# Rational Inference of Beliefs and Desires From Emotional Expressions

## Yang Wu, Chris L. Baker, Joshua B. Tenenbaum, Laura E. Schulz

*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

## Abstract

We investigated people's ability to infer others' mental states from their emotional reactions, manipulating whether agents *wanted*, *expected*, and *caused* an outcome. Participants recovered agents' desires throughout. When the agent observed, but did not cause the outcome, participants' ability to recover the agent's beliefs depended on the evidence they got (i.e., her reaction only to the actual outcome or to both the expected and actual outcomes; Experiments 1 and 2). When the agent caused the event, participants' judgments also depended on the probability of the action (Experiments 3 and 4); when actions were improbable given the mental states, people failed to recover the agent's beliefs even when they saw her react to both the anticipated and actual outcomes. A Bayesian model captured human performance throughout ($rs \geq .95$), consistent with the proposal that people rationally integrate information about others' actions and emotional reactions to infer their unobservable mental states.

*Keywords:* Theory of mind; Emotions; Facial expressions; Mental state inferences; Bayesian models

## 1. Introduction

In July 2014, 715 million people watched as Germany beat Argentina in the final game of the soccer World Cup championship. When Mario Goetze kicked the ball to score the winning goal, almost every one of those faces expressed an emotional reaction to the event. Intuitively, the spectators' facial expressions were influenced both by how strongly they believed that the ball would—or would not—go through the goal posts and how much

they wanted Goetze to score the goal. Some faces were apprehensive or upset: fans of Argentina who expected (with varying levels of confidence) that Goetze would score a goal. Other faces were hopeful or delighted: fans of Germany who believed (again with different degrees of certainty) that they were about to win the match. Could you, as an observer, have looked at the faces of the fans and inferred their desires and beliefs?

Research suggests that in simple contexts, even very young children can infer others' desires given information about their beliefs and vice versa (see Baillargeon, Scott, & He, 2010; Saxe, Carey, & Kanwisher, 2004; and Wellman, Cross, & Watson, 2001, for reviews). If, for instance, observers know an agent's desire (e.g., to get a ball) and see her action (reaching for a box), they can infer her beliefs (that the ball is in the box); similarly, if observers know an agent's beliefs (that the ball is in the box) and see her action (reaching for the box), they can infer her desire (to get the ball). Indeed, given sufficiently rich information about an agent's actions (i.e., if someone checks one location and then changes course and heads to another), people can infer beliefs and desires simultaneously (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). Recently, computational models have begun to formalize these and many other aspects of theory of mind (e.g., Baker, Saxe, & Tenenbaum, 2009; Baker et al., 2017; Frank & Goodman, 2012, 2014; Frank, Goodman, & Tenenbaum, 2009; Goodman & Stuhlmüller, 2013; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Kao, Wu, Bergen, & Goodman, 2014; Lucas et al., 2014; Shafto, Eaves, Navarro & Perfors, 2012; Shafto, Goodman, & Frank, 2012; Zaki, 2013).

However, the assumptions governing much of this literature may underestimate the difficulty of inferring mental states in the real world. When we observe strangers, we are typically ignorant of both their beliefs and desires and we rarely get to observe uniquely informative sequences of actions. At the same time, more information may be available to observers than merely observable actions and the context in which they occur. As the World Cup example suggests, people often have emotional reactions to both anticipated and actual events. Although emotions themselves are not observable, their effects on people's facial expressions typically are. Here, we investigate the hypothesis that people's emotional response to events provides rich evidence about unobservable mental states that would otherwise be ambiguous. We look at whether people can use information about an agent's emotional reactions (and actions if any) to recover her beliefs and desires, and we compare people's judgments with the predictions of an ideal observer model.

Given the vast literature on both emotion and theory of mind, some justification is required for suggesting that the question of adults' ability to recover mental states from emotional expressions remains unresolved. Note, however, that to the degree that the literature on emotion and theory of mind has been connected, the vast majority of studies have focused on people's ability to infer others' emotions from behavioral cues, mental state knowledge, and contextual information. Thus, for instance, participants have been asked to predict what emotion someone would feel upon learning that a close friend betrayed a secret (Smith & Lazarus, 1993), or on being called into the boss's office after learning that the company is planning massive layoffs (Skerry & Saxe, 2015). Here, we are interested in the inverse problem: the conditions under which people can use contextual cues and emotional expressions to recover someone's beliefs and desires about the

outcome of an event, both when the person is merely a spectator of the event (as in the World Cup example) and when she is causally responsible for it.

We begin with a review of the developmental literature because the relationship between emotion understanding and other aspects of theory of mind has perhaps been most extensively investigated in early childhood. Infants begin to represent the relationship between an agent's goals and their emotions within the first year of life. Thus, for instance, 10 month olds look longer when an agent responds negatively than positively to achieving a goal (although the negative response does not lead to longer looking if the agent failed to achieve the goal; Skerry & Spelke, 2014). By 2 years of age, children explicitly predict that someone will be happy if she gets what she wants and sad if she does not (Stein & Levine, 1989; Wellman & Woolley, 1990; Yuill, 1984).

By contrast, the connection between emotional expressions and others' beliefs emerges relatively late: Only between the ages of 4 and 6 years do children expect an agent to be surprised if her beliefs are falsified and to be happy if she falsely believes that her desires will be fulfilled (Baron-Cohen, 1991; Hadwin & Perner, 1991; Harris, Johnson, Hutton, Andrews, & Cooke, 1989; Wellman & Banerjee, 1991). Moreover, children's ability to represent the emotions commensurate with true and false beliefs lags behind their ability to infer the beliefs themselves (Bender, Pons, Harris, & de Rosnay, 2011; De Rosnay, Pons, Harris, & Morrell, 2004; Hadwin & Perner, 1991; Harris et al., 1989; Pons, Harris, & de Rosnay, 2004; Ruffman & Keenan, 1996; Wellman & Bartsch, 1988). For instance, 4 and 5 year olds may correctly represent Red Riding Hood's false belief (that her grandmother is in bed), but incorrectly infer that she is scared (Bradmetz & Schneider, 1999). Explicit categorization of emotion concepts also emerges relatively late in development (see, e.g., Widen, 2016; Widen & Russell, 2008, 2010).

As clear from the above, most developmental studies of emotion have focused on what children understand *about* emotional expressions; fewer studies have asked what children can learn *from* emotional expressions, including whether children can use others' emotional expressions to recover their beliefs and desires. However, current research suggests that this ability emerges more slowly over development. Thus, for instance, infants as old as 14 months fail to use an agent's emotional reaction (i.e., positive or negative) to infer which of two food containers she wants, although they can predict which container she will reach for from the direction of her gaze (Vaish & Woodward, 2010). Similarly, 14 month olds fail to use an agent's positive and negative emotional reactions to infer that an agent likes a food the child does not, although, at 18 months, toddlers succeed (Repacholi & Gopnik, 1997). By 2 years of age, children can use an agent's emotional reaction to say explicitly whether she is looking at something she does or does not want (Wellman, Phillips, & Rodriguez, 2000).

Such inferences refer to others' desires; inferences about others' beliefs undergo more protracted development. Even children as old as 6 years rarely refer to others' beliefs in explaining their emotional reactions (Rieffe, Terwogt, & Cowan, 2005). The exceptions are that 4 and 5 year olds use beliefs to account for fearful or atypical emotional reactions (e.g., saying "She thought it was a ghost" if a character looks scared after hearing a noise or "She thought it would be something else" if someone looks

sad on opening a gift; Rieffe et al., 2005; see also Wellman & Banerjee, 1991). However, the interpretation of these findings is complicated by the fact that young children have learned a number of scripts connecting familiar events and emotions (e.g., between getting a puppy and being happy or dropping an ice cream cone and being sad; Barden, Zelko, Duncan, & Masters, 1980; Denham, Zoller, & Couchoud, 1994; Fabes, Eisenberg, McCormick, & Wilson, 1988; Gnepp, McKee, & Domanic, 1987; Harris, Olthof, Terwogt, & Hardman, 1987; Trabasso, Stein, & Johnson, 1982; Widen & Russell, 2010). Thus, children might link fear with a belief in ghosts, or sadness with disappointment in a gift (Rieffe et al., 2005) without necessarily being able to recover mental states from emotions broadly.

Perhaps the strongest evidence that children connect beliefs to emotional responses comes from studies showing that children invoke others' representations of past experiences to explain their current emotions (Harris, 1983; Harris, Guz, Lipian, & Man-Shu, 1985; Lagattuta & Wellman, 2001; Lagattuta, Wellman, & Flavell, 1997; Taylor & Harris, 1983). Thus, for instance, between 4 and 6 years, children expect people to feel more intensely about recent events than past ones, and they recognize that people will be happy if they remember positive events and forget negative ones (Harris, 1983; Harris et al., 1985; Taylor & Harris, 1983). Children also understand that particular events in an individual's past can lead to idiosyncratic emotional reactions; for instance, 4 and 5 year olds explain that a girl may be sad on seeing a puppy if her own puppy ran away (Lagattuta et al., 1997; Lagattuta & Wellman, 2001; see also Lagattuta, 2005).

In the real world, however, observers typically have no more access to others' past history of emotional experiences than to their beliefs and desires. Theory of mind is a challenging inference problem because the only information available is often only that which can be observed in the environment and the agent's behavior. Precisely for this reason, others' emotional reactions might be a particularly valuable cue to their mental states. However, the question of whether—absent specific prior knowledge about the individual—people can use emotional reactions and contextual information to jointly recover others' beliefs and desires remains largely unanswered (though see Wu & Schulz, 2017, for some recent evidence in 5 year olds).

Thus, we now turn to the adult literature. There is of course a large body of work on emotion and emotional expressions per se (see e.g., Ekman, 1992; Barrett, 2011; Barrett, Lewis, & Haviland-Jones, 2016; Russell, 2003, for reviews). However, unlike the developmental literature, this work has remained relatively disconnected from research on theory of mind (i.e., inferences about agent's beliefs and desires). One exception, and the work that perhaps best connects emotion to other cognitive states, is appraisal theory: a theory suggesting that an individual's evaluation of events plays a crucial role in eliciting and differentiating her emotional responses to those events (e.g., Lazarus, 1991; Ortony, Clore, & Collins, 1988; Scherer, 1984). Different appraisal theories differ in the appraisal dimensions that are at stake (e.g., the probability of an outcome, the desirability of an outcome, the immediacy of an outcome, etc.; see Moors, Ellsworth, Scherer, & Frijda, 2013, for a review). However, appraisal theories are united in assuming that an agent's beliefs and desires affect her evaluation of events and thus the emotional reactions she generates.

Appraisal theory is a scientific theory of how emotions are generated within the individual. It does not attempt to describe the analogous *intuitive* theory: how the individual herself might think about the causes of her emotional states, or how naïve observers might use someone's emotional reaction to infer her beliefs and desires. Nonetheless, many studies suggest that in addition to identifying others' emotions by their facial expressions (e.g., Ekman, 1992), vocalizations (e.g., Bachorowski & Owren, 2003), posture, and gait (e.g., Dael, Mortillaro, & Scherer, 2012), adults' emotion inferences depend on information about others' perceived expectations and attitudes towards events (Clore & Ortony, 2013; Ortony, Clore, & Collins, 1988; Scherer & Meuleman, 2013; Zaki, Bolger, & Ochsner, 2008). As in the developmental literature, however, such work has focused almost uniformly on how the appraisal of events affects the prediction and interpretation of emotional responses (see e.g., Fontaine, Poortinga, Setiadi, & Markam, 2002; Fontaine, Scherer, Roesch, & Ellsworth, 2007; Skerry & Saxe, 2015) rather than how contextual information and emotional reactions to events might inform adults' judgments about others' beliefs and desires about those events.

Here, we propose that people infer others' unobservable mental states from their emotional reactions using an intuitive theory of emotions, structurally analogous to appraisal theories in assuming that emotional reactions are probabilistically affected by agents' beliefs and desires about events. We focus specifically on whether an agent did or did not *believe* the outcome would occur, did or did not *want* the outcome to occur, and did or did not act to *cause* the outcome to occur. We focus on these three factors, not to imply that they are exhaustive, but because a primary goal of the current research is to provide a formal account of the role of emotional reactions in theory of mind, and beliefs, desires, and intentional action are at the heart of traditional models of theory of mind. Additionally, empirical work suggests that attributions of desirability, expectedness, and causal responsibility capture much of the variance in people's emotional reaction to events (see e.g., Scherer & Meuleman, 2013; Scherer, Schorr, & Johnstone, 2001; Skerry & Saxe, 2015). In addition to manipulating these factors, we independently vary the amount of evidence participants have about the agent's emotional reaction across experiments. Insofar as people are updating their beliefs from the data, they should draw stronger inferences when more evidence is available.

Because our focus in this paper is on the inference from observable emotional reactions to mental states involved in the cognitive appraisal of events, we can remain agnostic about an issue that has been the focus of many previous investigations: the inference from observed correlates of emotional reactions (e.g., specific facial expressions) to classifications of emotions themselves (e.g., Carroll & Russell, 1996; Crivelli, Russell, Jarillo, & Fernández-Dols, 2016; Gnepp, 1983; Izard, 1994; Scherer, Banse & Wallbott, 2001; Sievers, Polansky, Casey, & Wheatley, 2013; see Barrett, Mesquita, & Gendron, 2011; and Keltner, Tracy, Sauter, Cordaro, & McNeil, 2016, for reviews). There is considerable debate about whether the expression of emotion is universal, to what extent body language affects the interpretation of facial expressions, and the ways the expression and interpretation of emotions is affected by socio-cultural context (e.g., Carroll & Russell, 1996; Darwin, 1872/1965; Ekman & Friesen, 1971; Elfenbein, Beaupré, Lévesque, & Hess, 2007;

Lee & Anderson, 2016; Matsumoto & Willingham, 2009; Meeren, van Heijnsbergen, & de Gelder, 2005). However, these debates need not be of primary concern here. We take as a premise that at least within a well-specified context and shared cultural knowledge, people can probabilistically infer some emotional content from facial expressions. Our question is whether humans can integrate this content with information about the broader context and agents' actions (when applicable) to jointly infer agents' beliefs and desires.

We begin by specifying a simple probabilistic generative model of how an agent's appraisal of a situation—her beliefs and desires about an event—might lead to an emotional reaction to information about that event. This generative model forms the core of a Bayesian account of people's naïve theory of emotional responses, letting us consider how an ideal observer might reason backward from an agent's emotional reaction to the beliefs and desires that generated it. We then conduct a series of closely related experiments to quantitatively calibrate the model and test the inferences it supports.

## 2. Computational model

We take a Bayesian approach (Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) to characterizing the structure of the intuitive causal theory relating classical components of theory of mind (beliefs, desires, and actions) to observable emotional responses. Our approach is specifically inspired by research describing aspects of social reasoning as Bayesian inference (e.g., Baker et al., 2017, 2009; Frank & Goodman, 2012, 2014; Frank et al., 2009; Goodman & Stuhlmüller, 2013; Hamlin et al., 2013; Kao et al., 2014; Lucas et al., 2014; Ong, Zaki, & Goodman, 2015; Shafto, Eaves et al., 2012; Shafto, Goodman et al., 2012; Zaki, 2013).

We start by building a generative model including all the variables in our study. The generative model builds on the traditional theory of mind framework. We specify that an agent's beliefs and desires about an event are probabilistic causes of her emotional reaction to the event (if she is an observer of events) and also of her actions (if she is causally responsible for the event). See Fig. 1(a). *Belief* and *Desire* themselves are generated from a context-specific prior reflecting people's commonsense expectations of what beliefs and desires the agent is likely to have in a given context. Because all conditions of each experiment occur in an identical context, context does not play a differentiating role here and is not otherwise specified in the model. *Belief* and *Desire* cause *Action* in accord with a principle of rationality: An agent is expected to take actions that would lead to her desires being fulfilled given her beliefs. We integrate emotions with this framework by adding the agent's emotional reaction ($Reaction_0$) before she knows the outcome of the event. This emotional reaction is determined by whether the expected outcome (of her *Action* if relevant) given her *Belief* would fulfill her *Desire* (as illustrated by the blue arrows in Fig. 1(a)). We add another emotional reaction ($Reaction_1$) when the agent knows the final outcome of the event. This reaction is determined by whether the final *Outcome* fulfills her *Desire*, whether it confirms her previous *Belief*, and whether she is responsible for (i.e., her *Action* causes) the outcome (as illustrated by the red
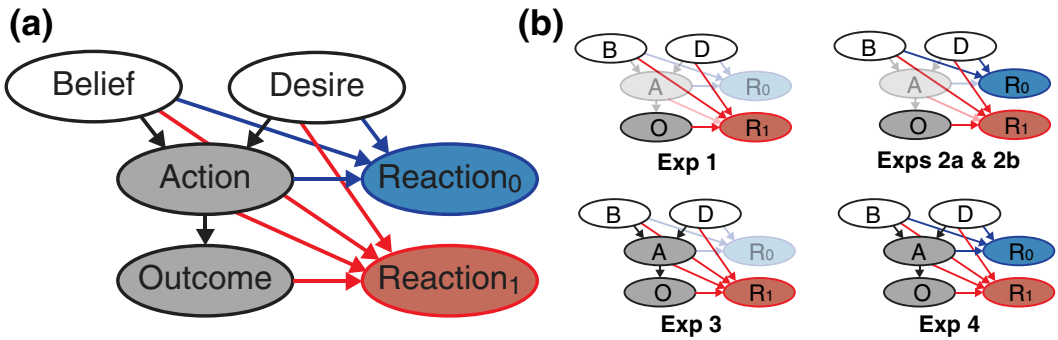
Fig. 1. (a) Template for Bayesian network models of people's intuitive theory of emotional responses and its integration with theory of mind. Arrows indicate hypothesized causal relationships between mental states, actions, outcomes, and emotional reactions. This generative model starts with people's representation of an agent's *Belief* and *Desire* about an event, generated from a context-specific prior for the relevant beliefs and desires in each scenario. The agent's *Belief* and *Desire* lead to an *Action* following the principle that agents act to fulfill their desires based on their beliefs about the world (the principle of rational action). The agent's *Action* causes an *Outcome*. *Reaction$_0$* is the agent's emotional reaction to the expected outcome based on her *Desire* and *Belief* and, if she acts, her *Action* (the blue arrows). *Reaction$_1$* is the agent's emotional response when she knows the outcome. This is influenced by the *Outcome*, her *Desire*, *Belief* and, if she is responsible for it, her *Action* (the red arrows). (b) Different subnetworks can characterize people's intuitive theory in different contexts. When the outcome is caused by an external cause rather than the agent's action (Experiments 1, 2a, and 2b), the *Action* (as well as any arrow directly connected with this node) drops out; when the agent's emotional reaction to the anticipated outcome is not observed, *Reaction$_0$* (as well as arrows directly connected with it) drops out (Experiments 1 and 3).

arrows in Fig. 1(a)). Nodes (as well as arrows connected with them) corresponding to any variable not present in a given scenario can be removed (see Fig. 1(b)). For example, when the outcome is caused by an external cause rather than the agent's action (Experiments 1 and 2), *Action* and all arrows connected with it drop out. When the agent's emotional reaction to the expected outcome is not observed (Experiments 1 and 3), *Reaction$_0$* and all arrows connected with it drop out.

The model for each of the experiments can thus be spelled out in detail. In Experiment 1 the agent observes the outcome of an event that she does not cause. The directed graph (Fig. 1(b), Exp. 1) indicates that the agent's emotional reaction ($R_1$) is affected jointly by her desires, beliefs, and the outcome. Experiment 2 is identical except that the agent reacts to both the expected and actual outcomes. Her emotional reaction to the expected outcome ($R_0$) is affected by her desires and beliefs; her reaction to the actual outcome ($R_1$) is affected by her desires, beliefs, and the outcome (Fig. 1(b), Exp. 2). Experiments 3 and 4 are similar to Experiments 1 and 2 except that the agent is causally responsible for the event, acting to bring it about. In Experiment 3 (as in Experiment 1) the agent reacts only to the actual outcome and the directed graph indicates that her emotional reaction ($R_1$) is affected jointly by her desires, beliefs, action, and the outcome (Fig. 1(b), Exp. 3). In Experiment 4 (as in Experiment 2) the agent reacts to both the anticipated and actual outcomes. The graph indicates that her emotional reaction to the anticipated outcome ($R_0$) is affected only by her desires,

beliefs, and action; her reaction to the actual outcome ($R_1$) is affected by her desires, beliefs, the action, and the outcome itself (Fig. 1(b), Exp. 4).

The informational content in these causal relationships can be expressed in terms of probability distributions over each variable in the network, conditioned on its parents. For instance, considering the case where all nodes and arrows are present, our Bayesian model predicts that backward inferences of *Belief* and *Desire* given observable information (e.g., *Action*, *Outcome*, and *Reactions*) decompose into a product of terms corresponding to each of the forward causal dependencies via Bayes' rule:

$$P(B,D|A,O,R_0,R_1) \propto P(R_1|B,D,A,O) \times P(R_0|B,D,A) \times P(A|B,D) \times P(B,D), \quad (1)$$

where we have abbreviated each variable by its first letter. To determine whether people's generative causal knowledge supports inferences about belief and desire from emotional expressions, actions and contextual cues, as predicted by our model, we elicit participants' judgments about each of the four components of the right-hand side of Eq. 1. We compute the normalized products of the forward distributions according to Eq. 1. We then compare the model's posterior distributions to an independent group of participants' backward inferences from the observable information to the agent's belief and desire. Our Bayesian model can account for our manipulations across the four experiments: when the agent does not act to cause the outcome (Experiments 1 and 2), $P(A|B,D)$ drops out from the right side of Eq. 1; when the reaction to the anticipated outcome (*Reaction_0*) is not observed (Experiments 1 and 3), $P(R_0|B,D,A)$ drops out. We also compare our model with several alternative models.

Our model is similar both in spirit and in its technical approach to a recent proposal by Ong et al. (2015) for how to capture intuitive theories of emotion in a causal, generative inference framework. They show how a similar model compellingly captures a range of phenomena about how people map between observed events (i.e., the outcome of bets on a Roulette wheel) and emotional reactions, including the integration of multiple cues to an emotional response. Critically however, people do not react to observed events; they react to a *mental representation* of those events, a representation that is affected jointly by their beliefs and desires. Ong et al. (2015) showed that people could recover emotions when the agent's mental states were not in question and all information was observed (i.e., the goal was to make money and the expectedness of the event was established by the distributions on the Roulette wheel). However, the beliefs and desires that determine people's emotional reactions to outcomes are often variable and unknown, and distinct combinations of beliefs and desires can generate different emotional reactions even to identical actions and outcomes. This study focuses on how we might use emotional reactions, even to identical events, to recover these distinct combinations of beliefs and desires.

## 3. Behavioral experiments

We test our Bayesian model with four behavioral experiments that vary the desirability and expectedness of the event within experiments and the causal relationship of the agent to the event and the amount of information participants have about the agent's emotional

reaction across experiments. Thus, in Experiments 1 and 2, the agent is merely an observer of events; in Experiments 3 and 4, she causes the events. Participants see the agent's reaction only to the event outcome in Experiments 1 and 3, but they see her reactions to both the anticipated and actual outcomes in Experiments 2 and 4. To test whether the model is robust to minor variations in the stimuli, we run internal replications of two of the experiments, comparing morphed versus pure facial expressions in Experiments 2a and b; and photographs versus movies in Experiments 3 and 3 Supplemental.

### 3.1. Experiment 1

In Experiment 1 (and all the experiments to follow), we use a scenario in which an agent has an unspecified belief and desire. We provide information about the outcome of events and the agent's emotional reaction to the outcome and then look at whether participants can use this information to recover the agent's beliefs and desires. We then compare the behavioral results to the model predictions.

#### 3.1.1. Method

*3.1.1.1. Design and materials*: We created an emotionally charged scenario in which an agent, Grace, learns that a plane has crashed on a route often flown by her coworker John. Grace's desire and belief are unspecified but constrained to two possibilities: Grace either wants John to die or live, and believes John is either on the flight that crashed or on a different, safe flight. There are two possible outcomes: John lives or dies. (See Supporting Information Text 1.1 for the complete scenario.)

The eight possible combinations of Grace's belief, desire, and the outcome yield Conditions 1–8 of the experiment. See Fig. 2(a). To generate Grace's emotional reaction in each condition, we used facial morphing software to create photograph stimuli. Consistent with the developmental literature (e.g., Hadwin & Perner, 1991; MacLaren & Olson, 1993; Repacholi & Gopnik, 1997; Skerry & Spelke, 2014; Stein & Levine, 1989; Wellman & Banerjee, 1991; Wellman & Woolley, 1990; Yuill, 1984),[1] we assumed that if the outcome was consistent with Grace's desire, her expression should be largely positive (and if inconsistent, largely negative), and that if the outcome was consistent with Grace's belief, her expression should not include surprise (but if inconsistent, it should). Since compound facial expressions combine muscle movements involved in the subordinate categories (Du, Tao, & Martinez, 2014), we created compound emotional reactions (e.g., in Condition 5, happily surprised) by morphing the corresponding two basic facial expressions (i.e., happy and surprised). See Supporting Information Text 2.1.1 and Table S1 for more details.

Fig. 2. (a) Design of Experiments 1, 2a, 3, and 4. The beliefs *Crash* and *Safe* refer to the plane-crash scenario while *Poison* and *Sugar* refer to the chemical-factory scenario. (b) Given the plane-crash scenario, participants' model calibration judgments on an unnormalized 0–100 scale for (i) the prior probability of Grace's belief and desire, and the conditional likelihoods of (ii) $Reaction_0$ and (iii) $Reaction_1$ (photograph stimuli). (c) Analogous judgments for the chemical-factory scenario. Error bars indicate 95% confidence intervals. (Note that we were unable to track down the copyright permissions for the original photographs used. Figures throughout this paper show hand-drawn pencil sketches from our photograph stimuli.)

**(a) Design of Experiments 1, 2a, 3 and 4**

| Desire&Belief | Die&[Crash/Poison] | | Live&[Safe/Sugar] | | Die&[Safe/Sugar] | | Live&[Crash/Poison] | |
|---|---|---|---|---|---|---|---|---|
| Reaction$_0$ (Exps 2a&4 only) | | | | | | | | |
| Outcome | Die | Live | Die | Live | Die | Live | Die | Live |
| Reaction$_1$ | | | | | | | | |
| Condition | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

**(b) Experiments 1&2a (Plane-crash scenario)**

Mental States: Die&Crash, Live&Safe, Die&Safe, Live&Crash

**(i) Prior**
$\hat{P}(B, D)$

**(ii) Reaction$_0$ likelihood** (Exp 2a only)
$\hat{P}(R_0 \mid B, D)$

**(iii) Reaction$_1$ likelihood**
$\hat{P}(R_1 \mid B, D, O)$
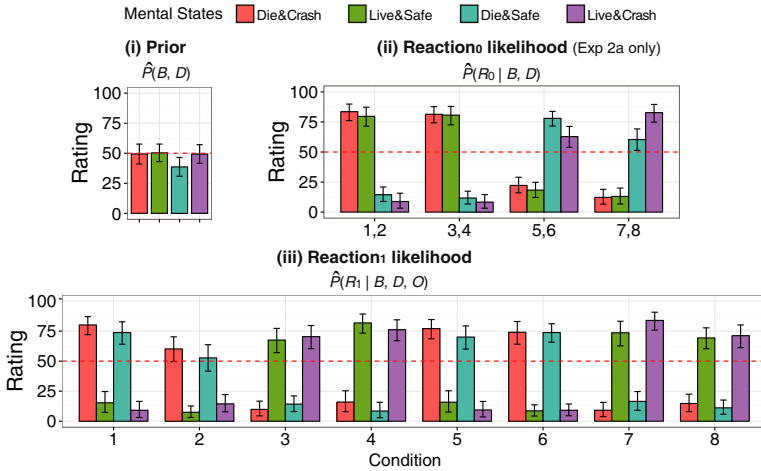


**(c) Experiments 3&4 (Chemical-factory scenario)**

Mental States: Die&Poison, Live&Sugar, Die&Sugar, Live&Poison

**(i) Prior**
$\hat{P}(B, D)$

**(ii) Action likelihood**
$\hat{P}(A = \text{True} \mid B, D)$

**(iii) Reaction$_0$ likelihood** (Exp 4 only)
$\hat{P}(R_0 \mid B, D, A = \text{True})$

**(iv) Reaction$_1$ likelihood**
$\hat{P}(R_1 \mid B, D, A = \text{True}, O)$
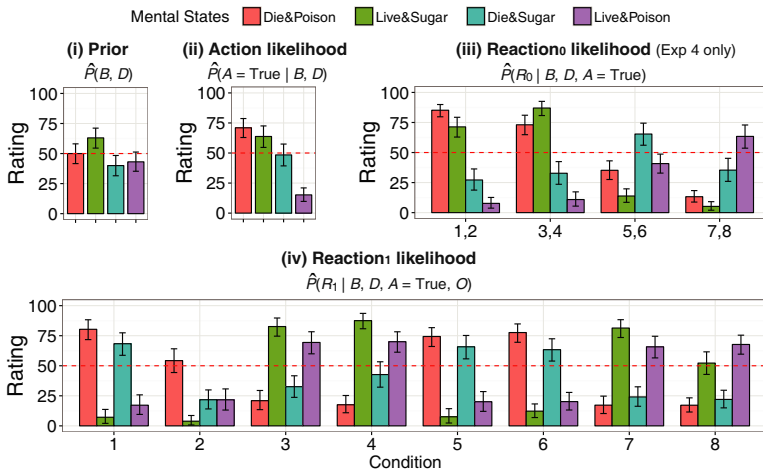


*3.1.1.2. Participants and procedure*:  All participants in this and the following experiments were recruited on Amazon Mechanical Turk. Participation was restricted to individuals with HIT approval rate of 95% or higher. A range of ethnicities and socioeconomic backgrounds

reflecting the diversity of the marketplace was represented. We preset the sample size for each group of participants at $n = 60$, sufficient for 97% power assuming a medium effect size (Cohen's $d = 0.50$). On average, 12% of the participants were dropped due to responding to less than half of the test questions or failing catch questions (designed to evaluate participants' comprehension of the scenario; see Supporting Information Text 1 for details). All remaining participants were included in the final analyses; the resulting minimum power to detect an effect in any experiment was 91%.

To test the predictions of the model, three separate groups of participants were recruited. Groups one and two were asked for judgments used to calibrate the model; the third group was the test group.

The first group ($n = 57$) judged the prior plausibility of each combination of Grace's desire and belief given the context, $P(D,B)$. The four possible combinations are as follows: (a) Grace wants John to die and believes John was on the flight that crashed (Die&Crash); (b) Grace wants John to live and believes John was on a safe flight (Live&Safe); (c) Grace wants John to die and believes John was on a safe flight (Die&Safe); and (d) Grace wants John to live and believes John was on the flight that crashed (Live&Crash).

The second group of participants ($n = 45$) was asked to judge the plausibility of Grace's facial reactions given her belief, desire, and the event outcome specified in each condition, $P(R_1|B,D,O)$. All the forward judgments in this study were elicited on a 0–100 scale and thus are not strictly speaking conditional probabilities. We treat them as relative estimates of the corresponding probabilities, which are effectively normalized and converted to probabilities when processed through the Bayesian analysis of Eq. 1 to produce the model's posterior probability predictions.

The test group ($n = 52$) was asked to predict Grace's belief and desire given the event outcome and her reaction to this outcome, $P(B,D|O,R_1)$. All the mental state inferences in the study were collected on a 0–100 scale but normalized to sum to 1 over all four possible belief-desire combinations. See Supporting Information Text 3 for details.

### 3.1.2. Results and discussion

*3.1.2.1. Model calibration*: The prior probability of each combination of desire and belief was relatively uniform (Fig. 2(b)(i)), indicating that, as intended, the task instructions led people to consider all possible mental states. (See Supporting Information Text 4.1 for detailed analyses.) Similarly, participants' judgments about the relative plausibility of the different emotional expressions were consistent with our assumption that Grace should have a positive expression if she wanted the outcome to occur and a negative expression if she did not. However, contrary to our assumptions, participants did not strongly distinguish the conditions under which Grace would or would not look surprised. Consider, for example, the emotional expression in Condition 1. This expression was treated as equally plausible for two cases where John died: both the scenario in which Grace wanted John to die and believed John was on the flight that crashed (Die&Crash), and the scenario in which Grace wanted John to die and believed John was on a safe flight (Die&Safe). Thus, participants seemed to expect Grace's facial expression to reflect her

desires but not her beliefs. Fig. 2(b)(iii) shows participants' conditional likelihood ratings for each of the eight emotional reactions as a function of Grace's desire and belief, given the event outcome from the corresponding condition. (See Supporting Information Text 4.4 for detailed analyses.)

*3.1.2.2. Mental state inferences*: Our primary question of interest was whether people could infer Grace's belief and desire in each of the eight conditions. We built a mixed-effects model, using Mental State and Condition as fixed factors and Subject as a random factor. There was no main effect of Condition ($F(7, 1561) = 0.18$, $p = .989$) but a significant main effect of Mental State ($F(3, 1561) = 166.12$, $p < .001$) and a significant interaction between Condition and Mental State ($F(21, 1561) = 4.35$, $p < .001$). We then looked at the main effect of Mental State in each condition and found a significant main effect of Mental State in each of the eight conditions (all $F$s > 7.54, all $p$s < .001). We further looked at whether participants rated the target mental state (i.e., the combination of desire and belief actually used to generate the facial expression) significantly higher than the other three mental states. This resulted in 24 comparisons across the eight conditions and the $p$ values reported here and in the following experiments were all corrected using the Bonferroni method.

Participants successfully rated the target combination of beliefs and desires higher than the other possibilities in Conditions 1 and 4 (all $z$s > 3.77, all $p$s < .004). However, in the remaining conditions, they failed to infer the agent's beliefs and recovered only the agent's desires, rating the target mental states significantly lower than the mental state with the correct desire but incorrect belief ($z = -4.63$, $p < .001$) in Condition 5, and failing to differentiate between the two mental states with the correct desire but different beliefs in Conditions 2, 3, 6, 7, and 8 (all $|z|$s < 2.06, all $p$s > .953). Thus, overall, participants successfully inferred the agent's desires but struggled to infer her beliefs. See Fig. 3(a) for the results by condition and Fig. 4(a) for the target and non-target responses averaged across conditions.

Similar results were found when we used one-sample $t$-tests (two tailed) to analyze the data. Here, we looked at whether any of the four combinations of mental states was rated significantly above 50 (i.e., the middle point of the 0–100 scale, where 0 indicated "completely implausible" and 100 indicated "completely plausible"). This resulted in 32 comparisons and the $p$ values reported here and in the following experiments were also corrected, using the Bonferroni method. Participants uniquely rated the target mental states significantly above 50 in Conditions 1 and 4 ($t_1(50) = 7.00$, $p_1 < .001$; $t_4(50) = 4.56$, $p_4 < .001$). They were biased towards the mental state with the correct desire but incorrect belief (Die&Crash) in Condition 5 ($t(51) = 6.47$, $p < .001$), and they failed to distinguish between the two mental states with the correct desire but different beliefs in Conditions 2, 3, 6, 7, and 8 (none of these ratings differed significantly from 50: all $|t|$s < 2.17, all $p$s = 1.000; mental states with the incorrect desire were rated significantly below 50: all $t$s < -3.67, all $p$s < .018).

The model predictions were generated according to Eq. 1 (omitting the *Action* and *Reaction*$_0$ terms; see Supporting Information Text 5.1), using the independent raters'
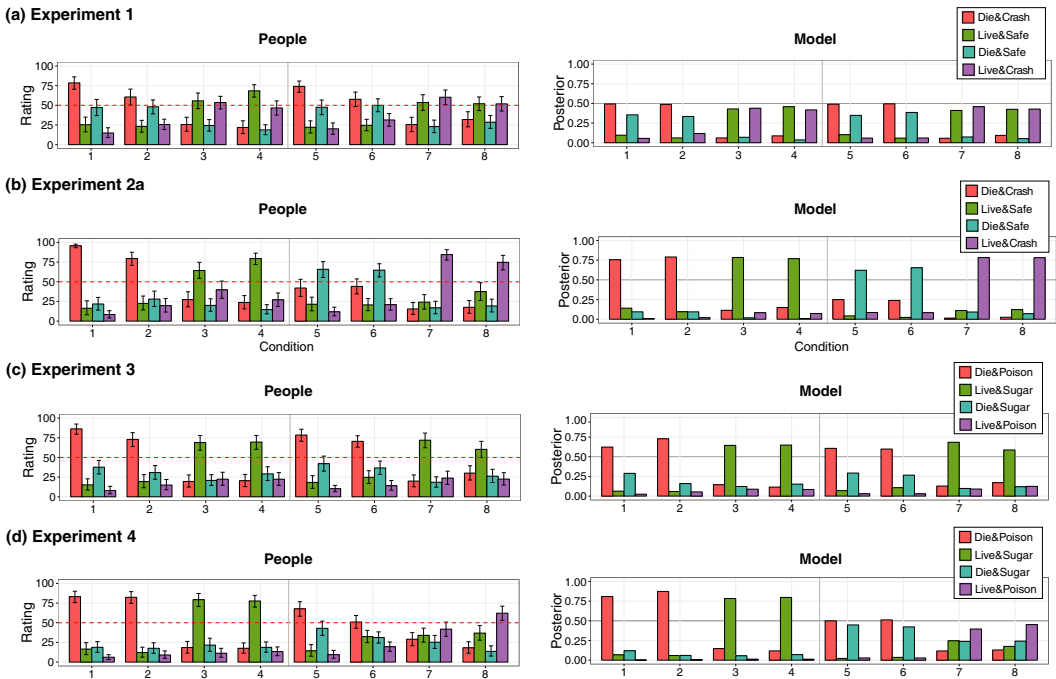
Fig. 3. People's mental state inferences on an unnormalized 0–100 scale and model predictions in Experiments 1, 2a, 3, and 4. Error bars indicate 95% confidence intervals.

judgments of the prior probability of each combination of belief and desire and the likelihood of each facial expression. (See Fig. 3(a)) The model predictions correlated highly with people's inferences ($r = .954$).

In sum, human judgments were rational with respect to the model predictions but reflect limitations on people's ability to infer other's mental states: Participants successfully recovered the agent's desires but struggled to infer her beliefs. This pattern of results is consistent with previous research suggesting that belief inferences are more difficult than desire inferences for both children and adults (Saxe et al., 2004; Wellman et al., 2001; see Apperly & Butterfill, 2009; Astington & Gopnik, 1991; Wellman, 2014, for reviews and discussion).

Note, however, that participants in Experiment 1 saw Grace's reaction only at a single time point: on observing the final outcome of the event. Arguably, if people could see Grace's emotional expression in response to the *anticipated* as well as the actual outcome, they might be able to use the presence or absence of a change in valence to infer the veracity of her beliefs. We test this hypothesis in Experiment 2a.

Additionally, one might wonder why participants appeared insensitive to the presence or absence of surprise in judging the likelihood of the facial reactions and, in parallel, resisted using surprise cues in the facial expressions to infer Grace's beliefs when asked to do so. These two behaviors, in two independent groups of participants, are consistent with each other if people are generally making rational Bayesian inferences from
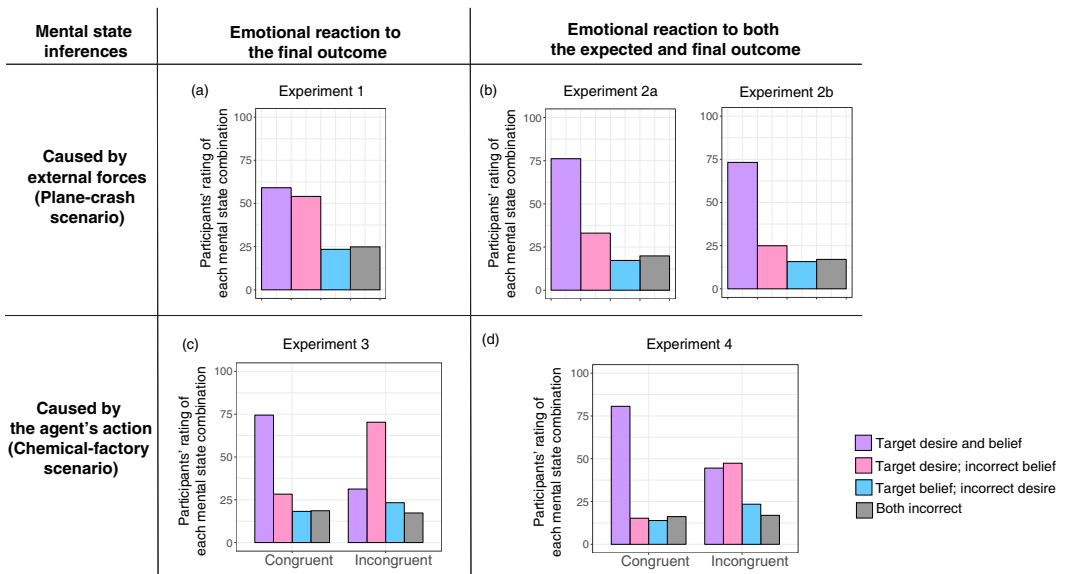
Fig. 4. Participants' mental state inferences averaged across conditions. In each plot, the first bar (purple) indicates the average rating of the target combination of desires and beliefs used to generate the facial expressions. The following three bars indicate the average ratings of each of the three non-target combinations. The pink bar indicates the target desire but incorrect belief; the blue bar indicates the target belief but incorrect desire; the gray bar indicates the incorrect desire and incorrect belief. In Experiments 1, 2a, and 2b, responses are averaged across all conditions. In Experiments 3 and 4, responses are averaged across the four conditions where the agent's action and emotional reaction provide either Congruent or Incongruent information about the agent's mental states.

emotional expressions back to mental states, but each was surprising to us empirically. We return to this question in Experiment 2b.

## 3.2. Experiment 2a

In Experiment 2a, we replicate Experiment 1 but show participants one additional emotional expression: Grace's reaction to anticipating the outcome of the event ($Reaction_0$). We hypothesized that if Grace looked happy about the outcome she expected but sad about the outcome she observed (or vice versa), participants would infer that Grace's initial belief was false (and that if her expression remained the same, that her initial belief was true).

### 3.2.1. Method
*3.2.1.1. Design and materials*: Experiment 2a was identical to Experiment 1 except that Grace's emotional reaction to the expected outcome was also observed. For Conditions 1, 4, 6, and 7, where the expected and actual outcomes match, we set the valence of $Reaction_0$ to match the valence of $Reaction_1$; for the remaining conditions where Grace has a false belief

(i.e., there is a mismatch between the expected and actual outcomes), we flipped the valence between $Reaction_0$ and $Reaction_1$. See Supporting Information Text 2.1.2.

*3.2.1.2. Participants and procedure*: To calibrate the model, participants ($n = 50$) rated the likelihood of $Reaction_0$, $P(R_0|B,D)$. Because the eliciting conditions for the other model calibration judgments (i.e., the prior probability of mental states and the likelihood of $Reaction_1$) were identical to those in Experiment 1, the judgments from Experiment 1 were used to calibrate the model here as well.

The test group ($n = 57$) inferred the probability of each combination of Grace's belief and desire given the event outcome and Grace's reactions to the anticipated and observed outcomes, $P(B,D|O,R_0,R_1)$. See Supporting Information Text 3.

### 3.2.2. Results and discussion

*3.2.2.1. Model calibration*: The likelihood of $Reaction_0$ is reported in Fig. 2(b)(ii). The positive expressions (those used in Conditions 1–4) were rated higher given the two mental states that Grace's desire would be fulfilled according to her belief (Die&Crash and Live&Safe) than given the two mental states that her desire would not (Die&Safe and Live&Crash). The negative expressions (those used in Conditions 5–8) showed roughly the opposite pattern. That is, as we had assumed, participants expected the agent to express positive emotions when the expected outcome given her belief would fulfill her desire, and negative emotions when it would not (see Supporting Information Text 4.3 for detailed analyses).

*3.2.2.2. Mental state inferences*: People's inferences are shown in Fig. 3(b). See also Fig. 4(b) for the overall pattern. We ran the same analyses as in Experiment 1. Mixed effects model analyses revealed no main effect of Condition ($F(7, 1688) = .28$, $p = .961$) but a significant main effect of Mental State ($F(3, 1688) = 357.75$, $p < .001$) and a significant interaction between Condition and Mental State ($F(21, 1688) = 4.80$, $p < .001$). A significant main effect of Mental State was found in each of the eight conditions (all $F$s > 15.05, all $p$s < .001). Participants rated the target mental states significantly higher than the other mental states in all conditions (all $z$s > 3.43, all $p$s < .014).

A similar pattern was found using one-sample $t$-tests. Participants uniquely rated the target mental states used to generate the facial expressions above 50 in Conditions 1, 2, 4, 6, 7, and 8 ($t_1(53) = 38.90$, $p_1 < .001$; $t_2(54) = 6.87$, $p_2 < .001$; $t_4(55) = 7.92$, $p_4 < .001$; $t_6(55) = 3.45$, $p_6 = .035$; $t_7(54) = 9.86$, $p_7 < .001$; $t_8(55) = 5.22$, $p_8 < .001$), and showed a non significant trend in the same direction in the remaining two conditions ($t_3(54) = 2.760$, $p_3 = .253$; $t_5(56) = 3.075$, $p_5 = .104$; all other mental states were rated significantly lower than or equal to 50: all $t$s < −1.42, all $p$s < 1.000).

These responses were well predicted by the model (generated according to Eq. 1, with $Reaction_0$ and $Reaction_1$ terms but no *Action* term; see Supporting Information Text 5.2). The model's posterior probability $P(B,D|O,R_0,R_1)$ favored the target mental states from which the reactions were generated in all conditions (see Fig. 3b); the correlation between the model predictions and people's inferences was high ($r = .953$).

Given the presence or absence of a change in valence between the expected and observed outcome, people were able to infer both the agent's beliefs and desires, and people's responses were well predicted by the Bayesian model. However, we are left with the question of why participants did not use the presence or absence of a surprised reaction to the outcome alone to infer the agent's beliefs in Experiment 1. In Experiment 2b, we run a replication of Experiment 2a using slightly different facial expressions to try to shed more light on the unanticipated finding.

### 3.3. Experiment 2b

In Experiments 1 and 2a, the agent's response to violations of her belief contained a mix of valence and surprise. In Experiment 2a, participants successfully recovered the agent's beliefs and desires from such morphed facial expressions. However, they may have done so only using the valence information, rather than the surprise cue. Suggestive evidence that this is the case comes from the model calibration judgments: When participants were asked to rate the relative plausibility of the different emotional expressions ($Reaction_1$ likelihood), they failed to distinguish expressions with and without surprise (see Fig. 2(b)(iii)).

One possibility is that participants simply failed to detect the presence or absence of surprise in the facial expressions. Especially since surprise was blended with valence information, the latter may have obscured the former to the point that people simply could not perceive surprise in these stimuli. To test this possibility, we conducted a follow-up study (Experiment 2b Supplemental) asking a separate group of participants to rate the degree to which Grace's facial reactions contained surprise and other basic emotions (e.g., happiness, sadness, anger, etc.). Inconsistent with this possibility, in the absence of the background scenario, participants were able to identify the absence or presence of surprise in the faces at a level roughly equivalent to the other emotions (Supporting Information Text 6).

Since people could identify the absence or presence of surprise in the facial expressions, why didn't they use this information to draw inferences about the content of Grace's beliefs? Another possibility, suggested by some versions of appraisal theory, is that in some contexts, surprise may function as an intensifier of valence: If, for instance, a desirable event is unexpected, surprise might magnify the felt happiness (Ortony, Clore, & Collins, 1988). In our scenarios, people may have interpreted the surprise only as an intensifier of valence, attenuating their responses to surprise per se. If this is the case, people may be more sensitive to the link between surprise and the veracity of beliefs when surprise is not blended with valence.

To test this, as well as to establish the degree to which our previous results are robust to minor variations in the stimuli, in Experiment 2b, we use only basic (de-morphed) emotional expressions matching the primary components of the morphed faces throughout. Conditions in which Grace's expectations are fulfilled result in facial expressions in which the valence corresponds to her desires (positive if desired; negative if not). Conditions in which Grace's expectations are violated result in facial expressions expressing

surprise without any valence information or expressing valence information without any surprise information. See Fig. 5(a). We predict that the results of Experiment 2a will be replicated, using unmorphed facial expressions; in particular, we predict that in the conditions where participants see the agent's valenced response to the anticipated outcome ($Reaction_0$) and her surprised response to the observed outcome ($Reaction_1$), they will successfully recover Grace's beliefs as well as her desires.

### 3.3.1 Method

*3.3.1.1. Design and materials*: The design was similar to Experiment 2a except that all the emotional reactions were unmorphed expressions. See Fig. 5(a). For $Reaction_1$, we replaced the original morphed expressions with the prototypical facial expressions matching the primary valence components of those faces (see Supporting Information Table S1: Components (%); the primary valence components were underlined). This generated Conditions 1, 2a, 3a, 4, 5a, 6, 7, and 8a. Besides valence, some of the morphed faces contained another key component—surprise. We created additional conditions in which these expressions were replaced by purely surprised faces, yielding Conditions 2b, 3b, 5b, and 8b. For $Reaction_0$, we re-used $Reaction_1$ from Conditions 1, 4, 6, 7, where the expected and actual outcomes matched.

*3.3.1.2. Participants and procedures*: To calibrate the model, we measured people's judgments about the likelihood of the new set of stimuli. Participants ($n = 58$) rated each of the four facial expressions responding to the expected outcome ($Reaction_0$) given Grace's belief and desire, $P(R_0|B,D)$. A separate set of participants ($n = 58$) judged each of the 12 facial expressions ($Reaction_1$) given Grace's belief, desire, and the outcome specified in each condition, $P(R_1|B,D,O)$.

The test participants ($n = 55$) judged Grace's belief and desire given the outcome of the event and Grace's facial reactions before and after she knew the outcome, $P(B,D|O, R_0,R_1)$.

### 3.3.2. Results and discussion

*3.3.2.1. Model calibration*: For $Reaction_0$ and the valenced $Reaction_1$, the estimated likelihoods were similar to those found in Experiments 1 and 2a (see Fig. 5(b)). For the surprised reactions, participants' judgments varied with the outcome. When John survived (*Outcome*: live), participants, as intended, judged the surprised faces more likely given false beliefs than true beliefs. However, counter to our intention, when John died (*Outcome*: die), participants judged that the surprised response was equally probable whether Grace expected the death or not (possibly because death may always be perceived as shocking even when it is in some sense anticipated). (See Supporting Information Texts 4.3 and 4.4 for detailed analyses.)

*3.3.2.2. Mental state inferences*: Participants' mental state inferences are reported in Fig. 5(c). See also Fig. 4(b) for the overall pattern. There was no main effect of Condition ($F(11, 2490) = .34$, $p = .976$) but a significant main effect of Mental State

**Experiment 2b**
**(a) Design**



**(b) Likelihoods of facial reactions**



**(c) Mental state inferences**



Fig. 5. Design and results of Experiment 2b. (a) Design. (b) Participants' model calibration judgments on an unnormalized 0–100 scale for the conditional likelihoods of (i) *Reaction*$_0$ and (ii) *Reaction*$_1$. (c) Participants' mental state inferences on an unnormalized 0–100 scale and model predictions.

($F(3, 2490) = 498.35$, $p < .001$) and a significant interaction between Condition and Mental State ($F(33, 2490) = 5.32$, $p < .001$). The main effect of Mental State was significant in all conditions (all $F$s > 3.00, all $p$s < .032). In 11 of the 12 conditions, participants rated the target mental state significantly higher than the other mental states (all $z$s > 4.98, all $p$s < .001); the exception was Condition 5b (all $|z|$s < 2.97, all $p$s > .108).

Converging results were found using one-sample *t*-tests. In the conditions where participants saw valenced facial reactions to the expected and observed outcomes, we replicated the finding from Experiment 2a that participants successfully recovered both the

agent's belief and desire ($t_1(53) = 17.71$, $p_1 < .001$; $t_{2a}(53) = 8.19$, $p_{2a} < .001$; $t_{3a}(53) = 4.112$, $p_{3a} = .007$; $t_4(53) = 6.00$, $p_4 < .001$; $t_6(53) = 5.13$, $p_6 < .001$; $t_7(53) = 5.81$, $p_7 < .001$; $t_{8a}(53) = 3.87$, $p_{8a} = .014$ and with a non significant trend in Condition 5a, $t(53) = 3.30$, $p = .082$). Similarly, when participants saw a valenced response to the expected outcome and a surprised response to the actual outcome, they successfully recovered the target mental states in Conditions 2b and 8b ($t_{2b}(53) = 5.44$, $p_{2b} < .001$; $t_{8b}(53) = 5.72$, $p_{8b} < .001$) and showed a non significant trend in the same direction in Condition 3b ($t(53) = 2.57$, $p = 0.631$; all other mental states were rated significantly below 50: all $t$s $< -3.83$, all $p$s $< .016$). Again, the exception was Condition 5b (mental states Live&Safe and Die&Safe were rated not significantly different from 50: both $|t|$s $< 3.12$, both $p$s $> .141$; mental states Die&Crash and Live&Crash were rated significantly below 50: both $t$s $< -3.91$, both $p$s $< .013$).

These behavioral responses were also predicted by our model. Model predictions were generated according to Eq. 1, with *Reaction$_0$* and *Reaction$_1$* terms but no *Action* term; see Supporting Information Text 5.2. The correlation between the model predictions and people's inferences was high ($r = .950$). See Fig. 5(c).

Thus overall, the results mirrored those in Experiment 2a, both with respect to people's ability to successfully infer others' mental states, and the model's ability to predict people's inferences. Nonetheless, they raise the question of why participants failed to recover the agent's beliefs and desires in Condition 5b. In this condition, Grace wanted John to die but believed he was on the safe flight. John, unexpectedly, did die, and Grace expressed surprise, but participants failed to use her surprised expression to infer that she had (falsely) believed that he was safe. Participants' likelihood judgments (see Conditions 3b and 5b in Fig. 5(b)(ii)) suggest the possibility that people may generally be surprised by someone's death and thus the surprised expressions may not be reliably informative about others' underlying beliefs. However, participants succeeded in the other condition involving a surprised response to death (Condition 3b, where Grace wanted John to live, believed he was on the safe flight, and was surprised at his death); thus, we cannot definitively explain the failure in the single condition. However, participants' ability to recover the target mental states in 11 of the 12 conditions suggests that the primary findings of Experiment 2a replicated overall. Taken together, Experiments 1, 2a, and 2b suggest that in this relatively constrained, forced-choice context, people can recover others' desires from their emotional reaction to events, but they can recover others' beliefs only when they observe reactions to both expected and observed outcomes. As noted, this is consistent with previous findings, suggesting that both children and adults are better at inferring others' desires than beliefs (Apperly & Butterfill, 2009; Astington & Gopnik, 1991; Saxe et al., 2004; Wellman, 2014; Wellman et al., 2001). It is also consistent with previous work suggesting that expressions of surprise can (at least when unmixed with valence) be an important cue to beliefs (Hadwin & Perner, 1991; Wellman & Banerjee, 1991). This study additionally highlights the role of a presence or absence of a change in valence as an important cue to others' beliefs: When there is a change of valence between when someone anticipates and observes an outcome, people infer a false belief; when there is no change, people infer a true belief.

In Experiments 3 and 4, we look at more complex cases of emotion inference, cases in which the agent causes (as well as observes) the events to which she is reacting. Previous computational work on theory of mind has either looked at the relationship between agents' actions, beliefs, and desires (e.g., Baker et al., 2009, 2017) without considering emotions, or it has looked at the relationship between an agent's emotional reactions and outcomes (Ong et al., 2015) without manipulating actions, beliefs, or desires. Here, we bridge these lines of work to provide a more unified account of theory of mind, looking at how people integrate observed actions, outcomes, and emotional reactions when making joint inferences about beliefs and desires. Experiments 3 and 4 are similar to Experiments 1 and 2a, respectively, except that in Experiments 3 and 4, the agent's actions cause the outcome to occur.

## 3.4. Experiment 3

In Experiment 3, as in Experiment 1, participants observe the agent's emotional reaction only to the final outcome of an event. In contrast to Experiment 1, the outcome of the event does not result from an external cause, but from the agent's action. Here, we look at how changing the causal role of the agent influences people's mental state inferences and whether our model captures human judgments.

### 3.4.1. Method
*3.4.1.1. Design and materials*: We use a scenario adapted from previous research (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010) in which two coworkers are visiting a chemical factory. One coworker (Grace) finds an unlabeled container of white powder and puts some of the powder in her colleague John's coffee. Grace's desire and belief are unspecified but constrained to two possibilities: Grace either wants John to die or live, and believes the powder is either poison or sugar. There are two possible outcomes: John either lives or dies after drinking the coffee. (See Supporting Information Text 1.2 for details.)

We use the same stimuli as in Experiment 1, with the same assumptions: If the outcome is consistent with Grace's desire, she should express positive emotions (and if inconsistent, negative); if the outcome is consistent with her belief, she should be unsurprised (and if inconsistent, surprised; see MacLaren & Olson, 1993; Hadwin & Perner, 1991; Repacholi & Gopnik, 1997; Skerry & Spelke, 2014; Stein & Levine, 1989; Wellman & Banerjee, 1991; Wellman & Woolley, 1990; Yuill, 1984; but see Krettenauer et al., 2008).

Additionally, to see to what extent the results were robust to details of the stimuli, we generated a separate set of 6-second movie stimuli (see https://osf.io/cdrbp/?view_only=b 3cb225cdbdc498caa900e7431322fda) by asking a professional actor, blind to the experimental hypotheses, to generate his own facial reactions given information about Grace's belief, desire, action, and the event outcome specified in each condition (see Supporting Information Text 2.2); we refer to this as Experiment 3 Supplemental.

In each of the eight conditions, Grace acts to put the powder into John's coffee. However, the prima facie likelihood of this action is different given different combinations of beliefs and desires. See Fig. 2(a). In Conditions 1–4, the observed action of putting powder into John's coffee is likely given Grace's stipulated belief and desire (e.g., if she thinks the powder is poison and wants John to die, it is likely that she would put the powder in his coffee). Thus, the mental state inferences supported by Grace's action are congruent with the mental state information used to generate Grace's emotional reaction. We categorize these conditions as "congruent" conditions. Conversely, in Conditions 5–8, the same action is performed but it is unlikely given Grace's stipulated belief and desire (e.g., if Grace thinks the powder is poison and wants John to live, it is unlikely that she would put the powder in his coffee). In these cases, the action is prima facie unlikely given the beliefs and desires used to generate Grace's emotional reaction; the plausibility of the action depends on entertaining hypotheses about the context external to the information provided in the stories (e.g., if she wants him to live and nonetheless puts what she believes to be poison in his coffee, she must have been at gunpoint or otherwise coerced; if she wants him to die and nonetheless puts what she believes to be sugar in his coffee, she must be biding her time and wanting to appear helpful). We categorize these conditions as "incongruent" conditions. We are interested in both the congruent and incongruent conditions because we want to see how people weigh and integrate different sources of potentially complementary or contradictory information when reasoning about others' mental states.

*3.4.1.2. Participants and procedure*: As in the preceding experiments, we used independent groups of participants to calibrate the model. Participants ($n = 57$) judged the prior over mental states, $P(B,D)$, and how likely it was that Grace would put the powder in John's coffee given each combination of Grace's belief and desire, $P(A|B,D)$. Separate groups of participants ($n = 55$) rated the likelihood of the photograph stimuli given Grace's belief, desire, action, and the event outcome specified in each condition, $P(R_1|B, D,A,O)$, and ($n = 51$) rated the likelihood of the movie stimuli.

The test participants ($n = 49$ for the photograph stimuli; $n = 52$ for the movie stimuli) judged the probability of each combination of Grace's belief and desire given her action, the event outcome, and her emotional reaction to the outcome, $P(B,D|A,O,R_1)$. See Supporting Information Text 3 for details.

### 3.4.2. Results and discussion
*3.4.2.1. Model calibration*: For ease of comparison with the preceding experiments, we report the results of the photograph stimuli first and in full. We provide the results of the movie stimuli second, and details can be found in Supporting Information Text 7. The prior probability of each combination of desire and belief was relatively uniform (Fig. 2(c)(i)). As anticipated, the action likelihood was in general higher for the mental states in the congruent conditions (Die&Poison, Live&Sugar) than in the incongruent conditions (Die&Sugar, Live&Poison) (see Fig. 2(c)(ii) and Supporting Information Texts 4.1 and 4.2 for detailed analyses). Participants' likelihood judgments for the photograph

stimuli in this scenario were similar to those in Experiment 1, reflecting the robustness of people's relative insensitivity to surprise when morphed with valence (see Fig. 2(c)(iv) and Supporting Information Text 4.4 for detailed analyses).

*3.4.2.2. Mental state inferences*: Participants' mental state inferences based on the photograph stimuli are reported in Fig. 3(c). See also Fig. 4(c) for the overall pattern. The analyses were identical to those in previous experiments. There was no main effect of Condition ($F(7, 1511) = .61$, $p = .748$) but a significant main effect of Mental State ($F(3, 1511) = 170.27$, $p < .001$) and a significant interaction between Condition and Mental State ($F(21, 1511) = 25.56$, $p < .001$). The main effect of Mental State was significant in all conditions (all $F$s > 13.91, all $p$s < .001). In contrast to Experiment 1 (in which participants inferred desires but did not differentiate between the two beliefs), in Experiment 3, participants rated the target combination of beliefs and desires higher than all other combinations in the congruent conditions (Conditions 1–4: all $z$s > 6.64, all $p$s < .001). In the incongruent conditions (Conditions 5–8), participants correctly chose the desire corresponding to the valence of the facial expression. However, instead of either choosing the belief used to generate the emotional expression or failing to distinguish the two beliefs (as in Experiment 1), participants chose the belief congruent with the inferred desire given the action, rating it higher than the target in all four conditions (all $z$s > 5.76, all $p$s < .001). Consider Condition 8, for example; this was the condition in which Grace wanted John to live, believed the powder was poison, and John unexpectedly lived. On seeing the outcome, Grace's expression was both positive and surprised. Participants (correctly) inferred that Grace wanted John to live but (incorrectly) inferred that Grace believed the powder was sugar. That is, even though Grace's reaction to the final outcome was *surprised*, participants favored the belief that the powder was sugar, a belief that rendered the outcome unsurprising but also rendered it congruent with Grace's desires given her action (i.e., that she wanted him to live and put the powder in his coffee).

One-sample *t*-tests showed similar results. Participants uniquely rated the target mental state significantly above 50 in the congruent conditions (Conditions 1–4: $t_1(48) = 11.00$, $p_1 < .001$; $t_2(49) = 4.97$, $p_2 < .001$; $t_3(49) = 3.99$, $p_3 = .007$; $t_4(48) = 4.30$, $p_4 < .001$). In the incongruent conditions, only the mental state with the correct desire and the belief congruent with that desire given the action was rated above 50 in Conditions 5–7 ($t_5(49) = 7.24$, $p_5 < .001$; $t_6(49) = 5.45$, $p_6 < .001$; $t_7(49) = 4.54$, $p_7 < .001$), with a non-significant trend in the same direction in Condition 8 ($t(49) = 1.92$, $p = 1.000$; by comparison, the other three mental states were rated significantly below 50, all $t$s < −4.25, all $p$s < .001).

Model predictions were generated using the independent raters' judgments of the prior probability of each combination of mental states, the likelihood of the action, and the likelihood of the facial reactions according to Eq. 1 (but omitting the *Reaction$_0$* term, see Supporting Information Text 5.3), $P(B,D|A,O,R_1)$. Fig. 3(c) shows the model predictions of people's inferences about the mental states underlying the photograph stimuli. Like people, the model gave the highest probability to the desire that was in fact used to generate the emotional reaction. However, also like people, the model predicted the beliefs

that were congruent with the desires given the action in all conditions (i.e., failing to distinguish the beliefs in Conditions 1 and 2 from Conditions 5 and 6, or Conditions 3 and 4 from Conditions 7 and 8; see Fig. 3(c)). These predictions result from conditioning on the observed *Action*; the conditional action likelihood favors Die&Poison and Live&Sugar, biasing the posterior inferences toward combinations of mental states that are congruent with acting in all conditions. The model's inferences correlated well with the behavioral results ($r = .985$).

We conducted the same analyses for the movie stimuli (Experiment 3 Supplemental). The behavioral results replicated those from the photograph stimuli in all respects (see Supporting Information Text 7), including the insensitivity to the link between surprise and belief in people's likelihood judgments. The correlation between the model predictions and participants' mental state inferences was 0.908. These results suggest that the findings are robust to variations in the stimuli.

Experiment 3 suggests that people perform a particularly sophisticated kind of mental state inference: integrating observed emotional reactions with actions to jointly infer beliefs and desires. Critically, note that neither inferences from the observed action alone nor from the emotional reaction alone can explain the pattern of results in Experiment 3. In Experiment 1 (where the agent did not act), participants recovered the agent's desires but largely did not differentiate the two candidate beliefs. By contrast, in Experiment 3 (where the agent did act), participants recovered both the agent's desires and beliefs in the four congruent conditions (Conditions 1–4), but in the incongruent conditions (Conditions 5–8), they were biased towards the beliefs congruent with the desires given the actions. This does not imply, however, that a participant's inferences can be explained by a model of theory of mind that excludes the agent's emotional reactions and includes only her actions. Grace's context and action were identical throughout; nothing distinguished Conditions 1 and 3, or 2 and 4 except Grace's emotional reaction. Nonetheless, participants inferred distinct combinations of desires and beliefs. Again, our Bayesian model captured participants' judgments.

## 3.5. Experiment 4

Experiment 4 is identical to Experiment 3 except that (as in Experiments 2a and b) we give participants information about the agent's reactions to both the expected and observed outcomes. We predict that this additional evidence may help people recover the target mental states in the incongruent conditions so that people should be more likely to recover the target mental states in Experiment 4 than Experiment 3. However, if people integrate the evidence with the likelihood of the agent's actions, then they should still have some difficulty recovering the target mental states in the incongruent conditions (when the actions are unlikely given these mental states). Thus, we additionally predict that people's ability to recover the target mental states in the incongruent conditions of Experiment 4 (where Grace acts to generate the outcome) should be more fragile than in Experiments 2a and 2b (where she merely observes the outcome). As in the preceding studies, we look at whether our model quantitatively captures human performance.

### 3.5.1. Method

*3.5.1.1. Design and materials*: We used the same chemical-factory scenario as in Experiment 3 and the same photograph stimuli used in Experiment 2a.

*3.5.1.2. Participants and procedure*: To calibrate the model, participants ($n = 58$) rated the likelihood of $Reaction_0$, $P(R_0|B,D,A)$. Otherwise, the model calibration judgments from Experiment 3 were reused here because the eliciting conditions for all the other model calibration judgments (i.e., the prior probability of mental states, the likelihood of actions, and the likelihood of $Reaction_1$) were identical to those in Experiment 3.

The test participants ($n = 53$) judged the probability of Grace's belief and desire given her action, the outcome of her action, and her reactions to the anticipated and observed outcomes, $P(B,D|A,O,R_0,R_1)$. See Supporting Information Text 3.

### 3.5.2. Results and discussion

*3.5.2.1. Model calibration*: The likelihood of $Reaction_0$ is reported in Fig. 2(c)(iii). Similar to the calibration results in Experiment 2a, the positive expressions (those used in Conditions 1–4) were rated higher for the two mental states in which Grace's desire would be fulfilled by her action based on her belief (Die&Poison, Live&Sugar) than those in which it would not (Die&Sugar, Live&Poison). The negative expressions used in Conditions 5–8 showed roughly the opposite pattern. That is, as we had assumed, participants expected the agent to express positive emotions when the expected outcome of her action would fulfill her desire, and negative emotions when it would not (see Supporting Information Text 4.3 for detailed analyses).

*3.5.2.2. Mental state inferences*: People's mental state inferences are reported in Fig. 3(d). See Fig. 4(d) for the overall pattern. The mixed effects model showed no main effect of Condition ($F(7, 1600) = .36$, $p = .923$) but a significant main effect of Mental State ($F(3, 1600) = 260.53$, $p < .001$) and a significant interaction between Condition and Mental State ($F(21, 1600) = 22.93$, $p < .001$). The main effect of Mental State was significant in all conditions (all $F$s $> 12.01$, all $p$s $< .001$) except Condition 7 ($F(3, 153) = 2.63$, $p = .052$). Further analyses showed that, as in Experiment 2a, participants rated the target mental state significantly higher than the other mental states in the congruent conditions (Conditions 1–4: all $z$s $> 10.54$, all $p$s $< .001$). However, as predicted, the action likelihood affected participants' responses in the incongruent conditions so that, in contrast to Experiment 2a, participants struggled to recover the agent's mental states in the incongruent conditions. Participants successfully rated the target mental state (i.e., the combination of belief and desire that was used to generate the emotional reaction) higher than the other three mental sates in Condition 8. However, in Conditions 5 and 6, they correctly identified the target desire but were biased towards the belief that was congruent with the action, rating this mental state combination higher than the target (both $z$s $> 3.72$, both $p$s $< .005$). In Condition 7, they did not differentiate the target mental state from the other three mental states (all $|z|$s $< 2.66$, all $p$s $> .190$).

A similar pattern was found using one-sample $t$-tests. As in Experiment 2a, participants uniquely rated the target mental state significantly above 50 in the congruent conditions (Conditions 1–4: $t_1(52) = 8.89$, $p_1 < .001$; $t_2(51) = 8.35$, $p_2 < .001$; $t_3(52) = 6.86$, $p_3 < .001$; $t_4(51) = 7.26$, $p_4 < .001$). In the incongruent conditions, there was a non significant trend towards correctly identifying the target mental state only in Condition 7 ($t(51) = -1.75$, $p = 1.000$; the other three mental states were rated significantly below 50: all $t$s $< -3.51$, all $p$s $< .030$). Participants uniquely rated the mental state with the correct desire and the belief congruent with the action significantly above 50 in Condition 5 ($t(52) = 3.70$, $p = 0.017$) and showed a non significant trend in the same direction in Condition 6 ($t(52) = .22$, $p = 1.000$ with the other three mental states rated significantly below 50: all $t$s $< -4.39$, all $p$s $< .001$). In Condition 8, the two mental states with the correct desire were rated at chance (both $|t|$s $< 2.72$, both $p$s $> .286$); the remaining two mental states were rated significantly below 50 (both $t$s $< -8.47$, both $p$s $< .001$).

We can compare people's judgments with the predictions of our Bayesian model, this time incorporating $R_0$: $P(B,D|A,O,R_0,R_1)$ (see Supporting Information Text 5.4). Again, the correlation between the model predictions and human judgments ($r = .950$) was high.

Together with the previous experiments, the results of Experiment 4 suggest that people integrate observed actions and emotional reactions to produce probabilistic inferences about others' beliefs and desires. Given only an agent's emotional reaction to the outcome of an observed event, participants were able to recover the agent's desires, but not her beliefs (Experiment 1). However, given her emotional reaction to both the expected and actual outcome of an observed event, participants successfully recovered both the agent's beliefs and desires (Experiment 2). Adding information about the agent's actions had a paradoxical effect, making participants both more *and* less able to recover the agent's mental states. When the inferred beliefs and desires were congruent with the agent's action, a single emotional reaction sufficed for participants to recover both mental states (cf: the failures of Experiment 1 and the successes in Experiment 3, Conditions 1–4). However, when the beliefs and desires were improbable given the agent's action, participants were unable to recover them, even given information about the agent's emotional reaction to both the observed and expected outcome (cf: the failures in Experiment 4 and the successes in Experiment 2, Conditions 5–8). See Fig. 4. Collectively, these results suggest that people integrate information about an agent's emotional reactions and their actions.

## 4. Comparison with other models

This integration is well characterized by our probabilistic inference model. In our ideal observer model, inferences about others' beliefs and desires from observations of their behavior (e.g., their emotional expressions and actions) are based on inverting a forward model of how beliefs and desires generate that behavior. How does our model compare with alternative models?

In the spirit of classic accounts of theory of mind that do not take into account emotional reactions, can a model (No-Emotion model) that combines the prior probabilities of mental states with only the likelihood of the agent's actions predict the mental state judgments in our studies? What about the complementary alternative, a model (No-Action model) that looks only at how beliefs and desires determine emotional reactions to outcomes without taking into account how these mental states also inform agents' actions? Alternatively, perhaps people's inferences are not based on a causal model at all, but rather on some learned associations between event features and types of mental states (Event-Features model)? In this section, we compare each of these alternative models with our full Bayesian model.

## 4.1. No-Emotion model

This model is based on the possibility that mental state inference is not integrated with an intuitive theory of emotion and is strictly the provenance of classical "rational actor" theory of mind. That is, for the purposes of mental state inference, people may represent beliefs and desires as determinants only of agents' actions (i.e., the classic theory of mind model) without taking into how these mental states might cause emotional reactions. To evaluate this account, we generated new model predictions by dropping all of the emotional reaction terms (i.e., $P(R_0|B,D,A)$, $P(R_1|B,D,A,O)$) in our original Bayesian model. The correlations between these model predictions and the behavioral data were 0.147, 0.114, 0.085, 0.528, and 0.379, for Experiments 1, 2a, b, 3, and 4, respectively. All of these correlations were significantly lower than those of the full Bayesian model (all $ps < .05$), according to a bootstrapped hypothesis test, randomly sampling 1/4 of the data points in each of the 10,000 iterations. This suggests that a model that fails to consider emotional reactions is not sufficient to capture people's inferences in this task. Intuitively, the failure of the No-Emotion model should be unsurprising given that participants successfully recovered agents' beliefs and desires in the absence of any actions by the agent (e.g., Experiment 2a and 2b) and distinguished mental states that were equally consistent with rational action (Die&Poison and Live&Sugar) in the congruent conditions of Experiments 3 and 4.

## 4.2. No-Action model

The No-Action model reflects a complementary proposal to the No-Emotion model, namely that when emotional reactions are observed, mental state inference becomes purely the provenance of a naïve theory of emotion, independent of a theory of how these same mental states determine agents' actions. To test this proposal, we drop the action term ($P(A|B,D)$) from the original Bayesian model. The model predictions do not change for Experiments 1, 2a and 2b (where the agent merely observes the events) but do change for Experiments 3 and 4 (where there is an action performed by the agent). The correlations between the model predictions and the behavioral data were 0.843 and 0.893 for Experiments 3 and 4, respectively. Using the same bootstrapped hypothesis test described above, the correlation was significantly lower than that of the full Bayesian model in

Experiment 3 ($p$ = .018) and was as high as that of the full model in Experiment 4 ($p$ = .144). The relatively good performance of the No-Action model in Experiments 3 and 4 compared to the No-Emotion model is not surprising given that the emotional reactions differed in every one of the eight experimental conditions, whereas the action did not vary at all. Consequently the action term only scales the overall model predictions for each distinct mental state (Fig. 2(c)(ii)), independent of condition, whereas the emotional-reaction term differentially influences model predictions for every mental state in every condition (Fig. 2(c)(iii) and (iv)). Taken together, across all our experiments, only the full Bayesian model that considers both actions and emotional reactions as informative effects of underlying mental states provides a complete account of people's judgments. Again, intuitively, this can be seen in the behavioral results in which adding information about the agent's actions made participants relatively more capable of distinguishing (congruent) beliefs and desires from a single emotional reaction (Experiment 3 vs. 1, Conditions 1-4) but less capable of distinguishing desires and beliefs incongruent with the actions even when given the agent's reaction to both the expected and observed outcome (Experiment 4 vs. 2, Conditions 5–8).

### 4.3. Event-Features model

As noted, people might not invoke a causal model of agent's minds at all, but instead use "model-free," data-driven cues derived from past experience. That is, people may learn from experience that some features of events (including agents' emotional reactions to them) statistically relate to certain types of mental states and use those learned statistics to make predictions about new events. For example, in Experiment 1, the event features may include whether the agent performs an action, what the outcome is, and the perceptual features of her emotional reaction; these features, not constructed as causal models per se, may be integrated in a regression-style model with learned weights to generate the probable mental state as an output.

To formally evaluate this Event-Features account, we built a feature-based regression model that attempted to directly predict people's mental-state inferences across Experiments 1–4. The features used were the action (i.e., whether the agent acts to cause the event), outcome (i.e., whether John lives or dies), and the perceptual emotion features (i.e., happy, sad, angry, surprised, fearful, disgust, unhappy) of our photograph stimuli (see Supporting Information Text 6 and Tables S1 and S2). Because the perceptual features were not independent (e.g., sad and happy features were negatively correlated), we performed dimensionality reduction using principal component analysis (PCA) on the features of $Reaction_0$ and $Reaction_1$. This yielded a basis of two principal components for $Reaction_0$ and three principal components for $Reaction_1$. We trained the model to map these features to desired outputs using multinomial regression. The desired outputs were the sum of participants' judgments of each of the four mental states in every condition (44 conditions in total across the four experiments).

We used bootstrap cross-validation (BSCV) (Cohen, 1995) to evaluate the performance of this model-free account. We generated 10,000 random, non overlapping splits of all 44

experimental conditions into training sets of 33 conditions, and testing sets of 11 conditions. For each training set, we used multinomial regression to map the features to the human data. We then computed the Pearson correlation of the model with the human data for the corresponding test set, using the parameters fit from the training set. The median correlation on the test data was 0.583 (95% CI 0.25 0.77). For model comparison, we also bootstrapped the correlation of the Bayesian model using the same random test sets. The median bootstrapped correlation of the Bayesian model was 0.957 (95% CI 0.92 0.98). The correlation of the model-free account with the human data was significantly lower than that of the Bayesian model, according to a bootstrapped hypothesis test ($p < .001$).

We do not mean to suggest that event features learned through experience play no role in mental state understanding. However, our results argue strongly against the sufficiency of a purely model-free, data-driven account. Together with the results of the No-Emotion model and the No-Action model, we suggest instead that our ability to recover others' beliefs and desires requires richly structured, generative models of others' mental states, actions, and emotional reactions to events.

## 5. General discussion

The current results suggest both the sophistication and limitations of people's ability to recover mental states from observed emotional reactions. On one hand, people successfully recovered an agent's previously unknown beliefs and desires in some conditions of all the experiments, and all the conditions of one experiment (Experiments 2a and b). Moreover, across four separate experiments and variations in both experimental scenarios and stimuli, participants' inferences were also consistent with our ideal observer model (Experiments 1–4). This is impressive given that the inferences participants were asked to make in this study were arguably more complex than those in many previous studies of theory of mind: The context (and actions when applicable) were identical in all conditions, participants had very sparse evidence for the agent's emotional reactions, and participants were asked to simultaneously infer the agent's beliefs and desires. On the other hand, despite a very restricted hypothesis space—only two possible beliefs and two possible desires—people were only able to infer unique combinations of the agent's beliefs and desires when they observed the agent's emotional reaction to both an expected and observed outcome (Experiments 2a and 2b) or when the agent's action and emotional reaction were likely given the target beliefs and desires (the congruent conditions of Experiments 3 and 4).

Given that the inferences were made about a stranger, and the outcome, context and action were not in themselves differentially informative (constraints that hold for many real world scenarios), the results suggest that observed emotional expressions provide a valuable entrée into mental state inferences. However, it is equally noteworthy that participants were unable to reliably infer others' beliefs when the mental states were unlikely given the action. As noted, a large body of research suggests that belief inferences are challenging, even for adults (Saxe et al., 2004; Wellman et al., 2001; see Apperly &

Butterfill, 2009; Astington & Gopnik, 1991; and Wellman, 2014, for reviews and discussion). The current results suggest that people have particular difficulty in attributing beliefs that imply that someone consciously acted in a way that is inconsistent with her desires. Although such contexts may be relatively rare, they are far from non existent (e.g., consider cases of coercion, addiction, or compulsion). The results of this study (in particular the incongruent conditions of Experiment 3) suggest that in such contexts, we may confabulate beliefs and desires that are consistent with an observed action even when the agent's emotional expression might otherwise belie this judgment. More broadly, however, the results of the current studies suggest that the principle of rational action—the assumption that agents act in ways that are consistent with their desires given their beliefs (see Gergely & Csibra, 2003, for a review)—can act as a double-edged sword: It may (misleadingly) bias our inferences towards mental states that are probable given the agent's action; however, that same bias may support our ability to draw accurate inferences from sparse data when the information we have is consistent but limited.

In this study, we failed to find any difference between morphed facial expressions and basic emotions (Experiment 2a vs. 2b) or photographs and movies as cues to mental states (Experiment 3 vs. Experiment 3 Supplemental). Intuitively, richer sources of information about an agent's emotional reactions seem likely to support richer, and more accurate, mental state inferences. At the same time, the prevalence of genuinely mixed emotions, and people's tendency to mask emotions in social contexts, might complicate real-world inferences about others' mental states. Future research might look at how different kinds of information about emotional reactions (e.g., facial expressions, vocalizations, body postures, and dynamic changes in these expressions over time), and pressure to conceal or reveal emotions might affect mental state inferences.

Future research might also look at the impact of cultural variability on our findings. There have been fierce debates about the universality of both the expression of emotions and the interpretation of emotional expressions across cultures (e.g., Darwin, 1872/1965; Ekman & Friesen, 1971; Matsumoto & Willingham, 2009; Elfenbein et al., 2007). The degree to which cultural differences impact people's inferences about mental states from emotional reactions remains an important area for future research. We suspect that although culture will surely affect which emotional reactions and actions people think are probable given particular beliefs and desires, the ability to draw inferences about others' beliefs and desires given information about their actions and emotional reactions is likely to be universal.

People's ability to distinguish mental states based on emotional expressions varied across the four experiments; however, participants' inferences in all four studies were quantitatively well fit by our model (all correlations at a level of $r = .950$ or above, corresponding to at least 90% of the variance explained). By including different terms in Eq. 1 (corresponding to different nodes in the graphical model of Fig. 1), the model was able to characterize the inferences people made from an agent's emotional reaction to an outcome she only observed (Experiments 1 and 2) and an outcome she caused (Experiments 3 and 4) and from both single emotional reactions (Experiments 1 and 3) and reactions to both expected and observed outcomes (Experiments 2 and 4). Similar principles

of Bayesian inference have been shown to govern fast and accurate inferences in perception, language processing, and other core domains of cognition (Chater, Tenenbaum, & Yuille, 2006). These models have been especially powerful as quantitative accounts of perceptual cue integration both within and across sensory modalities (Battaglia & Schrater, 2007; Beierholm, Quartz, & Shams, 2009; Ernst & Banks, 2002; Körding & Wolpert, 2004; Weiss, Simoncelli, & Adelson, 2002). The principles of Bayesian inference have also been proposed as a potential unifying framework for cue integration in social cognition (Wolpert, Doya, & Kawato, 2003; Zaki, 2013). A recent study has tested this in the emotion domain, showing that emotion cue integration (i.e., reasoning about emotions from facial expressions, utterances and outcomes) can be well characterized by Bayes' rule (Ong et al., 2015). Our study bridges theory of mind research and emotion attribution, suggesting that mental-state inferences from multiple cues (i.e., context, actions, outcomes, and emotional reactions) may be likewise the product of evolutionarily or developmentally tuned perceptual machinery that computes accurate inferences under uncertainty by integrating multiple sources of information in near-optimal ways.

An important limitation of our present model is that although it captures the high-level structure of the causal relationships between beliefs, desires, actions, outcomes, and emotional reactions in people's intuitive psychology, it does not represent the fine-grained functional form of these relationships. We have not attempted to specify the precise mechanism by which people represent the causal relationship between mental states, contextual variables, and specific emotional reactions; these fine-grained dependencies are represented only implicitly in our framework in the components of the forward model (the terms on the right-hand side of Eq. 1). Explicitly modeling how people represent these fine-grained generative relationships remains an important task for future work.

Importantly, however, the present work suggests that the high-level causal structure of these relationships is sufficient to produce accurate quantitative "inverse" models of mental-state inference. It appears that our naïve theory of emotional reactions is structurally and causally intertwined with our theory of mind in a way that allows both forward prediction from an agent's beliefs and desires to her emotional expressions, and backward inference from emotional expressions to beliefs and desires, with a degree of quantitative internal coherence suggestive of highly optimized probabilistic inference mechanisms.

## Acknowledgments

## Note

1. We are grateful to an anonymous reviewer for pointing out that people's judgments about emotional responses to goal fulfillment are not always this straightforward.

In particular, older, but not younger, children recognize that someone who fulfills her goal by committing a moral violation may be remorseful rather than happy; thus, younger children accept "happy victimizers," whereas older children judge a moral violation more harshly if the perpetrator is happy rather than sad after committing it (e.g., Nunner-Winkler & Sodian, 1988; Krettenauer, Malti, & Sokol, 2008, for review).

## References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970.

Astington, J. W., & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology*, *9*(1), 7–31.

Bachorowski, J. A., & Owren, M. J. (2003). Sounds of emotion. *Annals of the New York Academy of Sciences*, *1000*(1), 244–265.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behavior*, *1*(0064).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113* (3), 329–349.

Barden, R. C., Zelko, F. A., Duncan, S. W., & Masters, J. C. (1980). Children's consensual knowledge about the experiential determinants of emotion. *Journal of Personality and Social Psychology*, *39*(5), 968–976.

Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child Development*, *62*(2), 385–395.

Barrett, L. F. (2011). Was Darwin wrong about emotional expressions? *Current Directions in Psychological Science*, *20*(6), 400–406.

Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (Eds.). (2016). *Handbook of emotions*. New York: Guilford.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286–290.

Battaglia, P. W., & Schrater, P. R. (2007). Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *The Journal of Neuroscience*, *27*(26), 6984–6994.

Beierholm, U. R., Quartz, S. R., & Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of Vision*, *9*(5), 1–9.

Bender, P. K., Pons, F., Harris, P. L., & de Rosnay, M. (2011). Do young children misunderstand their own emotions? *European Journal of Developmental Psychology*, *8*(3), 331–348.

Bradmetz, J., & Schneider, R. (1999). Is little red riding hood afraid of her grandmother? cognitive vs. emotional response to a false belief. *British Journal of Developmental Psychology*, *17*(4), 501–514.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, *70*(2), 205–218.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Clore, G. L., & Ortony, A. (2013). Psychological construction in the OCC model of emotion. *Emotion Review*, *5*(4), 335–343.

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. Vol. 139. Cambridge, MA: MIT Press.

Crivelli, C., Russell, J. A., Jarillo, S., & Fernández-Dols, J. M. (2016). The fear gasping face as a threat display in a Melanesian society. *Proceedings of the National Academy of Sciences*, *113*(44), 12403–12407.

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, *12*(5), 1085–1101.

Darwin, C. (1965). *The expressions of the emotions in man and animal*. Chicago: University of Chicago Press. (Original work published 1872)

De Rosnay, M., Pons, F., Harris, P. L., & Morrell, J. (2004). A lag between understanding false belief and emotion attribution in young children: Relationships with linguistic ability and mothers' mental-state language. *British Journal of Developmental Psychology*, *22*(2), 197–218.

Denham, S. A., Zoller, D., & Couchoud, E. A. (1994). Socialization of preschoolers' emotion understanding. *Developmental Psychology*, *30*(6), 928–936.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454–E1462.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion*, *7*(1), 131–146.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fabes, R. A., Eisenberg, N., McCormick, S. E., & Wilson, M. S. (1988). Preschoolers' attributions of the situational determinants of others' naturally occurring emotions. *Developmental Psychology*, *24*(3), 376–385.

Fontaine, J. R., Poortinga, Y. H., Setiadi, B., & Markam, S. S. (2002). Cognitive structure of emotion terms in Indonesia and The Netherlands. *Cognition & Emotion*, *16*(1), 61–86.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*(12), 1050–1057.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive psychology*, *75*, 80–96.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

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

Gnepp, J. (1983). Children's social sensitivity: Inferring emotions from conflicting cues. *Developmental Psychology*, *19*(6), 805–814.

Gnepp, J., McKee, E., & Domanic, J. A. (1987). Children's use of situational information to infer emotion: Understanding emotionally equivocal situations. *Developmental Psychology*, *23*(1), 114–123.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Hadwin, J., & Perner, J. (1991). Pleased and surprised: Children's cognitive theory of emotion. *British Journal of Developmental Psychology*, *9*(2), 215–234.

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

Harris, P. L. (1983). Children's understanding of the link between situation and emotion. *Journal of Experimental Child Psychology*, *36*(3), 490–509.

Harris, P. L., Guz, G. R., Lipian, M. S., & Man-Shu, Z. (1985). Insight into the time course of emotion among Western and Chinese children. *Child Development*, *56*(4), 972–988.

Harris, P. L., Johnson, C. N., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition & Emotion*, *3*(4), 379–400.

Harris, P. L., Olthof, T., Terwogt, M. M., & Hardman, C. E. (1987). Children's knowledge of the situations that provoke emotion. *International Journal of Behavioral Development*, *10*(3), 319–343.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*(2), 288–299.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Keltner, D., Tracy, J., Sauter, D. A., Cordaro, D. C., & McNeil, G. (2016). Expression of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed.) (pp. 467–482). New York: Guilford.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Krettenauer, T., Malti, T., & Sokol, B. W. (2008). The development of moral emotion expectancies and the happy victimizer phenomenon: A critical review of theory and application. *International Journal of Developmental Science*, *2*(3), 221–235.

Lagattuta, K. H. (2005). When you shouldn't do what you want to do: Young children's understanding of desires, rules, and emotions. *Child Development*, *76*(3), 713–733.

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development*, *72*(1), 82–102.

Lagattuta, K. H., Wellman, H. M., & Flavell, J. H. (1997). Preschoolers' understanding of the link between thinking and feeling: Cognitive cuing and emotional change. *Child Development*, *68*(6), 1081–1104.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist*, *46*(8), 819.

Lee, D. H., & Anderson, A. K. (2016). Form and function in facial expressive behavior. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed.) (pp. 495–509). New York: Guilford.

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

MacLaren, R., & Olson, D. (1993). Trick or treat: Children's understanding of surprise. *Cognitive Development*, *8*(1), 27–46.

Matsumoto, D., & Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology*, *96*(1), 1–10.

Meeren, H. K., van Heijnsbergen, C. C., & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16518–16523.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124.

Nunner-Winkler, G., & Sodian, B. (1988). Children's understanding of moral emotions. *Child Development*, *59*(5), 1323–1338.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge, England: Cambridge University Press.

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *1* (2), 127–152.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Rieffe, C., Terwogt, M. M., & Cowan, R. (2005). Children's understanding of mental states as causes of emotions. *Infant and Child Development*, *14*(3), 259–272.

Ruffman, T., & Keenan, T. R. (1996). The belief-based emotion of suprise: The case for a lag in understanding relative to false belief. *Developmental Psychology*, *32*(1), 40–49.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293, 317). Hillsdale, NJ: Erlbaum.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology*, *32*(1), 76–92.

Scherer, K. R., & Meuleman, B. (2013). Human emotion experiences can be predicted on theoretical grounds: Evidence from verbal labeling. *PLoS ONE*, *8*(3), e58166.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. New York, NY: Oxford University Press.

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

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*(4), 341–351.

Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, *110*(1), 70–75.

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology*, *25*(15), 1945–1954.

Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.

Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and the emotions. *Cognition & Emotion*, *7*(3–4), 233–269.

Stein, N. L., & Levine, L. J. (1989). The causal organisation of emotional knowledge: A developmental study. *Cognition & Emotion*, *3*(4), 343–378.

Taylor, D. A., & Harris, P. L. (1983). Knowledge of the link between emotion and memory among normal and maladjusted boys. *Developmental Psychology*, *19*(6), 832–838.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Trabasso, T., Stein, N. L., & Johnson, L. R. (1982). Children's knowledge of events: A causal analysis of story structure. *The Psychology of Learning and Motivation*, *15*, 237–282.

Vaish, A., & Woodward, A. (2010). Infants use attention but not emotions to predict others' actions. *Infant Behavior and Development*, *33*(1), 79–87.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

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

Wellman, H. M., & Banerjee, M. (1991). Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology*, *9*(2), 191–214.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, *30*(3), 239–277.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wellman, H. M., Phillips, A. T., & Rodriguez, T. (2000). Young children's understanding of perception, desire, and emotion. *Child Development*, *71*(4), 895–912.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

Widen, S. C. (2016). The development of children's concepts of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed.) (pp. 307–318). New York: Guilford.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*(2), 291–312.

Widen, S. C., & Russell, J. A. (2010). Differentiation in preschooler's categories of emotion. *Emotion*, *10*(5), 651.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1431), 593–602.

Wu, Y., & Schulz, L. E. (2017). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*. https://doi.org/10.1111/cdev.12759 First published online on 8 March.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–6758.

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology*, *2*(1), 73–81.

Zaki, J. (2013). Cue integration a common framework for social cognition and physical perception. *Perspectives on Psychological Science*, *8*(3), 296–312.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two the interpersonal nature of empathic accuracy. *Psychological Science*, *19*(4), 399–404.

## Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

**Fig. S1.** Prototypical facial expressions of the six basic emotions.

**Fig. S2.** The structure of the tasks.

**Fig. S3.** The likelihood of the movie stimuli, people's mental state inferences and model predictions in Experiment 3 Supplemental.

**Table S1.** The creation and assessment of the photograph stimuli. (See Supporting Information Text 2.1.1 for detailed explanation of the components of these faces.)

**Table S2.** The assessment of the prototypical facial expressions used in Experiment 2b.