Joint inferences of belief and desire from facial expressions

Yang Wu (yangwu@mit.edu), Chris L. Baker (clbaker@mit.edu),
Joshua B. Tenenbaum (jbt@mit.edu), Laura E. Schulz (lschulz@mit.edu)

Department of Brain and Cognitive Sciences, MIT
77 Massachusetts Avenue, Cambridge, MA 02139 USA

Abstract

Theory of mind research has looked at how learners infer an agent's unobservable mental states from observable actions. However, such research has tended to neglect another observable source of data: the agent's reactions to events. In particular, the agent's facial reactions might provide important information about her mental states that are otherwise ambiguous given her actions. Here we present a Bayesian framework and a behavioral study testing how adults use an agent's facial reactions to reason backward about her beliefs and desires. We found that participants' joint inferences of belief and desire from facial expressions were predicted by a Bayesian model analysis, based on integrating the likelihoods of the observed facial reactions and the observed action with their prior over mental states. We argue that people's naïve theory of emotional reactions is structurally and causally intertwined with theory of mind in a way that allows forward prediction and backward inference.

Keywords: Theory of mind; appraisal theory; emotion; facial expression; Bayesian inference

Introduction

Human beings are adept at inferring others' mental states given sparse observations. One of the mysteries that has intrigued cognitive scientists for decades is what representations make this inference so efficient and accurate. Many studies in theory of mind have focused on how people infer beliefs and desires from observed actions. For example, if Sally reaches for a container, what can we infer about her beliefs and desires? A powerful basis for such inferences is the assumption of rational action – that agents act to fulfill their desires as efficiently as possible in accordance with their beliefs about the world. If we know that Sally believes there are cookies within the container, we may infer that Sally wants cookies based on her reaching behavior; conversely, if we know that Sally wants cookies, it is plausible that she believes there are cookies in the container.

Studies suggest that even infants can infer the desire underlying an observed action (when the belief is directly or indirectly given by the context) or the belief underlying an action (when the desire is directly or indirectly given) (e.g. Csibra, Bıró, Koós, & Gergely, 2003; Gergely, Nádasdy, Csibra, & Bíró, 1995; see Gergely & Csibra, 2003 for a review). Other work suggests that adults and older children can jointly infer an agent's desires and beliefs given a sequence of actions in which the agent approaches and retreats from potential goals (Baker, Saxe, & Tenenbaum, 2011; Richardson, Baker, Tenenbaum, & Saxe, 2012).

Often, however, the information available to observers about agents' actions, beliefs, and desires may be much more limited. We may for example arrive in the middle of a scene, seeing someone we do not know engage in a single action. For instance, we might see someone look up as another person approaches. This information is relatively sparse; inferring the agent's belief and desire from the action is nearly impossible, much like trying to solve one equation with two unknown variables. The observation of a simple action is not informative enough to discriminate different mental states.

However, these kinds of actions are typically accompanied by another kind of observable response: an emotional reaction. Emotional reactions – often manifest as facial expressions – intuitively seem to provide rich evidence about agents' mental states, arguably simplifying the theory of mind inference. If, for example, the agent frowns, we might infer that she knows the person and doesn't like him; if she smiles we might infer that she knows and likes him, and if she has no emotional response, we will probably infer that she doesn't know the approaching person at all.

There is of course a large literature looking at emotion per se (e.g. Ekman, 1992; Lazarus, 1991; Vuilleumier, 2005). However, this literature has remained relatively disconnected from the theory of mind literature. Perhaps the most relevant work connecting emotion to other cognitive states comes from appraisal theory, a theory of emotion suggesting that an individual's evaluation of events plays a crucial role in eliciting and differentiating emotions (e.g., Ellsworth & Scherer, 2003). Although different appraisal theories differ in the appraisal dimensions that are at stake (e.g. desirability, certainty, causal attribution, coping potential), most of these theories make reference to the agents' beliefs and desires (either explicitly or implicitly) as influences on people's evaluation of events and thus the generation of their emotional reactions (e.g. Lazarus, 1991; Ortony, 1990; Scherer, 1984).

Critically however, 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 either the individual herself might think about the causes of her emotional states, or how observers might use an agents' emotional reactions to reason backward about the mental states (the beliefs and desires) that generated them via appraisal. This is our goal here. We hypothesize that people have an intuitive theory of emotional responses that is at least coarsely analogous to appraisal theory, and they can use this intuitive theory to integrate observations of actions, outcomes and emotional responses to make rational inferences about agents' mental states. We model this intuitive theory formally and quantitatively evaluate its predictions with human judgments in four experiments.

We begin by specifying a simple probabilistic generative model of how an agent's appraisal of a situation – her belief and desire about an event – might lead to an emotional reaction. We then use that to analyze how an observer might reason backward (in a Bayesian fashion) from the emotional reaction to the belief and desire that generated it. Our focus in this paper is on the backward inference from the emotional reaction to the mental states (belief and desire) involved in the cognitive appraisal of the event. To preserve this focus, we restrict our study to emotional reactions revealed on others' faces. We use facial expressions because they are directly observable, because they can change dynamically over time, and because the understanding of facial expressions is less constrained by verbal fluency than understanding emotion words or descriptions.

Additionally, facial expressions have been well-studied in the literature. Considerable work has looked both at the relationship between facial expressions and emotions and at how people can use facial expressions to infer emotional states (e.g. Calder et al., 2003; Ekman, 1993). Some studies in this area (Carroll & Russell, 1996; Meeren, van Heijnsbergen, & de Gelder, 2005) suggest that given appropriate contextual cues, normal adults are very good at inferring emotions from facial expressions, but they may struggle if the facial expression is not well-predicted by the context. Here we do not look at whether people can infer emotions from facial expressions and context cues; instead we study how, combined with theory of mind, facial reactions can provide information about beliefs and desires—the abstract causal factors underlying action, emotion, and facial expressions.

Computational model

We take a Bayesian approach to characterizing the structure of the knowledge relating emotional reactions to classical theory of mind representations (beliefs, desires and actions). Our approach is inspired by research describing aspects of social reasoning as Bayesian inference over generative models of the ways in which mental states cause behavior (Baker, Saxe, & Tenenbaum, 2009; 2011). Fig. 1 expresses our model as a Bayesian network, which specifies the structure of the causal processes by which beliefs and desires influence emotional reactions. We focus on scenarios in which an emotionally charged event occurs that is the *Outcome* either of the agent's *Action* or some external cause. The agent's reactions to the *Outcome*, and possibly also the *Action*, are observed. The forward blue arrows to the *Reaction*$_0$ node and the forward red arrows to *Reaction*$_1$ node capture their causal dependence on the *Belief*, *Desire*, *Action*, and *Outcome* nodes. *Reaction*$_0$ depends on the agent's evaluation of the expected outcome (prior to observing the actual outcome), while *Reaction*$_1$ depends on the agent's evaluation of the observed *Outcome* once it has occurred. The model also specifies how beliefs, desires, and actions are generated according to the familiar theory of mind schema in which *Belief* and *Desire* cause *Action*, and *Belief* and *Desire* themselves are generated from a context-specific prior.

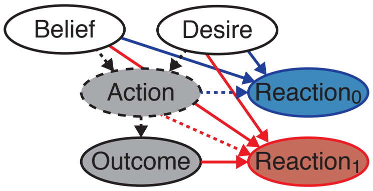


Figure 1 Graphical model illustrating the relationship between theory of mind and emotional reactions. Based on different substructures of the model, we modeled people's backward inferences from emotional reactions to belief and desire, varying whether only *Reaction*$_1$ is observed (red arrows; Exps. 1,3) or both *Reaction*$_0$ and *Reaction*$_1$ are observed (blue&red arrows; Exps. 2,4), and whether the agent acts to cause the outcome (including dotted arrows; Exps. 1,2) or merely observes it (excluding dotted arrows; Exps. 3,4).

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. Given these distributions, the model predicts that backward inferences about *Belief* and *Desire*, given observable information (e.g., *Action*, *Outcome*, and *Reactions*) decompose into a product of forward causal dependencies via Bayes' rule:

$$\begin{aligned} &P(\mathit{Belief}, \mathit{Desire} \mid \mathit{Action}, \mathit{Outcome}, \mathit{Reaction}_0, \mathit{Reaction}_1) \propto \\ &\quad P(\mathit{Reaction}_1 \mid \mathit{Belief}, \mathit{Desire}, \mathit{Action}, \mathit{Outcome}) \times \\ &\quad P(\mathit{Reaction}_0 \mid \mathit{Belief}, \mathit{Desire}, \mathit{Action}) \times \qquad (1) \\ &\quad P(\mathit{Action} \mid \mathit{Belief}, \mathit{Desire}) \times P(\mathit{Belief}, \mathit{Desire}). \end{aligned}$$

To determine whether people's generative, causal knowledge supports backward belief and desire inferences as predicted by our model, across several experiments, we elicit people's forward judgments about each component of the right-hand side of the equation, including the conditional probabilities $P(R_1|B,D,A,O)$, $P(R_0|B,D,A)$, and $P(A|B,D)$, and the prior $P(B,D)$ (abbreviating each variable by its first letter). We then compare people's backward inferences about belief and desire given observable information with the Bayesian model predictions, computed from the normalized product of the judged forward distributions, according to Eq. 1.

We tested this model with four behavioral experiments varying the context and the amount of information available to participants. In Exps. 1 and 2, people observed the agent perform an *Action* and generate a facial expression based on the *Outcome* (*Reaction*$_1$; see Fig. 1 and Eq. 1). Exp. 2 added an additional observation of the agent's reaction prior to observing the *Outcome* but after acting (*Reaction*$_0$), in order to test whether additional facial information would produce stronger inferences. In Exps. 3 and 4, the *Outcome* occurred due to an external cause, and no *Action* was performed by the agent. In these cases, only the agent's reactions were informative about her mental states, and we hypothesized that people's inferences would reveal more fine-grained facial processing. As before, we varied whether only *Reaction*$_1$ (Exp. 3), or both *Reaction*$_1$ and *Reaction*$_0$ (a reaction to initial news of the possible outcome, prior to observing it) were observed (Exp. 4). Our Bayesian model can account for these manipulations across Exps. 1-4 simply by removing terms from the product in Eq. (1) corresponding to any variable not present in a given scenario: when *Reaction*$_0$ is not observed (Exps. 1,3), $P(R_0|B,D,A)$ drops out of the product; when the agent does not act to cause the outcome (Exps. 3,4), $P(A|B,D)$ drops out.

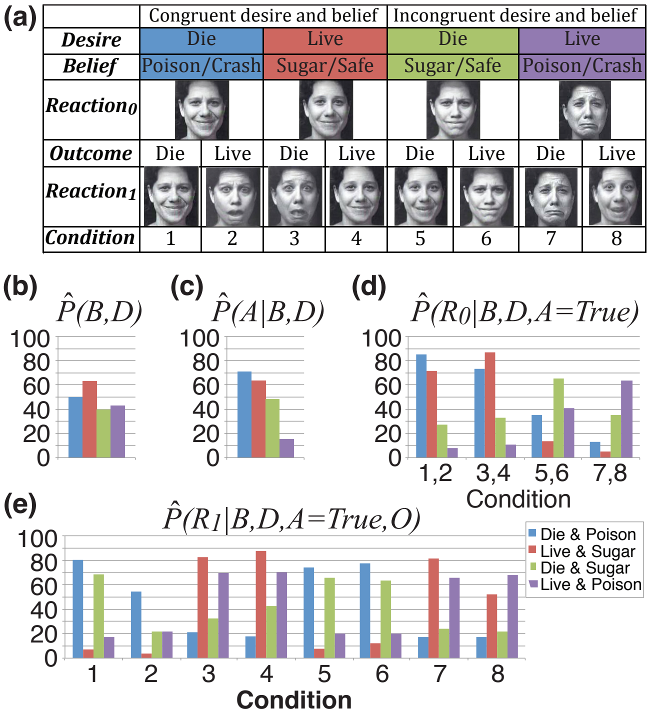
Experiment 1

Methods

Scenario We presented a scenario (adapted from 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 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 either believes the powder is poison or sugar.

Design and stimuli There are eight possible combinations of *Belief*, *Desire*, and *Outcome*, represented by *Conditions* 1-8 in Fig. 2(a). For each condition, we generated emotional reactions in two different ways. First, we used the facial morphing software *Fantamorph 5.4.0* to create a set of potential facial expressions. We manipulated two attributes in our morphed pictures (see Fig. 2(a): *Reaction*$_1$), based on the assumption that if the outcome was consistent with Grace's desire, her expression should be positive (and if inconsistent, negative), and that if the outcome was inconsistent with Grace's belief, her expression should be surprised (and if consistent, there should be no surprise). Second, to ensure that any effects we might find were not due to arbitrary features of the stimuli, we generated a separate set of stimuli by asking a professional actor, blind to the experimental motivation, to produce his own facial reactions given *Belief, Desire, Action* and *Outcome* in each condition. Eight short movie clips were filmed (http://web.mit.edu/yangwu/www/EmoToM).

Fig. 2(a) categorizes the conditions into two groups: "congruent" and "incongruent". In conditions 1-4, Grace's action of putting powder into John's coffee is naively congruent with the desires and beliefs used to generate her facial reaction (i.e., expected to achieve her desired outcome, according to her belief). In Conditions 5-8, the observed action is incongruent [1] with the desires and beliefs underlying the facial reaction; thus the observed action and facial reaction provide conflicting evidence about Grace's mental states.



(a)	Congruent desire and belief				Incongruent desire and belief			
Desire	Die		Live		Die		Live	
Belief	Poison/Crash		Sugar/Safe		Sugar/Safe		Poison/Crash	
Reaction₀								
Outcome	Die	Live	Die	Live	Die	Live	Die	Live
Reaction₁								
Condition	1	2	3	4	5	6	7	8

Figure 2 (a) Design of facial reaction stimuli for Exps. 1-4. The *Beliefs* Poison&Sugar refer to the chemical-factory scenario used in Exps. 1,2 while Crash&Safe refer to the plane-crash scenario used in Exps. 3,4; *Reaction*$_0$ was only used in Exps. 2,4. (b)-(e) Typical pattern of people's forward judgments (on an un-normalized 0-100 scale) about *Prior*, and *Action*, *Reaction*$_0$ and *Reaction*$_1$ *likelihoods* (results are shown for the chemical-factory scenario with the picture stimuli).

Participants and procedure All participants in this and following experiments were recruited on Amazon Mechanical Turk, with a drop rate of 13.1% (due to answering the scenario comprehension questions incorrectly or answering less than 50% of the test questions). The numbers of participants reported are those included in the final analyses.

Firstly we measured the prior over mental states given the scenario, $P(B,D)$, and the likelihood of Grace's action given each mental state, $P(A|B,D)$. Fifty-seven participants rated the prior plausibility of each combination of desire and belief: (1) Grace wants John to die and believes the powder is poison (in short, Die&Poison), (2) Grace wants John to live and believes the powder is sugar (Live&Sugar), (3) Grace wants John to die and believes the powder is sugar (Die&Sugar), (4) Grace wants John to live and believes the powder is poison (Live&Poison). The same participants also rated the likelihood of Grace's action given each of the four possible mental states, $P(A|B,D)$. All these and following judgments 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.

[1] Pilot work suggested that people were able to reason about mental states that were incongruent with observed actions and did imagine narratives outside of the scenario (e.g., for Die&Sugar, perhaps Grace was envisioning some other method of homicide; for Live&Poison, perhaps she was acting under coercion).

Next, we determined the likelihood of each reaction given *Belief*, *Desire*, *Action*, and *Outcome*, $P(R_1|B,D,A,O)$. One hundred and six participants provided forward predictions in each condition, judging the plausibility that the given desire, belief, action and outcome caused each of all the emotional reactions. Half of the participants rated the picture stimuli (n=55); the other half rated the movie stimuli (n=51).

Lastly, we tested people's backward inferences of *Belief* and *Desire* given *Action*, *Outcome* and *Reaction*$_1$, $P(B,D|A,O,R_1)$. One hundred and one participants were asked to judge the plausibility of the four mental states given Grace's action, the outcome, and *Reaction*$_1$ in each of the 8 conditions. Half of the participants were tested with the picture stimuli (n=49); the other half were tested with the movie stimuli (n=52). These judgments were also collected on a 0-100 scale but normalized to sum to 1 over all four possible belief-desire combinations, for comparison with model posterior probabilities.

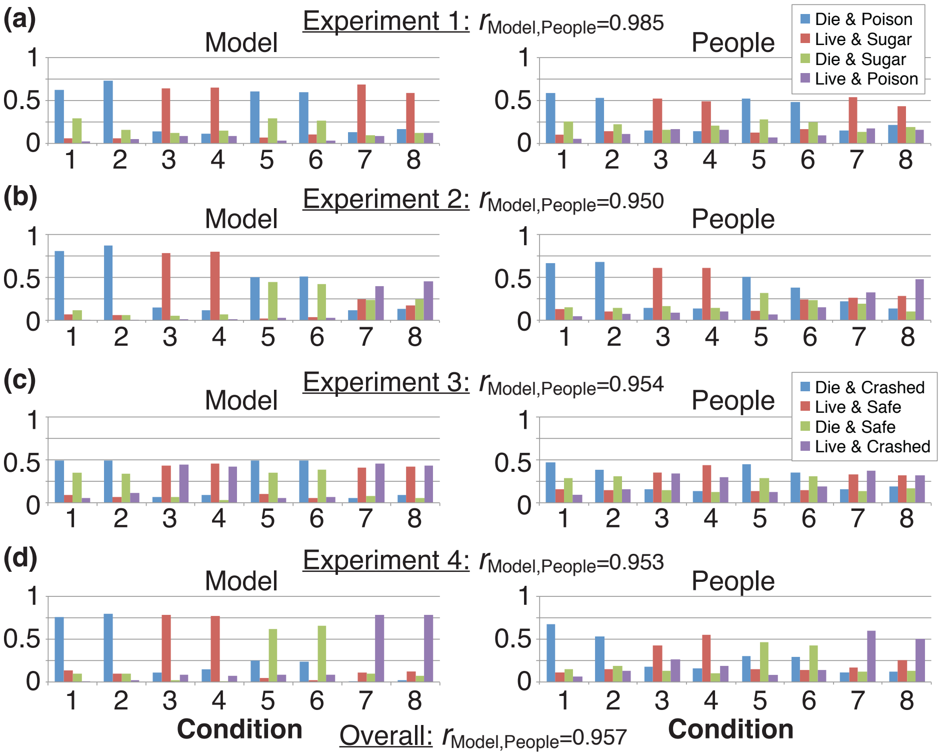


Figure 3 Comparison of model predictions and human backward inferences in Exps. 1-4.

Results and discussion

People's prior over mental states given the scenario was relatively uniform (Fig. 2(b)), indicating that the task instructions led them to consider all possible mental states. The action likelihood was rated higher for congruent mental states than for incongruent mental states (Fig. 2(c)).

For the picture stimuli, Fig. 2(e) arranges participants' conditional likelihood ratings for each value of *Reaction*$_1$ as a function of *Desire* and *Belief*, given the *Outcome* from the corresponding condition. In each condition, the likelihood of the emotional reaction was rated the highest for the desire from which the reaction was generated. However, in all but one condition (*Condition* 2), the two beliefs received roughly equal likelihood. For example, people judged that the emotional reaction in *Condition* 1 was as likely to have been produced by Die&Poison as Die&Sugar, suggesting that *Reaction*$_1$ was informative about desires, but not beliefs.

Fig. 3(a) shows model predictions of people's backward judgments for the picture stimuli, generated according to Eq. 1 (omitting the *Reaction*$_0$ term), using the forward distributions measured as described above. The model infers the desire underlying the reaction due to the *Reaction*$_1$ likelihood function. However, the model strongly predicts that people's belief inferences will be those most congruent with the desire inferences—for example Poison in *Conditions* 1,2,5,6, and Sugar in *Conditions* 3,4,7,8. These predictions result from conditioning on the observed *Action*; the conditional action likelihood favors Die&Poison or Live&Sugar, and this biases the backward posterior inferences toward congruent mental states.

People's backward inferences tested with the picture stimuli are reported in Fig. 3(a). They correlated strongly with the model predictions (r=0.985), consistent with Bayesian inference over structured causal knowledge, as measured in the forward tasks.

We performed the same analysis on the data from the movie stimuli as on the picture stimuli data. For the sake of brevity we will not present those results here, because they replicated those from picture stimuli in all respects; the correlation between model predictions and participants' judgments for these stimuli was r=0.908.

Critically however, for both the pictures and the movies, participants saw the agent's emotional reaction at a single time point: once Grace knows whether John lives or dies after drinking the coffee. In Exp. 2, we look at whether people's belief inferences are less biased by the action likelihood if additional emotional reactions are observed.

Experiment 2

Methods

Scenario Same as Exp. 1.

Design and stimuli We modified Exp. 1 by adding observations of Grace's facial reactions before observing the outcome (*Reaction*$_0$), based on the expected outcome according to her belief. We assume that *Reaction*$_0$ and *Reaction*$_1$ will be similar when the expected and actual outcomes match, but when Grace has a false belief (i.e., there is a mismatch between actual and expected outcomes), the valence of *Reaction*$_0$ and *Reaction*$_1$ will be different. For simplicity, for each pair of conditions sharing the same mental state (and expected outcome), we select *Reaction*$_0$ by reusing *Reaction*$_1$ from the condition where the expected and actual outcomes match (e.g., *Reaction*$_0$ in *Conditions* 1 and 2 reuses *Reaction*$_1$ from *Condition* 1; see Fig. 2(a)). We used only the picture stimuli in this and following experiments because they produced the same results as the movie stimuli in Exp. 1 and they are simple to manipulate.

Participants and procedure Fifty-eight participants rated the likelihood of *Reaction*$_0$, $P(R_0|B,D,A)$; 53 participants made backward inferences about the probability of each combination of Grace's *Belief* and *Desire* given *Action*, *Outcome*, *Reaction*$_0$ and *Reaction*$_1$, $P(B,D|A,O,R_0,R_1)$.

Results and discussion

The likelihoods of *Reaction*$_0$ are reported in Fig. 2(d). The most obvious result is that the two positively valenced reactions (those used in *Conditions* 1,2 and 3,4) were rated higher given congruent than incongruent mental states. The two negatively valenced reactions (those used in *Condition* 5,6 and 7,8) showed the opposite pattern. Since the congruency of the mental states determines whether the action would satisfy the desire according to the agent's belief, these results suggest that participants judged *Reaction*$_0$ likelihood based on the match between the valence and the expected satisfaction of desires.

Model predictions of people's backward inferences of belief and desire $P(B,D|A,O,R_0,R_1)$ were generated according to Eq. 1 (see Fig. 3(b)), predicting reliable inference of the desires underlying the reactions. The belief inferences predicted by the model were no longer dominated by the action likelihood, and were less certain (*Conditions* 5,6) or even flipped (*Conditions* 7,8) compared with Exp. 1 in the conditions where the valence contrast of the two emotional reactions supported a different belief than that favored by the likelihood of the observed action.

People's backward inferences are reported in Fig. 3(b). Participants' responses correlated highly with our model predictions (r=0.950).

In Exps. 1 and 2, backward inferences were lower for incongruent mental states (Die&Sugar and Live&Poison), due to the action likelihood. In Exps. 3 and 4 we remove any effect of the action likelihood by making Grace only an observer so that more fine-grained reasoning based on emotional reactions could be revealed. In Exp. 3 we provided participants with reactions at a single time point (*Reaction*$_1$) while in Exp. 4 we provided reactions at two time points (*Reaction*$_0$ and *Reaction*$_1$).

Experiment 3

Methods

Scenario We presented a scenario similar to the one used on the previous experiments, but instead of taking a tour in a chemical factory, Grace is watching TV and learns that a plane has crashed on a route often flown by her coworker John. Grace either wants John to die or live, and either believes John is on the crashed plane or a safe plane.

Design and stimuli Same as Exp. 1.

Participants and procedure Given the new scenario, 57 participants judged the prior over mental states $P(B,D)$, and 46 rated the likelihood of *Reaction*$_1$ given *Belief*, *Desire* and *Outcome*, $P(R_1|B,D,O)$.

Results and discussion

The elicited prior and *Reaction*$_1$ likelihood were similar to those from Exp. 1 (r=0.926). Model predictions of people's backward inferences of belief and desire, $P(B,D|O,R_1)$ were generated according to Eq. 1 (omitting the *Reaction*$_0$ and *Action* term). The model predicted that the desires would be consistently inferred due to the *Reaction*$_1$ likelihood function. However, the model predicted that people's belief inferences, without the action likelihood biasing them toward congruent mental states, would assign equal probability to the two possible beliefs (see Fig. 3(c)). People's backward inferences of belief and desire correlated highly with the model predictions (r=0.954, Fig. 3(c)).

These results suggest that when no action was performed and only reactions to actual outcomes were observed, people could recover the underlying desires, but were uncertain about the beliefs, as predicted by our Bayesian model.

Experiment 4

Methods

Scenario Same as Exp. 3.

Design and stimuli Same as Exp. 2.

Participants and procedure Fifty participants rated the likelihood of *Reaction*$_0$, $P(R_0|B,D)$, in the new context; 57 participants made backward inferences about the probability of each combination of Grace's *Belief* and *Desire* given *Outcome*, *Reaction*$_0$ and *Reaction*$_1$, $P(B,D|O,R_0,R_1)$.

Results and discussion

The likelihood ratings of *Reaction*$_0$ paralleled those in Exp. 2 (r=0.950). Model predictions of people's backward inferences were generated according to Eq. 1 (omitting the *Action* term). The predicted posterior probability of belief and desire, $P(B,D|O,R_0,R_1)$ suggested that in all conditions participants would not only reason backward about desires but also the beliefs from which the reactions were generated (see Fig. 3(d)). People's backward inferences qualitatively confirmed these predictions, and correlated highly with the model predictions (r=0.953).

As evident in Fig. 3(d), when participants were given emotional reactions over two key time points and there was no bias due to the action likelihood, people were able to infer each unique combination of beliefs and desires from the emotional reactions and the context. These responses were well-predicted by the model.

General discussion

We proposed a Bayesian framework for modeling people's joint inference of belief and desire from emotional reactions. To test our model, we measured people's forward judgments about the prior over mental states, the likelihood of performing an action, and the likelihood of emotional reactions; we then fed the forward data into our model, which accurately predicted people's backward inferences across multiple experiments and scenarios.

Across four experiments, our model predicted different patterns of backward inferences (see Fig. 3). In Exp. 1, the model predicted that people could infer agents' desires based on their observed emotional reactions to actual outcomes, and beliefs based on the observed action. In Exp. 2, given additional reactions at a different time point (after acting but before knowing the outcome), the model predicted that people's belief inferences would be less biased by the action when contrasting reactions suggested an alternative belief. In Exps. 3 and 4, when no action was performed by the agent, the model predicted that desires could be inferred but that beliefs could only be inferred when reactions at two time points were available. These model predictions closely captured people's backward inferences across the four experiments (r=0.957).

Our study also probes people's naïve understanding of the relationship between mental states and facial expressions. Our original hypotheses were that the valence dimension of facial expressions could reveal the state (satisfied or unsatisfied) of desires, and the surprise dimension could reveal the veracity (false or true) of initial beliefs. However, our results support the former but not the latter hypothesis. A possible explanation could be that since surprise plays a role in intensifying valence (e.g. if a desirable event is unexpected, the surprise magnifies the felt happiness; Ortony, 1990), the combination of surprise and valence is perceptually obscured with intensely valenced emotions. Thus, people do not take the perceived surprise as a reliable cue for an initial false belief. Additionally, surprise may often be fleeting, hard to catch and easy to hide, perhaps explaining why people do not infer a true belief from the absence of a surprised facial expression. Further research is needed to advance our understanding of the relationship between expressions of surprise and attributions of belief.

Our study does suggest that observing facial expressions over multiple time points can be informative about agent's belief. The absence of a valence change in the facial expression between the expected and the actual outcome suggested a true belief, and the presence of a valence change, a false belief.

At least in adults, our naïve theory of emotional reactions appears to be structurally and causally intertwined with theory of mind in a way that allows forward prediction from an agent's beliefs and desires to her facial expressions, and backward inference from facial expressions to beliefs and desires. In future research we hope to investigate the ways in which our ability to infer mental states from emotional expressions in childhood changes over development.

Although our present model captures the structure of the causal relationship between beliefs, desires, and emotional reactions, the functional form is represented only implicitly in the forward predictions elicited within our experiments. In ongoing research, we are modeling people's knowledge of how emotions arise from beliefs, desires, actions, and outcomes, and how facial reactions express these emotional states – intuitive versions of classical problems studied by psychologists. As a first step, this account accords well with the scientific appraisal theory of emotions, suggesting that the appraisal process is shared by both the scientific study of emotions and people's intuitive theories.

Acknowledgments

This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

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

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-second annual conference of the cognitive science society* (pp. 2469-2474).

Calder, A. J., Keane, J., Manly, T., Sprengelmeyer, R., Scott, S., Nimmo-Smith, I., & Young, A. W. (2003). Facial expression recognition across the adult life span. *Neuropsychologia*, *41*(2), 195-202.

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.

Csibra, G., Bıró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111-133.

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

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of affective sciences*, *572*, V595.

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165-193.

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

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.

Ortony, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Richardson, H. L., Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2012). The Development of Joint Belief-Desire Inferences. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 923-928).

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences*, *9*(12), 585-594.

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.