was neither predictive of the results, nor was its interaction with participants' observed proportion or reliability attribution (respectively, $\chi^2_{df=1} = 1.8$, $p > .05$; $\chi^2_{df=1} = 2.3$, $p > .05$; $\chi^2_{df=1} = 2.0 \times 10^{-3}$, $p > .05$).

Table 2
Results of the Major Predictions of the Integrative Learner Model Simulations

Prediction	Description	Result
1	The more knowledgeable Zach is, the more influential he should be.	$\chi^2_{df=1} = 16$, $p < .001$
2	The more reliable Zach is, the more influential he should be.	$\chi^2_{df=1} = 22$, $p < .001$
3	Zach's influence is dependent upon him being both reliable and knowledgeable (interaction).	$\chi^2_{df=1} = 4.6$, $p < .05$
4	The learner's estimate of Zach's observations should mediate the effect of his knowledgeability.	$z = 3.5$, $p < .05$
5	When Zach is highly confident his bet should be more influential than when he has little confidence or doesn't provide his confidence.	$\chi^2_{df=1} = 16$, $p < .001$
6	Zach's bet can have a paradoxical effect when he has low confidence and his bet is consist with strong direct evidence.	$t_{238} = 2.74$, $p < .05$
7	When Zach's bet is inconsistent with his observations, learner's should infer he is unreliable.	$\chi^2_{df=1} = 16$, $p < .001$

Note. All statistics are derived from relevant mixed-effects models.

Participants were influenced by Zach's bet, which reflected the knowledge he derived from his observations. However, the influence of his bet, even when he saw six times as many races as the participant, was modest. Interestingly, although this small effect is in line with the model's prediction, it did not resonate with the authors' intuitions: If you saw horse *x* win zero out of five races and learn that Zach would bet on him, how many observations would Zach need to have seen before you would believe him and bet on horse *x* (i.e., estimate horse *x*'s chances greater than .5)? Our initial intuition was that Zach would have to have seen twice as many. This was corroborated by the free responses of 34 Amazon Mechanical Turk responders. The mean number of observations Zach would have to have seen to persuade them was 12.3 when they saw zero of five, with the amount decreasing as their observations became more consistent with his bet. Hence, though participants' behavior conformed closely to that of the normative integrative learner model in Experiment 1, explicit intuitions were highly inaccurate. To put this another way, participants' online behavior was consistent with an agent reasoning rationally over the causal structure of the integrative learner model (see Figure 8, Panels c and d), but the same reasoning was not used to explicitly predict the outcome of the experiment. This décollage is an interesting subject for future exploration.

The full integrative model predicts the modest observed effect of Zach's bet because a single bet only licenses a limited inference about Zach's observations. Although Zach betting multiple times would license a slightly stronger inference, the model predicts that the real bottleneck is the amount of information a binary choice confers (when made at a 50% cutoff point). However, with the addition of simple, low or high confidence ratings, the model predicts that participants should be able to infer what Zach saw with greater precision and, subsequently, weigh his high confidence bets more heavily. Experiment 2 tests this prediction.

In Experiment 3 we examine a surprising result observed in the Zach-consistent trials—we find that Zach's bet on horse *x* causes participants to think that horse *x* winning is less likely than if they had not seen his bet and simply had the same direct evidence. This paradoxical effect of Zach's bet seems to clash with the measured way in which participants integrated his bet on the inconsistent trials, which was well-captured by our Bayesian integrative model. However, this effect is consistent with our model if some learners spontaneously attributed Zach low confidence, similar to how some learners seemed to spontaneously attribute him low reliability. We explore this possibility in Experiment 3.

The 24% of participants who spontaneously thought that Zach was betting unreasonably also contributed to the relatively muted effect. When Zach does not bet reasonably, his action is not diagnostic of his observations. Although the predicted effect is evident in the results (see Figure 9), the data admits an alternative explanation: Participants with low estimates of horse

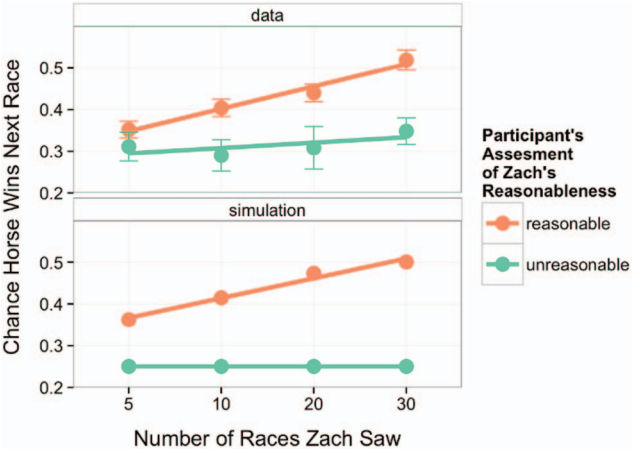

Figure 9. Model's and participants' mean estimate of horse x winning the next race when Zach bet on him, for each level of Zach's knowledgeability and split by participants' assessment of Zach's reasonableness. By marginalizing across the participants' observations, we can clearly see that Zach's knowledgeability and reliability interact—Zach's knowledge is only influential when he is thought to be reliable, as predicted by the model ($R^2 = .88$, $MSE = 2.5 \times 10^{-3}$). See the online article for the color version of this figure.

x's chances could simply be more likely to say that Zach is unreliable, as his bet is dissonant with their estimate. Conversely, those who think horse x's chances are good would be more likely to say Zach is reliable. This alternative explanation would not explain the observed interaction between reliability and knowledgeability, but the nature of the reliability inference merits further study. In Experiment 4, we explicitly manipulate the evidence of Zach's reliability to explore both the source and flexibility of participants' reliability estimates.

Experiment 2: Effect of Confidence

Method

All procedures were identical to those in Experiment 1, with one addition to the stimuli. In Zach's speech bubble, where he stated which horse he bet on, he now also gave his level of confidence—"I bet on horse x with [high/low] confidence." Zach had high confidence for all bets consistent with the participants' information and had moderate confidence on the first trial, which had ambiguous direct evidence (horse x wining two of four races). When Zach's bet was inconsistent with the participant's information, he would either have low confidence in the low confidence condition or have high confidence in the high confidence condition.

Participants

Eighty participants (mean age = 26 years, 59% female) were recruited through Amazon's Mechanical Turk crowd-sourcing service. They completed the experiment in 6 to 9 min for a small payment. Forty participants were randomly assigned to each of the high and low confidence conditions.

Model Specification

For this experiment, the integrative learner model was extended with a straightforward definition of confidence—what Zach thinks his chances of success are, as described by Equation 14. Given this definition we must assess how participants interpreted the utterances, "I bet on horse x with [high/low] confidence," that is, what learners think Zach thinks about the chance that his bet is correct given his stated confidence level. This corresponds to the utterance thresholds used in Equation 14. We conducted a separate experiment with 40 participants to measure this mapping. Participants were shown a (fic-

tional) previous "Zach," who made a bet on a horse with either high, moderate, or low confidence. Participants saw one trial at each level of confidence. No information beyond Zach's confidence was provided. Participants responded by providing the lower and upper bounds of an interval, via two free response boxes, which indicated their estimate of Zach's beliefs about the likelihood of his bet being correct. We calculated the median point at which participants thought Zach would say he had moderate and high confidence, which were 51% and 73%, respectively.[23]

The model used the same horse skill prior and alpha values used in Experiment 1. The probability that Zach was reliable was fixed to the proportion measured in this experiment (.83).

Results and Discussion

Did Zach's high confidence in his bet increase its influence on participants' subsequent estimate of horse *x*'s chances (vis-à-vis Prediction 5)? We tested for this effect by expanding the additive mixed-effects model described in Experiment 1 with a dummy coded confidence variable (and removing the knowledgeability variable because that was held constant in this experiment). As seen in Figure 10, Zach's confidence did have a significant effect ($\beta = .13$, $\chi^2_{df=1} = 11$, $p < .05$), which the integrative learner model captured ($R^2 = .99$, $MSE = 1.6 \times 10^{-3}$). When Zach had high confidence in his bet, it was 2.6 times as influential as when he had low confidence in it (comparing the impact of Zach's bet in each condition to the baseline scores). Compared with the corresponding trials in Experiment 1 in which Zach saw 10 races and gave no confidence information, Zach's high confidence bet was 2.1 times as influential and his low confidence bet was.9 times as influential.

All the significant predictor variables from Experiment 1 were again predictive of participants' estimate of horse *x*'s chances. Participants were receptive to the proportion of times horse *x* won in their five observed races ($\beta = .37$, $\chi^2_{df=1} = 34$, $p < .001$). Their assessment of Zach's reliability influenced how much they incorporated his bet ($\beta = .15$, $\chi^2_{df=1} = 8.7$, $p < .05$), and participants' estimate of Zach's observed proportion of *x* wins was also a significant predictor of his bet ($\beta = .17$, $\chi^2_{df=1} = 4.7$, $p < .05$).

Participants' assessment of Zach's reliability was not driving the effect of confidence: The proportion of participants who thought Zach was reliable did not differ between confidence conditions ($\chi^2_{df=1} = 0$, $p > .05$). However, Zach's perceived reliability did interact with the effect of his confidence, just as his perceived reliability interacted with the effect of his knowledgeability in Experiment 1. Adding the interaction of Zach's reliability and his confidence to the additive mixed model yields a significant predictor ($\beta = .15$, $\chi^2_{df=1} = 4.0$, $p < .05$). The integrative learner model also captures this interaction ($R^2 = .99$, $MSE = 1.6 \times 10^{-3}$ see Figure 11). Looking at the participants who said Zach bet unreasonably, there was no effect of Zach's confidence ($\beta = 8.7 \times 10^{-3}$, $\chi^2_{df=1} = 1.5 \times 10^{-2}$, $p > .05$). For participants who said Zach bet reasonably, his bet had a larger impact when he was more confident ($\beta = .15$, $\chi^2_{df=1} = 12$, $p < .001$).

As the integrative model predicted, when Zach had high confidence he was much more influential, even more iso than if he had seen three times as many races. This finding indicates that the modest effect of Zach's bet in Experiment 1 was not due to a general discounting of social information. The integrative model predicted the small effect of knowledgeability in Experiment 1 from the limited inferences about Zach's beliefs that his binary bet permitted. In the model, the addition of confidence enabled a stronger inference about Zach's beliefs, resulting in his bet being more informative, consistent with participants' responses. Again, participants' estimate of Zach's reliability mediated his influence.

We next examine the impact of Zach's confidence when his bet is consistent with strong evidence. Do his low confidence bets have a paradoxical effect on participants as the model predicts?

[23] For each participant, the upper and lower estimate of adjacent intervals were averaged. For example, the upper bound where Zach would say he had low confidence and the lower bound where he would say he had moderate confidence were averaged to yield the moderate confidence bound. The median of these within subject estimates was used.

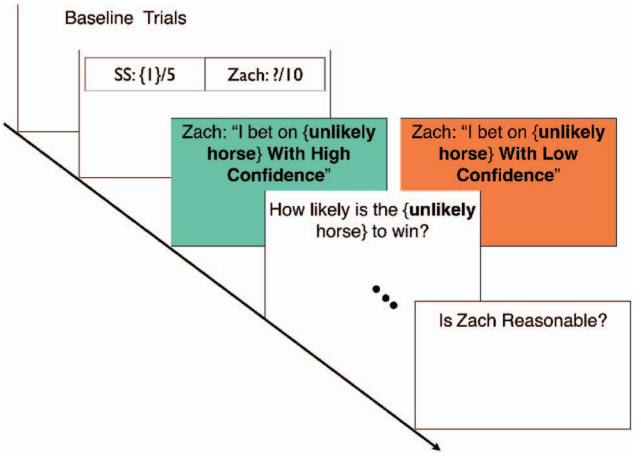

Figure 10. Outline of Experiment 2 and 3. Experiment 2 was identical to Experiment 1 except that Zach always saw 10 races and stated his confidence in his bet. Zach's confidence was manipulated between participants such that he had either "low" or "high" confidence in the inconsistent trials. As in Experiment 1, each participant saw six test trials, varying horse *x*'s record and Zach's bet. See the online article for the color version of this figure.

Experiment 3: Paradoxical Social Evidence Integration

As is shown in Figure 12, participants from Experiment 1 thought that horse *x* had worse chances when they heard Zach bet on horse *x* (with unspecified confidence) than in the corresponding baseline trial (where they had no social evidence). Looking at the figure from left to right, we observe that the "paradoxical" effect of social evidence increased as the direct evidence became stronger with the largest effect in the maximally consistent trials in which participants saw horse *x* win all five races.

This paradoxical updating of beliefs is inconsistent with the integrative learner model's predic-

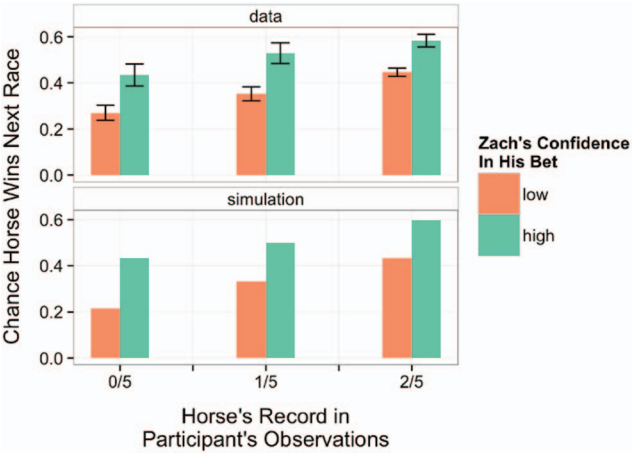

Figure 11. Model's and participants' mean estimate of horse *x* winning the next race at each level of participants' observations when Zach bets on him with either high or low confidence. Participants were influenced by Zach's high confidence bets, as quantitatively captured by the model ($R^2 = .99$, $MSE = 8.1 \times 10^{-4}$). See the online article for the color version of this figure.

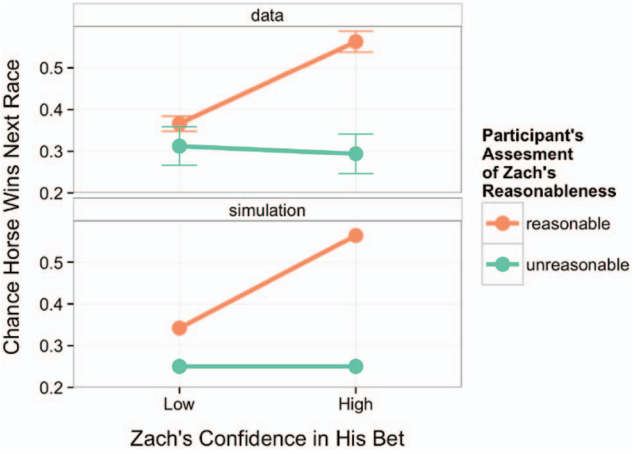

Figure 12. Model's and participants' mean estimate of horse *x* winning the next race when Zach bets on him with either high or low confidence and split by participants' assessment of Zach's reasonableness. By marginalizing across the participants' observations, we can clearly see the interaction or confidence and reasonableness. Only when Zach is thought to be reasonable does his increase in confidence increase his influence, as predicted by the model ($R^2 = .99$, $MSE = 1.6 \times 10^{-3}$). See the online article for the color version of this figure.

tion that seeing Zach's bet would produce a (very modest) increase in the learner's estimate of horse *x*'s chances (see the middle panel of Figure 12). The only way our framework would capture this pattern of reasoning would be if people were automatically thinking about Zach's confidence in his bet and some people thought he bet with low confidence.[24] We therefore tested whether paradoxical integration can be replicated by explicitly providing confidence cues (vis-à-vis Prediction 6) in the maximally consistent trials in which the effect expected to be the strongest. Crucially, the model predicts that the effect should occur when Zach bets with low confidence but *not* when he bets with high confidence, and only if Zach is also thought to be reliable.

Method

The procedure was similar to that of Experiment 2 in which Zach provides his confidence in his bets, but it focused on the maximally consistent trials in which Zach bet on horse *x* and participants had strong direct evidence supporting horse *x* winning (horse *x* won five of the five the trials they saw). The crucial comparison was between the maximally consistent trial in which Zach bets with high confidence and the maximally consistent trial in which he bets with low confidence. It is this comparison that should provide the clearest support for the model's prediction. However, it was important that participants knew the alternative confidence levels Zach could have stated prior to these two trials. Therefore, after the baseline trials, participants first saw three trials in which Zach bet with low, moderate, and high confidence. After these three trials, the two trials of interest were interspersed with other consistent trials to lessen demand characteristics. Thus, participants saw eight pseudorandom trials after the baseline trials. First, they saw the three "alternatives" trials in which Zach bet on horse *x* with each of the confidence levels. For these trials, the direct evidence was randomly selected among horse *x* winning two, three, or four out of five races. Then participants had one of the two trials of interest (randomly selected), which were followed by three randomly selected consistent trials in which Zach bet on horse *x* with randomly selected confidence and direct evidence randomly selected from horse *x* wins three or four out of five. The final trial was the remaining trial of interest.

[24] They automatically reasoned about Zach's reliability in Experiment 1.

Participants

Forty participants (mean age = 26 years, 59% female) were recruited through Amazon's Mechanical Turk crowd-sourcing service. They completed the experiment in 6 min to 9 min for a small payment.

Model Specification

The same model described in Experiment 2, and Equation 14 also applies to this experiment without modification. We used the same empirically derived priors described earlier: the horse skill prior and alpha values used in Experiment 1, and the meaning of the confidence levels from Experiment 2. The probability that Zach was reliable was fixed to the proportion measured in this experiment (.85).

To apply our model to Experiment 1, we similarly needed to determine the probability that Zach was confident $p(\sigma)$. We fixed the probability that Zach was confident to the proportion measured in a separate experiment. This experiment was a modified version of Experiment 1 in which we elicited 40 participants' confidence attributions. The experiment was identical to Experiment 1 except that after the third test trial participants were asked whether Zach had high or low confidence in his bet (a two alternative forced choice). On this random trial, 21 of the 40 participants (53%) said they though Zach had high confidence in his bet.

The resulting model deals with a fully specified Zach with known knowledgeability and with confidence and reliability spontaneously attributed to him. This *fully specified integrative learner model*, like the integrative learner model, has all parameters measured independently.

Results and Discussion

Participants who saw Zach bet on horse x with low confidence exhibited paradoxical integration, believing that horse x had worse chances than when they had not seen Zach's bet ($t_{39} = 2.5$, $p < .05$).[25] Conversely, when Zach bet on horse x with high confidence, there was no significant difference between his bet and the baseline ($t_{39} = -.83$, $p > .05$). We therefore see paradoxical updating from low confidence bets but not from high confidence bets (as expected from Prediction 6). We can see this effect in Figure 13 as the low-confidence point being deflected significantly below the black line representing the baseline score and the high-confidence point.[26]

We further expect that the paradoxical integration should only be evident when participants attribute low confidence to Zach *and* think that he is reasonable (as seen in the interaction in Experiment 2, Figure 11). Indeed, Zach's low-confidence bet on horse x only led to paradoxical integration in the reasonable condition (reasonable, $t_{33} = 3.5$, $p < .001$; unreasonable, $t_5 = -.42$, $p > .05$, as seen in the differential effect of confidence between the columns of Figure 13). The paradoxical integration from Experiment 1 was consistent with this pattern. Only participants who thought Zach was reasonable showed the discounting effect (for all conditions where Zach saw some number of races: reasonable, $t_{118} = 3.8$, $p < .001$; unreasonable, $t_{40} = 1.0$, $p > .05$).[27]

The observed evidence is therefore consistent with paradoxical integration being driven by low confidence attribution. If participants attributed either high or low confidence to Zach in the absence of explicit confidence cues (Experiment 1), then averaging this mixture should yield an intermediate estimate between conditions in which all participants knew Zach had high or low confidence.[28] Indeed, as seen in the "reasonable" column of Figure 13, participants' estimate when Zach's confidence was not stated (the point labeled *Unspecified* representing Experiment 1 data) fell between their estimates when Zach bet with high and low confidence.

In conclusion, the paradoxical discounting observed in Experiment 1 and its interaction

[25] All t tests conducted in this section were testing one-sided predictions derived from our model.

[26] Additionally, we replicate the main effect from Experiment 2 finding that when Zach bet on horse x with high confidence, participants thought horse x was more likely to win than when Zach had low confidence ($t_{39} = 3.9$, $p < .001$).

[27] Interestingly, Zach's reasonableness had the opposite effect than it had for the inconsistent trials in Experiment 1. When Zach is unreasonable and bet on horse x, participants had a *higher* estimate of x's chances that when he was reasonable.

[28] People may, in fact, have a finely graded sense of others' confidence and their initial attributions to them may reflect this, however we simplify our discussion by assuming that participants only thought Zach had high or low confidence.

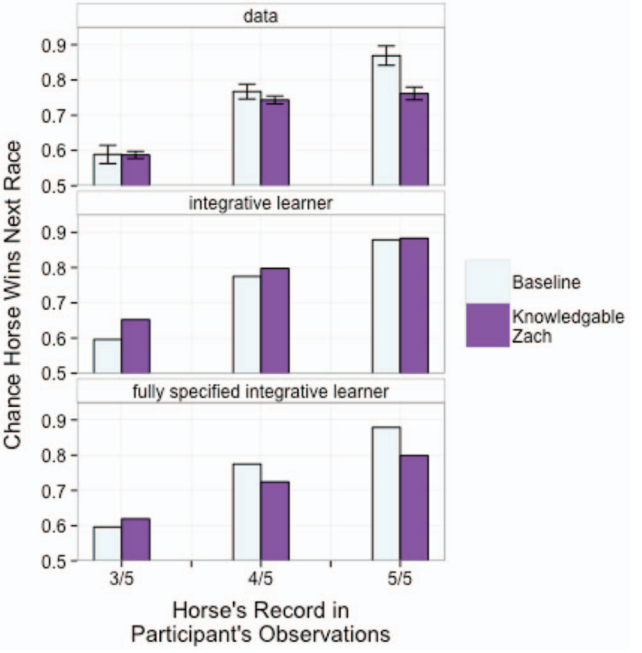

Figure 13. Participant's and model's mean estimate of horse *x* winning the next race when he dominated their observations. The baseline estimates based solely on direct information are compared with the *knowledgeable Zach* estimate, which are the estimates (averaged across all conditions where Zach saw some number of races) from Experiment 1 where Zach bet on horse x, consonant with the participants' observations. Participants' "paradoxical" integration of Zach's bet is evident. They thought horse x was less likely after seeing Zach bet on him. The integrative learner seen in the second panel does not predict this effect ($R^2 = .85$, $MSE = 3.7 \times 10^{-3}$), but the effect is captured by the fully specified integrative learner model ($R^2 = .97$, $MSE = 1.4 \times 10^{-4}$). See the online article for the color version of this figure.

with perceived reasonableness is well described by a model that posits that participants spontaneously fill in both Zach's reliability and his confidence ($R^2 = .90$, $MSE = 9.2 \times 10^{-4}$). Importantly, if we use this model to fit all the data from Experiment 1, then we see that its predictions only significantly diverge from the integrative model (see Equation 10 and Figure 8) when Zach's bet is consonant with very strong direct evidence, which is consistent with our data and Birnbaum and Mellers' (1983) finding. The model that includes learners attributing confidence to Zach therefore fits the Experiment 1 data in Figure 8 when Zach's bet was inconsistent ($R^2 = .94$) and the paradoxical updating observed when Zach's bet was consistent with direct observations.

In our last experiment, we tested whether participants update their initial attributions when provided with strong evidence, inferring Zach's latent states from his actions and the their own knowledge. Will evidence of Zach acting inconsistently with his observations cause people to revise their estimate of his reliability, consistent with our model's predictions?

Experiment 4: Inferring Reliability From Action

Method

Stimuli in Experiment 4 were similar to those in Experiment 1, with one major alteration (see Figure 15). In the first test trial, the *complete information trial*, the participants were able to see Zach's unobscured results table in addition to his bet. Zach always saw horse *x* win eight out of 10 times in this trial, which was consistent with the four out of five wins the participants saw. Participants were randomly placed

into either the consistent condition in which Zach's bet was supported by his observations or the inconsistent condition in which he bets against his observations.

The trial following the complete information trial, the *transfer trial*, was the same for both conditions and identical to the standard test trials in Experiment 1. The participants saw horse x win one of five trials and Zach bet on horse x. The transfer trial was followed by the same debriefing question from Experiment 1 assessing the participants' judgment of whether Zach was reliable.

Participants

Eighty participants (mean age = 31 years, 51% female) were recruited through Amazon's Mechanical Turk crowd-sourcing service. They completed the experiment for a small payment in 4 min to 7 min. Forty participants were randomly assigned to each of the Zach bets consistently or inconsistently conditions.

Model Specification

The integrative model from Experiment 1 was again used with all the same parameters except for the probability of Zach being reliable. In Experiment 1 and 2, the posterior probability of Zach being reliable was fixed to the measured proportion. For this experiment, we let the model infer Zach's reliability given his behavior in the complete information trial, like the participants (as described in Equation 16). The prior of Zach being reliable was set to the average measured proportion from Experiments 1 and 2 (.77).

Results and Discussion

The manipulation of the consistency of Zach's bet led to the predicted assessment of Zach's reliability (see Prediction 7). In the consistent condition, participants said Zach was reliable 88% of the time, compared with just 33% of the time in the inconsistent condition (see Table 3; $\chi^2_{df=1} = 23$, $p < .001$). The integrative learner model makes a similar inference, concluding that Zach is more likely to be reliable in the consistent trial ($MSE = 1.6 \times 10^{-3}$).

Does the assessment of Zach's reliability, inferred in the complete information trial, have an effect on how his bet is integrated in the transfer trial? Using a linear regression model with consistency condition as a predictor of horse x's chances in the transfer trial, we can conclude that the consistency manipulation had the predicted effect on the transfer trial ($t_{78} = 2.1$, $p < .05$). When Zach previously bet consistently, he was subsequently more influential than when he bet inconsistently as the integrative learner model predicted ($MSE = 3.8 \times 10^{-2}$; see Table 4).

In the integrative learner model, the complete information trial influences the transfer trial through the learner's estimate of Zach's reliability. Because Zach's reliability is assumed to be constant across his bets, forming a strong impression that he is reliable or unreliable should influence the amount of information his bets convey in subsequent trials. Indeed, the effect of Zach's (in)consistency in the complete information trial on judgments in the transfer trial should therefore be completely mediated through the manipulation-induced reliability attribution. A mediation analysis concluded that the effect of condition on the transfer trial was mediated by the manipulation-induced reliability attribution (Baron & Kenny, 1986). Including both condition and reliability attribution in the linear model yielded a highly significant effect for reliability ($\beta = .14$, $t_{77} = 3.0$, $p < .05$), and the previous effect of condition is absent ($\beta = .009$, $t_{77} = .20$, $p > .05$). The Sobel test showed significant reduction in the condition's effect on participants' estimate of horse x's chances ($z = 2.6$, $p < .05$).

Participants were more likely to infer that Zach was reliable when his bet was consistent with his information, and they subsequently use this inferred quality to determine how much they should learn from Zach's subsequent bets. It was the inferred reliability of Zach that drove them to listen to him 2.9 times more in the consistent condition than in the inconsistent condition.

Alternative Model Comparison

In this section we assess the empirical support for different key components of the full integrative learner model (Figure 1c; Equations 10, 15, and 17) to simpler models with these components "lesioned" (see Table 5). The integrative learner model posits that learners reason about Zach's bet with a generative model of behavior that describes how Zach's observations influence his beliefs and how his beliefs, desires, and reliability guide his

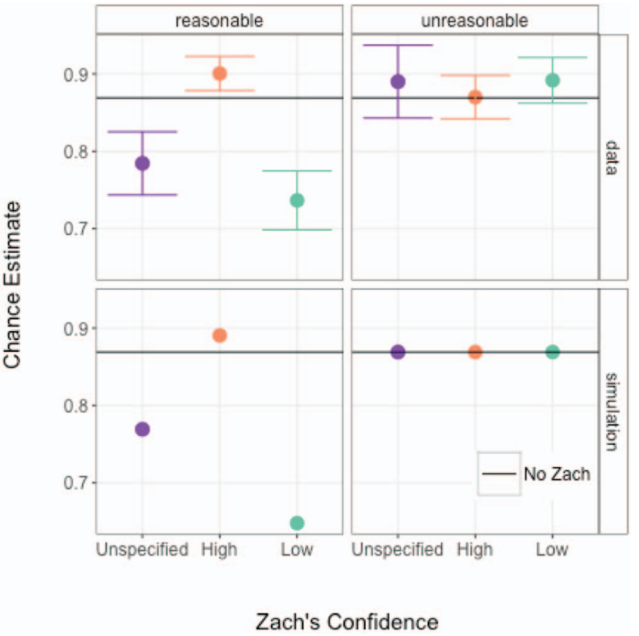

Figure 14. Participant's and model's mean estimate of horse *x* winning the next race when they observe him win his last five and Zach bets on horse x ($R^2 = .90$, $MSE = 9.2 \times 10^{-4}$). Zach saw 10 races and either bet with high or low confidence (Experiment 3) or with unspecified confidence (data from Zach saw 10 races condition of Experiment 1). Results from participants who thought Zach was reasonable are displayed in the left column, whereas those who thought he was unreasonable is displayed on the right. The horizontal black line represents participant's baseline estimate without seeing Zach's bet. The significant decrease in estimates of horse x's chances from participants who heard Zach bet with unspecified confidence compared with the baseline in the first column is an example of "paradoxical" social information integration (the downward divergence of the purple unspecified point from the black line). The model suggests that this is driven by participants spontaneously attributing low confidence and should only occur when participants also attribute reasonability, which is observed in the paradoxical integration of the low confidence estimate in the reasonable column but not the unreasonable column. See the online article for the color version of this figure.

actions (see Figure 1d). It interprets Zach's bet with this intuitive ToM and integrates it with direct observations to learn about the common cause—the horse's skill (see Figure 1c). Are all of the components of the integrative learner model necessary to explain the pattern of results across the three experiments? Specifically, do the data support a model that (a) integrates social and direct evidence, (b) interprets social evidence with a nested inference about Zach and his attributes, (c) spontaneously fills in Zach's unobserved attributes like reliability and confidence such that the Zach is "fully specified", (d) updates these spontaneous attributions in light of strong evidence. We first describe the simplified models that miss one or more of these components and then compare their ability to account for the data to the integrative learner model.

Introduction to Alternative Models

Each model corresponds to a row in Table 5 with a model number (M1, M2, etc.) where its ability to fit the data across our four experiments was assessed. The integration of direct and social evidence of the integrative learner–M7 was contrasted with the nonintegrative direct–M1 and pure-social–M2 models (Figure 1, Panels a and b, respectively). Each of these models exclusively relies on one source of evidence. The direct model simply infers the horse's skill from the direct observations, ignoring social cues,

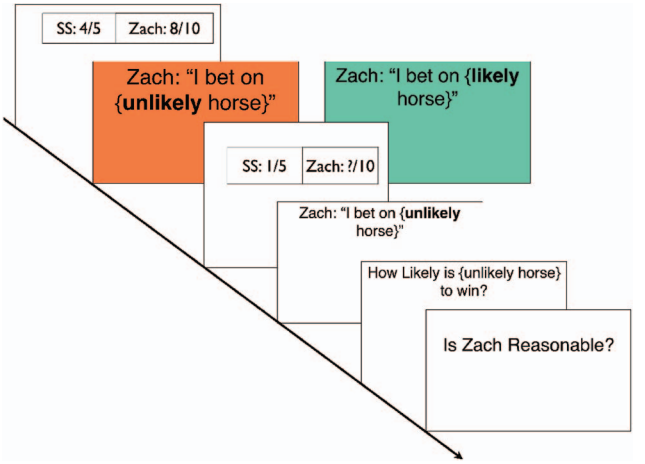

Figure 15. Outline of Experiment 4. Participants first observed the "transparent trial" in which Zach's observations are visible. They then see him bet (in)consistently with his evidence, depending on which (between-participants) condition they are in. Participant's then have a "transfer trial," identical to those in Experiment 1, in which Zach sees 10 races and they see horse x win one of five. See the online article for the color version of this figure.

whereas the pure-social model ignores direct evidence and instead uses Zach's bet to infer the horse's skill. The direct model uses the horse-skill prior empirically fit to independent data, as discussed in Experiment 1's *Model Specification* section. The best pure-social model shares all seven parameters of the integrative learner model, fit to independent data as described earlier.

The fully specified integrative learner–M7 posits that learners fill in Zach's unobserved attributes like his reliability and confidence. We contrast this with simple-reliability–M3 that assumes that learners do not attribute reliability (instead modeling all actors with a fixed degree of decision noise, fit as a free parameter). M7 claims that, for each attribute, there is heterogeneity in the population of learners, for example, some believe Zach is *a priori* reasonable with $p(\alpha)$, whereas others believe that he is unreasonable with $1 - p(\alpha)$. Similarly, some think *a priori* Zach bets with high confidence with $p(\sigma)$. The probability that learners make these attributions—$p(\alpha)$ and $p(\sigma)$—is fixed to the observed proportion in the data. Similarly, the interpretation of high and low confidence and a reliable and unreliable actor are fixed to independently collected data. The fully specified–M7 therefore has seven measured parameters, these six and the horse-skill prior, and no free parameters.

When Zach's confidence and reliability are not observed, the model's characterization of these attributes reflects its priors. The model therefore has to specify when this belief is updated, particularly reliability, which is defined as a stable attribute of Zach's that impacts subsequent decisions. An ideal Bayesian reasoner would always update his or her estimate of Zach's reliability with new evidence as the update-reliability–M6 does. Departing from this ideal reasoner, the flexibly update-reliability–M5/M7 posits that, perhaps because of resource

Table 3
Mean Probability Zach is Reliable With Standard Error of the Mean

Condition	Data	Simulation
Unreliable	.33 ± .074	.17
Reliable	.88 ± .052	.80

Table 4
Mean Estimate of Horse X's Chances With Standard Error of the Mean

Condition	Data	Simulation
Unreliable	.30 ± .023	.26
Reliable	.38 ± .034	.37

Table 5
Comparison Between Learner Models That Make Different Assumptions About the Structure and Content of the Representations Employed by Participants Across the Three Experiments

		Simple model				
Model	Kind	Figure 8	Figure 9	Figure 11	Figure 12	Average fit
1	Direct	.56	0	.38	0	.22
2	Pure-social	.36	.74	.57	.17	.46
7	Integrative	.96	.88	.99	.99	.96

	Integrative model										
	Reliability		Confidence								
	Attribute	Update	Attribute	Figure 8	Figure 9	Figure 11	Figure 12	Figure 13	Figure 14	Tables 2 and 3	Average fit
3	No		No	.96	.45	.99	.54	.85	.35	0	.59
4	Yes	Never	No	.96	.88	.99	.99	.85	.86	0	.79
5	Yes	Flexibly	No	.96	.88	.99	.99	.85	.86	.98	.93
6	Yes	Always	No	.73	.17	.74	.17	.85	.86	.98	.64
7	Yes	Flexibly	Yes	.96	.88	.99	.99	.97	.90	.98	.95

Note. Each column represents the R^2 of the model's predictions for the data represented in the named figure. The data in Figure 8 show the effect of the manipulation of Zach's knowledgeability. The data in Figure 9 Illustrate the interaction between perceived reliability and the manipulation of Zach's knowledgeability and reliability. Figure 11 demonstrates the effect of Zach's confidence, and Figure 12 shows that the dependence is on Zach being reliable. Finally, Tables 2 and 3 show that participants update their estimate of Zach's reliability when presented with clear, salient cues about it and use this updated estimate when learning from him in the future. All the models considered had no free parameters.

limitations, learners only update their beliefs about Zach's reliability when provided with strong evidence.[29] We contrast the ideal update-reliability–M6, the flexibly update-reliability–M5/M7, and the fixed-reliability–M4 that never updates its belief about Zach's reliability. These models share the same representation of Zach's reliability but differ in when this representation is updated.

Results and Discussion

Each model's fit to the three experiments was tested (presented in Table 5). We focused our analysis on the results summarized in Figures 8, 9, 10, 11, 12, and 13 and Tables 3 and 4 for continuity with our discussion in the experimental sections. Across these summaries of the four experiments, the fully specified integrative learner–M7 provided the best average fit with no free parameters. Looking at the pattern of failures of the simpler models reveals the components of M7 that capture different aspects of the learner's social reasoning about Zach and the world.

Participants' responses across the experiments reflected both the races they directly observed and Zach's bet. This integration of social and nonsocial cues is evident in the poor fit of the direct–M1 and the pure-social–M2. Participants were influenced by Zach's bet and showed sensitivity to Zach's knowledgeability (see Figure 8), confidence (see Figure 10), and reliability (see Figures 9 and 11 and Tables 3 and 4) that direct–M1 couldn't account for. Conversely, pure-social–M2 could not account for participants' sensitivity to direct evidence (see Figures 8 and 10). To account for the data we therefore need a model that integrates direct and social evidence like M3 through M7.

Participants' spontaneous assessment of Zach's reliability affected how much they were persuaded by his bet. This effect of reliability

[29] This situation-dependent updating of the learner's estimate of Zach's reliability was motivated by the data in Experiment 1's *Data Processing and Counterbalance Testing* section. The model itself does not specify what circumstances will lead learners to update their estimate of adviser attributes; we make the auxiliary hypothesis that this happens in our setting only when explicit and strong evidence is available. This is clearly an important direction for future work, which is further discussed in the *Reliability* subsection of the *Limitations and Future Work section*.

could not be captured by no-reliability–M3. The model's one free parameter allowed it to capture the main effect of Zach's knowledgeability (see Figure 8) and confidence (see Figure 10), but it couldn't account for the effect of participants' self-reported estimate of Zach's reliability (as is evident in Figure 9 and Figure 11) or the effect of manipulating evidence of Zach's reliability in Experiment 4. All models that had two levels of reliability that were spontaneously attributed to Zach (M3 through M7) were able to capture the influence of participants who have different beliefs about Zach's reliability. Similarly, the models that did not spontaneously attribute confidence to Zach (M3 through M6) were unable to explain the paradoxical discounting evident in Figure 12 and 13. Only fully specified integrative learner–M7 that spontaneously attributed both reliability and confidence could fully account for the data across the four experiments.

Comparing the predictions of M4 through M6—the integrative learner, constant-reliability, and update-reliability models—can reveal when learners updated their estimate of Zach's reliability. Fixed-reliability:M4 captured learners' social information integration across the first two experiments (see Figures 8, 9, 10, and 11). In these experiments the participants' behavior is consistent with the model's assumption that learners do not update their belief about Zach's reliability but do use this fixed belief to interpret Zach's bet. However, the model fails to account for the impact of the transparent trial in Experiment 4 in which participants update their belief about Zach's reliability and then use this updated belief when learning from Zach in subsequent trials (but fixed-reliability–M4 does account for how this updated estimate is used in subsequent trials). On the other hand, update-reliability–M6 cannot account for the participant's estimate in the first two experiments. On the trials in which Zach's bet is strongly inconsistent with the participants' direct observations, update-reliability–M6 infers that Zach is likely unreliable and discounts his bet. Participants' integration does not exhibit this effect, and instead is substantially influenced by Zach on these incongruent trials, as fixed-reliability model–M4 predicts.

The comparison of M4–M6 suggests that learners are capable of updating their estimate of Zach's reliability, but do so only in certain circumstances, like flexibly updating M5. However, the data presented here are insufficient to build a formal understanding of the circumstance in which learners should update their beliefs about latent actor attributes like reliability. One major difference between Experiment 1 and 2, in which participants did not update their belief about Zach's reliability, and Experiment 4, in which they did, was that the evidence in Experiment 4 was more diagnostic of Zach's reliability (according to update-reliability–M6). Given this difference, one hypothesis is that learners are flexible "resource-rational reasoners" who conserve computational resources—only reasoning about reliability when the added inferential complexity results in substantially improved accuracy (Lieder et al., 2012). Alternatively, learners could have a fixed cognitive limitation such that they are unable to perform a joint inference over the horse's skill, Zach's observations, and his reliability. Under this alternative hypothesis, learners update their belief in Zach's reliability in Experiment 4 not because of the strong inference it licensed, but because Zach's observations were known, which simplified the complexity of the inference to Equation 16.

In conclusion, the fully specified integrative learner–M7 provides the best fit to participants social reasoning and information integration across the four experiments (overall fit $R^2 = .96$). Simpler models that did not integrate social and direct evidence or fully specify Zach's attributes could not account for all the major effects in the data. Neither could models that did not selectively update Zach's reliability. The success of M7 is not likely due to overfitting the data. The full model's predictions across the four experiments had seven parameters fit to independent data (as described in each of their *Model Specification* sections), and these fit values were reused across the four experiments with no remaining free parameters (though the choice of which of the four experimental situations led to reliability updates could be seen as a single free parameter).

Insights From a Computational Account of Learning From Others

Learners' flexible social reasoning is evident in their ability to use what they know to infer what they do not. We saw this clearly across our

four experiments; learners used their model of Zach to interpret his actions to learn from him. Similarly, we saw in Experiment 4 that learners could learn about Zach's latent attributes when they had more certainty about what he saw (as Kushnir [2013] stated, learners can learn both from and about people; also see Landrum, Eaves, and Shafto [2015] complimentary treatment of "learning to trust and trusting to learn", p. 1).[30]

The fit of the fully specified integrative learner model across our experiments demonstrate that reasoning over a generative model of action and outcomes naturally captures these flexible patterns of inference. Further, by precisely defining the learner's ToM and what she knows about the situation—her direct evidence, prior knowledge, and theory of domain—our framework describes how learners learn from and about people and the factors that determine how much they learn from and about them. It allows us to distill learning from and about people into two interlocking patterns of social learning. First, learners use evidence judiciously, that is, they integrate different sources according to their diagnosticity. Because an actor's attributes determine their diagnosticity, learners use available information about them to weigh the relative value of his actions. Second, learners use available evidence to infer the unknown attributes of an actor and in the absence of evidence fill these attributes in. Our framework describes how these patterns inform one another—learners use available evidence (e.g., the discrepancy between the learner's initial estimate and the actor's estimate) to infer the unknown attributes of an actor, which, in turn, they use to weigh his actions.

In this section, we apply our framework to two experiments that exhibit these patterns of reasoning. The first experiment by Harvey and Fischer (1997) demonstrated that learners use all available information about an adviser to guide how they integrate that adviser's actions with their own beliefs. The second experiment by Yaniv (2004) shows that when information about an adviser's attributes is not present, learners flexibly use their understanding of the situation to reason about the adviser and how informative his actions are.

Learning From and About Advisers in JAS Experiments

Both Yaniv and Harvey and Fischer's experiments follow the same basic structure commonly used in the JAS literature. First they elicit an initial estimate from the learner. They then show the actor's estimate, and while displaying both the initial and actor's estimate, they elicit a final estimate.

The difference between the learner's estimates before and after receiving advice is their *advice utilization*. Researchers have consistently found that learners under-utilize advice across a range of situations and methods for calculating learners' advice utilization (as described previously in the section *Toward a Cohesive Account of Actor Attributes*). That is to say that learners weigh their initial opinions more than they "should." Here the amount that learners "should" incorporate advice is the amount that would result in the correct answer (e.g., how much the learner should have moved her estimate toward the actor's estimate to arrive at the correct year the Suez Canal opened). We can think of learners that approached this standard as ecologically normative.

However, from the perspective of a learner in these experiments, there is insufficient evidence to decide the extent to which they should rely on the available social information. For example, there is considerable uncertainty about both the actor's knowledgeability and her own. Our framework can explicitly define the learners' understanding of the actors' and their own attributes, as well as the learning situation given the uncertainty of each. The prediction of Bayesian inference over this formalization of the learner's model of the situation provides a new standard with which to evaluate their social evidence integration.

Using our framework to form a normative reference point for advice utilization paradigm we can see if, whereas learners do not incorporate advice as much as they should by the ecological standard, they conform to an ideal observer reasoning over an uncertain learning situation. We form such models for two experiments in the JAS tradition, which suggest that

[30] Or, as we will see in Yaniv's study in the following text, these inferences can happen simultaneously.

what was interpreted as an egocentric bias (from the ecological standard) can be understood as Bayesian reasoning over uncertain learning situations.

The two experiments follow the same experimental structure but differ in the information they provide the learner. Harvey and Fischer (1997) use training on a cue-based judgment task as a proxy for knowledgeability. They manipulated the amount of training given to learners (30, 100, or 240 training trial, between-subjects) and the amount of training the adviser had (within-subjects) such that each learner saw an adviser at each level of training. Harvey and Fischer's experiment provides a test of learner's ability to use an actor's attributes to determine how each cue should be used.

Yaniv and colleagues asked questions tapping historical knowledge (Yaniv 2004). Learners were only told that the adviser was another student participant of the study, which left them uncertain about the adviser's absolute and relative knowledgeability. As Yaniv pointed out, this leads to an epistemic asymmetry as learners have a better understanding of their own knowledgeability than the advisers. Yaniv's experiment therefore tests the framework's prediction that when information about an adviser's attributes is not present, learners flexibly use their understanding of the situation to reason about the adviser and how informative his actions are.

Sensitivity to the actor's relative knowledgeability: Judicious use of evidence.

Modeling Harvey and Fischer (1997). The learner in Harvey and Fischer's experiment is tasked with making final estimate q_f of continuous quantity q on a screen that displays her initial estimate q_l, her number of training trials (proportional to her knowledgeability k_l), the adviser's estimate q_a, and his number of training trials ($\propto k_a$). How should she integrate these two estimates with their associated knowledgeability cues? A learner using the ToM formalized in Equation 5, could reason about each estimate as the intentional choice of a person with belief b and reliability α, generalizing Equation 3:

$$p(q_a \mid b_a, k_a, \alpha_a) \propto e^{\alpha \cdot \mathbb{E}_{p(q \mid b_a, k_a)} U(q_a; q)}, \quad (17)$$

where utility of an estimate $U(q_a; w)$ is given by a squared error loss function and $p(q \mid b, k) \sim \mathcal{N}(b, xk)$, where x represents a free parameter for each level of k describing the increase in precision decision makers associate that level of training.[31] The three values of k that minimize the mean standard error of the model's predictions and the data produces the qualitative model predictions shown in the lower panel of Figure 16.

The learner could then integrate the adviser's estimate with her own initial estimate using a generalization of Equation 6 for each social source s in our case the learner l and adviser a:

$$p(q_f \mid q_l, k_l, q_a, k_a) \propto p(q_f) \sum_{s \in \{l,a\}} \sum_{\alpha_s, b_s} p(q_s \mid b_s, k_s, \alpha_s) p(b_s \mid q_f, k_l) p(\alpha_s), \quad (18)$$

where $p(b \mid q, k) \sim \mathcal{N}(q, xk)$. This model differs importantly from the integrative learner model. It describes learning about a continuous quantity from two social cues (instead of learning about a discrete variable from one direct and one social cue). It also models a situation in which the sources are not understood as updating their beliefs from unknown evidence. The sources beliefs are understood directly in terms of the underlying quantity with precision proportional to the source's training $p(b|q, k)$.

With the limited information about the domain and actor attributes the complex social inferences described in Equation 18 can be well approximated by a simple optimal cue integration model (widely used in perception, e.g., Knill & Richards, 1996; multimodal sensory integration, e.g., Bejjanki et al., 2011; social cognition, e.g., Zaki, 2013):

$$p(q_f \mid q_l, k_l, q_a, k_a) \propto p(q_f) p(q_l \mid q_f, k_l) p(q_a \mid q_f, k_a), \quad (19)$$

where each source's estimate is interpreted as a cue of the underlying q with Gaussian noise proportional to the source's training.

[31] Alternatively, the learner may not be reasoning with her initial estimate shown on the screen, rather she might be directly relying on memory traces of cue. Unfortunately, the data available in this experiment does not allow us to distinguish between this (or a number of other) alternative formalization of the learner's understanding of the situation. That is to say that there are multiple ideal observer models consistent with the experimental situation.

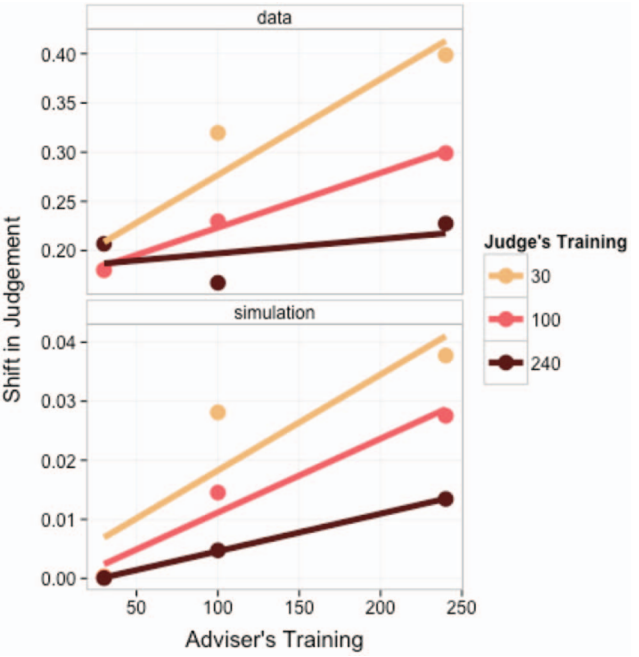

Figure 16. Participant's and model's percentage shift in judgment after advice as a function of participant's and advisers training (data from Harvey & Fischer, 1997, Figure 4a). There is a strong interaction between the participant's level of training and the adviser's level of training on how the participant used the adviser's estimate, which is captured by the model. More highly trained advisers were more influential, and this effect increased as the participants were less well trained. The model captures these qualitative patterns ($R^2 = .92$). See the online article for the color version of this figure.

Results. As seen in Figure 16, participants were sensitive both to their own level of knowledgeability and to their adviser's. Both Harvey and Fischer's (1997) data and our model's qualitative predictions are shown in terms of weight of advice—the difference between the original and final judgment expressed as a percentage of the difference between the original judgment and the advice. As predicted by the model, advice utilization reflects an interaction of the knowledgeability of the two sources ($R^2 = .92$). The more knowledgeable an adviser is, the more influential they are, but as learners themselves become more knowledgeable, the adviser's impact diminishes. When the learner thinks that she has low accuracy interpreting the cue, as in the top line of Figure 16, advisers are highly influential. However, when the learner thinks she can interpret the cue with high fidelity, an adviser (who sees the exact same cue) has muted influence.

Discussion. The model qualitatively captures the interaction of the learner's and adviser's training on advice integration, but the model consistently underestimates the extent to which the learner is influenced by the adviser (seen in the 20% increased utilization seen in all the data points when compared to the model in Figure 16). This is instructive of the differences between the ecological and the Bayesian learner reference points. Participants underutilize the advice in that they would have been more accurate if they had shifted more toward adviser's estimates. However, compared with the Bayesian learner model, learners erred toward overutilization, not underutilization. The Bayesian analysis indicates how learners trying to form accurate beliefs would integrate advice, and the fact that learners exceeded this amount is consonant with Harvey and Fischer's (1997) discussion that there are additional motivational

factors for advice utilization like sharing responsibility and accepting offered help.[32]

Inferring actor attributes from their actions and knowledge of the domain.

Modeling (Yaniv, 2004). Yaniv's experiment focuses on how two factors impact learner's advice utilization: the learner's knowledgeability and the distance of an adviser's estimate from the learner's initial estimate (the advice Δ). Yaniv divided his participants into high and low knowledge groups according to the accuracy of their preadvice judgments (median-split). The distance of advice was binned into near, intermediate, and far. As we shall see, our framework predicts that these two factors should interact, which results from the learner learning both from and about the adviser. This prediction emerges from applying the same model used for Harvey and Fischer's (1997) experiment, except that in Yaniv's experiment the learner does not have any information about the adviser's knowledge and therefore has to rely on internal cues of their own knowledgeability:

$$p(q_f \mid q_l, k_l, q_a) \propto p(q_f)p(q_l \mid q_f, k_l)\sum_{k_a} p(q_a \mid q_f, k_a)p(k_a), \quad (20)$$

where k, although potentially continuous, is discretized into the "high" and "low" knowledge categories Yaniv used in his analyses, and, as in Equation 20, $p(q \mid q_f, k) \sim \mathcal{N}(q_f, xk)$, where x represents a free parameter for each level of k describing the precision learners associate with sources with high and low knowledge.

The resulting model shows how learners can use the advice Δ to infer an adviser's knowledgeability, and how the strength of this inference depends on how accurate they think their estimate is. When the learner is highly knowledgeable and thinks that her estimate is close to the right answer, she can draw a strong inference that an adviser with an implausible estimate is not knowledgeable. She can then discount his estimate. A learner who guessed, and therefore does not think that her estimate was accurate, would not be able to make such a strong inference about an adviser based on the his advice Δ and would utilize advisers similarly, regardless of the distance of their estimate from hers.

Results. As seen in Figure 17, there was a main effect of dividing the learners into low and high knowledgeability groupings. Learners with high confidence were less influenced by advisers, regardless of the advice Δ. This is the same main effect observed in the model of Harvey and Fischer's (1997) experiment where participants had external cues of their knowledgeability, and is consistent with the model's assumption that learners' internal cues allowed for similar (coarse) awareness of their own knowledgeability. Looking beyond the main effect we see the striking interaction predicted by the model. High knowledge learners progressively discount advisers as their advice Δ increases. The advice utilization of low knowledge learners, in contrast, is unaffected by the adviser's advice Δ. The model qualitatively captured this interaction well ($R^2 = .98$).

Discussion. This pattern of results would be expected from learners who reason about an adviser's attributes and use their conclusions to titrate the relative influence of his estimates in their final judgment. The model prediction is therefore dependent upon the learner's estimate of the adviser's knowledge, which should mediate the interactions observed in Figure 17. Yaniv's experiments (Yaniv, 1997, 2004), like others in the JAS tradition, do not measure the learner's estimates of their own attributes or their adviser's. Because learner with high and low knowledgeability were simply grouped by average score it is unknown whether the learners had accurate trial by trial estimates of their own knowledgeability. Alternatively, they could have just had a coarse estimate akin to "I am [good/bad] at these kind of questions."

Learners with a coarse estimate could be systematically underutilizing advice by making poor inferences about advisers when they incorrectly think themselves knowledgeable. Although this learner may be integrating information rationally with respect to their beliefs about

[32] Motivational effects are outside the scope of our present analysis of learning about the world from social and direct evidence, but such an effect could be incorporated into our paradigm by including an explicit utility function for the learner that includes social goals (that are themselves defined using a ToM; see Ullman et. al [2010] for an example).

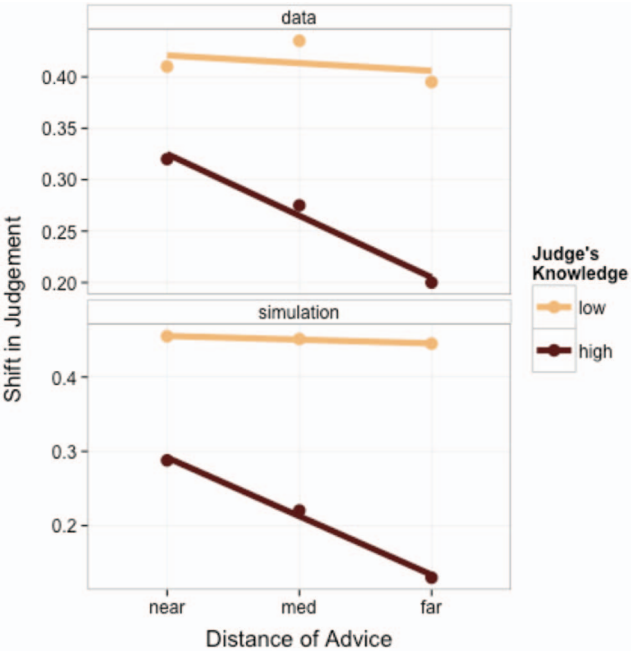

Figure 17. Participant's and model's percentage shift in judgment after advice as a function of participant's accuracy (median split; data from Yaniv, 2004). Knowledgeable participants use adviser's estimates less when the estimates are far from their own initial estimate. Less knowledgeable advisers do not show this pattern, instead using all estimates to the same extent. This interaction is consistent with the qualitative predictions of a model of participants in which they infer the adviser's knowledgeability using their own knowledge ($R^2 = .98$). When they feel knowledgeable, they make strong inferences about the adviser (shown in the lower curve of the graphs). See the online article for the color version of this figure.

the learning situation, their false beliefs lead them to overweight their own evidence. This points to an important limitation of the narrowly defined Bayesian learner we present. Our model simply indicates whether learner's combine information rationally given their beliefs about the situation.

Yaniv and colleagues (Yaniv, 1997; Yaniv & Kleinberger, 2000; Harvey & Fischer, 1997) explain the "discounting" observed in the experiment and others as arising from "an informational asymmetry inherent in any decision-making process that involves the use of advice" (Yaniv, 2004, p. 2). Indeed, in his experiment the learner has no direct evidence about the adviser's knowledgeability, but does seem to have a sense of their own knowledge. Our cognitive model formalizes this asymmetry and predicts the interaction observed in Figure 17.[33]

Conclusion From Application of Framework to JAS Studies

The two JAS experiments highlight two of the hallmarks of social learning. Learners in Harvey and Fischer's (1997) experiment were sensitive to parametrically varied (cues of) the sources knowledgeability. They flexibly used the available evidence about the relative diagnosticity of the two sources—judiciously integrating the two sources. Learners in Yaniv's experiment used their own knowledge to draw inferences about the adviser, filling in his unknown attributes. Our framework shows how understanding social evidence in terms of a

[33] This is an example of how formalizing theories with computational models illuminates the theory's subtle consequences, as Yaniv proposed in the asymmetry theory, but does not predict the observed interaction.

generative model of behavior can predict these patterns of inference.

General Discussion

How do people form a coherent view of the world from both social and direct information sources? Using a novel quantitative experimental paradigm, we found that people integrated social information and direct observations in accordance to the predictions of a sophisticated Bayesian computational model. People displayed a sensitivity to the actor's knowledgeability, reliability, and confidence consistent with our model's qualitative and quantitative predictions.

The core of the computational model is a formalization of an intuitive ToM—people's mental model of how intentional actors make decisions given their beliefs, desires, traits, and other attributes. This generative model of an actor describes what it means for him to be knowledgeable, reliable, and confident, and how these attributes guide reasoning about social evidence. As an explicit theory of the structure of people's ToM, the model makes a number of quantitative predictions about the role actor attributes play in our social reasoning. These predictions build upon previous findings of people's sensitivity to these actor attributes by providing a unified account of their relative impact and interrelationship.

In our computational model, the actor's influence is dependent on the precise extent of his knowledgeability. Experiment 1 confirmed this prediction. It found that people were influenced by actors in proportion to their knowledge and ignored actors with no knowledge. The evidence in our experiments was graded and quantitative, which allowed us to parametrically vary the actor's knowledge. This manipulation revealed that people integrated the (variably knowledgeable) actor's bet with their directly observed evidence as predicted by our Bayesian integrative model.

Experiment 2 found that people were more influenced by confident actors than those who expressed low confidence in their actions. Although this sensitivity to an actor's confidence is well documented, our experiment allowed us to measure the relative importance of an actor's confidence and knowledgeability. Both actor attributes are formalized in our model, providing quantitative predictions for their relative impact and relationship. For instance, the model predicted that an actor who saw 10 races would be more influential if he expressed confidence in his action than if he was three times as knowledgeable but didn't express confidence (from the *Predictions of the Integrative Learner* and *Learning from Confidence Cues* subsections of our *Computational Account of Learning From Others section*). Comparing Experiment 1 and 2, we see precisely this relationship: Confidence dramatically increases an actor's influence compared with increases in his knowledgeability.

Both the effect of Zach's knowledgeability and confidence are predicted to be modulated by his reliability. In the extreme, if the actor is completely unreliable, he should be ignored regardless of his knowledgeability and confidence. Comparing people who thought the actor was reliable to those who though he was unreliable in Experiment 1 and 2 reveals these predicted interactions. In Experiment 1 people who thought the actor was unreliable were not influenced by his knowledgeability. Those who thought the actor was reliable showed graded sensitivity to his knowledgeability. Similarly, in Experiment 2, only people who thought Zach was reliable showed sensitivity to Zach's confidence.

In both Experiment 1 and Experiment 2, the actor's reliability was not mentioned until the end of the experiment, when participants were asked to assess it. The fact that reliability judgments still mediated the effects of knowledgeability and confidence indicates that people spontaneously used their estimate of the actor's reliability to determine how informative his actions were. Similarly, we found evidence that participants spontaneously estimate an actor's confidence and use these estimates when learning from him. In Experiment 1, when Zach's bets were consonant with the participants' strong evidence for horse x, his bet had the paradoxical effect of decreasing their estimate of horse x's likelihood. Zach did not state his confidence in Experiment 1, but this paradoxical effect could be the result of some participants who spontaneously attribute Zach low confidence in his bet. Experiment 3 found that the paradoxical effect observed in Experiment 1 is consistent with participants who fully specify Zach's unobserved reliability and confidence.

Participants who fully specify Zach's unobserved attributes should not take their attributed values as given, but rather update their prior estimates with available information. Departing from a fully Bayesian approach, the behavioral evidence suggests that these estimates are formed and used but not always updated: The best-fitting model used people's prior expectations of the actor's reliability, held constant throughout the experiment, rather than updating in response to (weak, indirect) evidence from the actors actions. However, people did update these estimates in Experiment 4 where we showed stronger evidence—the actor placed a bet that was (in)consistent with his observations providing clear evidence that he was (un)reliable. People used this evidence to infer that the actor was (un)reliable in the proportion predicted by our model, and people used their updated estimate of the actor's reliability when subsequently learning from him. The experiments taken together therefore suggest that people spontaneously attribute actors with attributes like reliability to reason about them, but only update their attributions in certain circumstances. We discuss the implications more in the following passages.

Our model accurately predicted people's responses in all four experiments. The experiments manipulated the actor's attributes and probed people's beliefs about the actor and the world. The same core model was used to account for these manipulations and the different dependent variables, illustrating the productivity of our generative model of behavior (ToM). The model's flexibility emerges from this generative model, not free parameters fit to the specifics of the experiments; our model used the same fitted parameters for all its predictions, leaving none free. As discussed in the *Alternative Model Comparison* section, only the fully specified integrative model that reasoned about both direct and social evidence and filled in unobserved attributes of the actor could account for the pattern of data across these four experiments.

Connections to Other Work

The results of our computational analysis provide a new perspective into two documented irrationalities of advice utilization—that advice can have a paradoxical effect (Birnbaum & Mellers, 1983) and that it is systematically underutilized (e.g., Harvey & Fischer, 1997; Yaniv, 1997; Yaniv & Kleinberger, 2000). Our framework is built around a formal model of the learner's understanding of the adviser. Taking this perspective and formally considering the learner's characterization of the adviser—his knowledgeability, reliability, and confidence—reveals that the inexplicable behavior in both cases could emerge from learner's optimally integrating the social evidence given their characterization of the adviser.

Given the information provided to participants in Birnbaum and Mellers' (1983) experiment and our own Experiment 1, the actor selecting option *x* should have increased the participant's belief in *x*, but it has the opposite, paradoxical effect. Crucially, in neither case was the participant's estimate of the adviser's confidence measured. Our framework describes how this pattern would be expected from a learner who, in the absence of receiving confidence information, spontaneously attributes the adviser confidence. Experiment 3 supported this account of the paradoxical effect and could similarly explain the effect in Birnbaum and Mellers' experiment (which could be tested with manipulation of the adviser's confidence or measurement of participants unmanipulated attributions).

Similarly, our model indicates that what has been characterized as underutilization can be understood as Bayesian reasoning under uncertainty about the adviser's knowledgeability, reliability, and confidence. Indeed, in many of the classic experiments finding underutilization, the source or extent of the adviser's knowledge was uncertain for the learner, which our model predicts would significantly reduce the potential influence of the adviser (Yaniv, 1997; Yaniv & Kleinberger, 2000). However, we have shown that even knowing the adviser's exact knowledgeability, as people did in Experiment 1, does not ensure that he is highly influential.

This points to a strength of our computational approach: it can predict the precise cause of the low impact (seeming underutilization) of advisers—what the "bottleneck" in a given learning situation is. For Experiment 1, our model predicted that it was the nature of the actor's action—a discrete, binary choice—and the uncertainty about his reliability that limited his influence. To increase the adviser's influence, we widened the communication channel in Ex-

periment 2 and found the predicted jump in the adviser's influence. Similarly, our model can make predictions of what would further weaken an actor's influence. By providing strong evidence that the actor was unreliable in Experiment 4 we found that he was effectively ignored. Our modeling framework therefore provides a tool for both generating realistic learner-based normative reference points for advice utilization and for making predictions about which changes to a situation or adviser will have the largest impact on advice utilization. Both applications can buttress existing investigations in the advice-utilization and decision support systems literatures (Gönül et al., 2006; Harvey & Fischer, 1997).

Our framework uses the tools of computational cognitive psychology to synthesize two distinct intellectual traditions—epistemic trust and judgment adviser systems. The framework formalized the actor attributes identified in the epistemic trust literature and combined it with the insights and empirical methodologies for testing advice utilization of the JAS literature. The framework was able to captures some major findings in the respective fields, but a number of questions central to the epistemic trust and JAS literature remain to be explored. The developmental trajectory of the integration of social information with other evidence remains open. Similarly, we haven't addressed the quantitative dynamics of integrating information from multiple actors. Despite their different focuses, the epistemic trust and JAS fields pursue complimentary questions—how do people decide whom to trust, and how do they decide how far to trust them relative to other information sources. Our framework highlights the common cognitive substrate involved in these questions. As such we hope that our framework for understanding social information integration fosters connections between these complimentary literatures and fosters a broader dialogue about the social and cognitive dynamics underlying social learning.

Our computational framework uses a general probabilistic learning framework (Tenenbaum et al., 2011), sharing and elaborating on features of a number of previous accounts of social reasoning and learning (Baker et al., 2009, 2011; Goodman et al., 2009; Shafto et al., 2012). Our model shares the basic model of ToM as a formalization of belief-desire psychology with Baker et al. (2009, 2011) and Goodman et al. (2009; see Butterfield et al., 2008 for a synergistic approach using Markov random fields). However, this basic conception of ToM is elaborated in a number of ways. Our account used a graded notion of knowledgeability which allowed it to reason about actors with different epistemic relationships to the world (whereas Goodman et al. (2009) and Baker et al. (2009, 2011) assumed agents were completely knowledgeable). Our account also relaxed a foundational assumption of previous accounts—that agent's act rationally—allowing learners to reason about the extent to which an actor was reliable. Finally, our model formalized and explored the effect of an actor's confidence. We used this elaborated ToM to explain both how people learn about the world from actors, like Goodman et al. (2009) and Shafto et al. (2012), and how they infer latent states of the actor, like Baker et al. (2009, 2011) and Pantelis et al. (2014). The resulting model provides a unified account of how actor attributes guide social learning, bridging the epistemic trust and judgment adviser systems literatures (as Eaves and Shafto's [2012] model provided a unified account of pedagogical reasoning and epistemic trust).

Limitations and Future Work

Our experimental scenario provided a simple domain for measuring social information integration where the situation makes the nature of the actor's attributes transparent, for example, knowledge is simply the number of races observed. The scenario enabled a number of simplifying assumptions in our integrative learner model and allowed us to cleanly test the effect of actor attributes on social learning. Exploring more complicated situations where these assumptions no longer hold may expose additional complexity of people's naturalistic social reasoning. Each of these simplifications therefore represents an interesting avenue for further research; we next describe several of these possible extensions.

Amount of direct evidence. Across the four experiments the participants' knowledgeability from direct evidence, that is, the number of races they observed, was held constant. The model predicts that the actor's influence should diminish as the learner has more direct observations, consistent with Harvey and Fischer's (1997) findings. The model also predicts an interaction between

the effect of the actor and learner's knowledgeability: An actor's knowledgeability should have greater impact for less knowledgeable learners. Extending beyond Harvey and Fischer, the model predicts that learners with more knowledge would be able to draw stronger inferences about latent actor attributes, like reliability. For example, in Experiment 1, when the learner only had a few observations and the actor bet on the horse that was dominated in them, it did not license a strong inference. From the learner's limited observations, it was not unlikely that the actor saw a conflicting sample and was therefore acting reliably. However, a learner that saw many observations would have a better estimate of what the actor was likely to see, and his inconsistent bet would license a much stronger inference that he is unreliable. These predictions provide a further empirical test of the integrative learner model.

Reliability. The general conception of reliability is likely richer than was exposed in our experimental set-up. For instance, gradation in reliability was limited by the four-point scale we used to assess it. We then further binned responses into two categories—people who thought the actor was reliable or unreliable—to simplify the analysis. However, people's use of the scale and integration behavior indicated that their representation of reliability was more graded than the two simple categories. Additional explorations will be needed to determine how fine-grained reliability is. More importantly, our model treated reliability as simply the noise of a person's decision rule. It is likely that people also consider other ways in which actors could be (un)reliable, such as the way they form beliefs. We were unable to distinguishing between these two sources of (un)reliability from the limited evidence that people received in our experiments, but other paradigms could distinguish between these, and other, forms of reliability. There is likely a rich intuitive theory of reliability itself to be explored. For example, the actor could have a terrible memory and therefore could be reliable for recently acquired knowledge but uninformative at greater temporal remove. These more nuanced ways of being unreliable could be formalized in more elaborate models of mind; incorporating these more elaborate models of mind into an integrative social learner would highlight the unique behavioral signatures of each trait.

Another question suggested by our results is when learners update their attribution of traits to an actor. Across our four experiments, learners only updated their estimate of the actor's reliability in Experiment 4. One explanation for this is that the evidence provided about the actor's reliability was relatively strong (from our computational model's predictions) and salient (from the authors' intuition) relative to the evidence provided in Experiment 1 and 2. An alternate explanation identifies the more complex inference required to infer the actor's reliability from the evidence provided in Experiment 1 and 2 (see the *Alternative Model Comparison* section for further discussion of these theories). Both of these may amount to resource-efficient processing strategies for approximating a complex fully-Bayesian belief update. The current evidence cannot determine which explanation best accounts for the differential reasoning about reliability across our experiments. Future investigations are therefore necessary to determine the cause of this pattern of reasoning and interrelated questions like whether the effect of the strength and salience of evidence is graded or whether there is a threshold at which participants will begin to reason about (rather than just using) reliability and other traits.

Prior knowledge. The learner's understanding of the actor's knowledge and prior beliefs was constrained in our experiments. This allowed us to adopt a simple model of knowledge that will not necessarily hold in general. For example, we assumed learners thought the actor shared their prior distribution over horse skill.[34] We also assumed common knowledge of the causal structure of the domain, which meant that the actor's observations were the sole source of his knowledge. Although this correspondence is desirable for parametrically manipulating social evidence, people often learn from actors who have a different understanding of the domain. In fact, superior understanding of the domain is often precisely why others' opinions are sought, not their privileged evidence—even though Warren Buffet looks at the same stock fundamentals as everyone else, his bets are still

[34] Although expedient, it is unclear when such assumptions of correspondence between a learner and their model of an actor is justified. Are actors assumed to be "like me" until proven otherwise, or is there a separate default used for actors (Meltzoff, 2007)? This relationship between a learner's conception of herself and her understanding of others is a central question in social psychology (Malle, 2004; Ross, 1977). By explicitly representing the learner's beliefs and her beliefs about others, our modeling framework provides a new tool with which to approach this classic question.

highly influential. This situation corresponds to those often explored in the JAS literature, which has found that advice utilization is variable and dependent on a number of heuristic actor attributes like age and experience (Feng & MacGeorge, 2006). Developing a formal theory of how learners understand these ambiguously knowledgeable actors, and attribute knowledge on the basis of other traits, represents an important elaboration of the simple ToM we presented.

Communication. Finally, as we saw in Experiment 2, the nature of the actor's communication has a large impact on the influence of their message. In our experiments, we restricted the "communication" to an actor's goal-directed action, which is equivalent to watching the actor interact directly with the world (perhaps hesitantly or with confidence). Although our actor interacted directly with the world, actors are often learner-oriented, interacting with the learner with communicative, pedagogical, or nefarious intent. Learning from a teacher requires a model of how they make pedagogical decisions because the teacher's goal is not about the state of the world. It is about the beliefs of the learner. This requires the learner to think about the teacher who is, in turn, thinking about the learner. This more elaborate ToM has been formalized and tested by Goodman and Shafto (Eaves & Shafto, 2012; Shafto & Goodman, 2008; Shafto et al., 2012, 2014) in the "pure-social" case in which learners do not have alternative sources of information. Incorporating their more elaborate ToM into the integrative learner model would provide a good test of the extensibility of our framework. The resulting model could provide insights into how students learn from a mix of instruction and exploration. Does a student's self-confidence get in the way of her learning? What can teachers do to manage the student's perception of their attributes, like reliability and knowledgeability?

Conclusion

Other people represent an important source of information, but learning from them requires an understanding of how they tick—what they know, how they choose actions, and what they want. Formal models of ToM are therefore crucial to understanding social learning. We presented a highly extensible computational framework for modeling a learner's understanding of social sources and showed how it formalizes the integration of social and direct evidence by describing the three major attributes of social sources—their knowledgeability, confidence, and reliability. We found that this model explained the quantitative data from our horserace paradigm extremely well. Such rigorous models of social learning provide a framework for both understanding previous empirical investigations and generating future experiments. Developing our understanding of the theories and competencies underlying learning from others provides insights into the big questions of persuasion, pedagogy, and culture. How can I convince someone to heed my advice? How does a student's understanding of the teacher and situation impact their learning? How is knowledge transmitted from one generation to the next?

References

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

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In C. Laura, H. Christoph, & F. S. Thomas (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2469–2474). Boston, Massachusetts: Cognitive Science Society.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology, 51,* 1173–1182.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1,* 1–23.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE, 6*(5), e19812. http://dx.doi.org/10.1371/journal.pone.0019812

Birch, S. A., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science, 13,* 363–369.

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use oth-

ers' past performance to guide their learning. *Cognition, 107,* 1018–1034.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101,* 127–151.

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology, 38,* 958–966.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science, 8,* 535–543.

Butterfield, J., Jenkins, O. C., Sobel, D. M., & Schwertfeger, J. (2008). Modeling aspects of theory of mind with Markov random fields. *International Journal of Social Robotics, 1,* 41–51.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21,* 315–330.

Chow, V., Poulin-Dubois, D., & Lewis, J. (2008). To see or not to see: Infants prefer to follow the gaze of a reliable looker. *Developmental Science, 11,* 761–770.

Clément, F., Koenig, M., & Harris, P. (2004). The ontogenesis of trust. *Mind & Language, 19,* 360–379.

Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science, 12,* 426–437.

Eaves, Jr., B. S., & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. *Advances in Child Development and Behavior, 43,* 295–319. http://dx.doi.org/10.1016/b978-0-12-397919-3.00011-3

Feng, B., & MacGeorge, E. L. (2006). Predicting receptiveness to advice: Characteristics of the problem, the advice-giver, and the recipient. *Southern Communication Journal, 71,* 67–85.

Finetti, B. (1974). *Theory of probability; a critical introductory treatment*. New York, NY: Wiley.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review Psychology, 50,* 21–45.

Gardner, P. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9,* S55–S79.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems, 42,* 1481–1493.

Goodman, N. D., Baker, C., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In T. Niels & V. R. Hedderik (Eds.), In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2759–2764). Amsterdam, the Netherlands: Cognitive Science Society.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). *Church: A language for generative models*. Retrieved from https://arxiv.org/abs/1206.3255

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science, 5,* 173–184.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In Ron Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge, UK: Cambridge University Press.

Harris, P. L. (2007). Trust. *Developmental Science, 10,* 135–138.

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366,* 1179–1187.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70,* 117–133.

Hu, J. C., Buchsbaum, D., Griffiths, T. L., & Xu, F. (2013). When does the majority rule? Preschoolers' trust in majority informants varies by domain. In M. Knauff, N. Sebanz, M. Pauen, & I. Wachsmuth (Eds.), In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2584–2589). Austin, TX: Cognitive Science Society.

Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development, 8,* 263–283.

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science, 17,* 757–758.

Jungermann, H., & Fischer, K. (2005). Using expertise and experience for giving and taking advice. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (pp. 157–173). Mahwah, NJ: Lawrence Erlbaum.

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2010). Children's selective trust in native-accented speakers. *Developmental Science, 14,* 106–111.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. New York, NY: Cambridge University Press.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15,* 694–698.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76,* 1261–1277.

Kushnir, T. (2013). How children learn from and about people: The fundamental link between social cognition and statistical evidence. In M. Banaji, & S. Gelman (Eds.), *The development of social cognition* (pp. 191–196). New York, NY: Oxford University Press.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2014). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package*). R package version 2.0–11. Retrieved from https://cran.r-project.org/web/packages/lmerTest/index.html

Landrum, A. R., Eaves, B. S., Jr., & Shafto, P. (2015). Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences, 19,* 109–111.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In P. Bartlett, F. C. N. Pereira, Leon Bottou, Chris J. C. Burges, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25, pp. 2699–2707). Cambridge, MA: MIT Press.

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

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Matsui, T., Yamamoto, T., & McCagg, P. (2006). On the role of language in children's early understanding of others as epistemic beings. *Cognitive Development, 21,* 158–173.

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science, 10,* 126–134.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60,* 167–171.

Moore, C., & Davidge, J. (2009). The development of mental terms: Pragmatics or semantics? *Journal of Child Language, 16,* 633.

Nurmsoo, E., & Robinson, E. J. (2009). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. *Child Development, 80,* 23–27.

Olineck, K. M., & Poulin-Dubois, D. (2005). Infants' ability to distinguish between intentional and accidental actions and its relation to internal state language. *Infancy, 8,* 91–100.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., . . . Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition, 130,* 360–379.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology, 43,* 1216–1226.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Phillips, J. M. (1999). Antecedents of leader utilization of staff input in decision-making teams. *Organizational Behavior and Human Decision Processes, 77,* 215–242.

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology, 47,* 116–129.

Pillow, B. H., & Weed, S. T. (1997). Preschool children's use of information about age and perceptual access to infer another person's knowledge. *The Journal of Genetic Psychology, 158,* 365–376.

Pinheiro, J., & Bates, D. (2009). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Poulin-Dubois, D., Brooker, I., & Polonia, A. (2011). Infants prefer to imitate a reliable person. *Infant Behavior and Development, 34,* 303–309.

Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development, 61,* 973–982.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4,* 515–629.

Robinson, E. J., Butterfill, S. A., & Nurmsoo, E. (2011). Gaining knowledge via other minds: Children's flexible trust in others as sources of information. *British Journal of Developmental Psychology, 29,* 961–980.

Robinson, E. J., & Whitcombe, E. L. (2003). Children's suggestibility in relation to their understanding about sources of knowledge. *Child Development, 74,* 48–62.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10,* 173–220.

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development, 72,* 1054–1070.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1632–1637). Washington DC: Cognitive Science Society.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science, 7,* 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71,* 55–89.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62,* 159–174.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes, 84,* 288–307.

Sobel, D. M., & Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. *Child Development, 81,* 669–679.

Stock, H. R., Graham, S. A., & Chambers, C. G. (2009). Generic language and speaker confidence guide preschoolers' inferences about novel animate kinds. *Developmental Psychology, 45,* 884.

Stuhlmüller, A., & Goodman, N. D. (2013). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Journal of Cognitive Systems Research, 28,* 80–99.

Swol, V. M. L., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology, 44,* 443–461.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331,* 1279–1285.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems, 22,* 1874–1882.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false belief. *Child Development, 72,* 655–684.

Whalen, A., Buchsbaum, D., & Griffiths, T. L. (2013). How do you know that? Sensitivity to statistical dependency in social learning. In M. Knauff, N. Sebanz, M. Pauen, & I. Wachsmuth (Eds.), In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1593–1598). Austin, TX: Cognitive Science Society.

Whitcombe, E. L., & Robinson, E. J. (2000). Children's decisions about what to believe and their ability to report the source of their belief. *Cognitive Development, 15,* 329–346.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes, 69,* 237–249.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1–13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260–281.

Zaki, J. (2013). Cue integration a common framework for social cognition and physical perception. *Perspectives on Psychological Science, 8,* 296–312.

Received October 20, 2016
Revision received June 2, 2017
Accepted June 8, 2017 ■