

Emotion as information: Inferring the unobserved causes of others' emotional expressions

by

Yang Wu

B.S., Peking University (2012)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© 2018 Massachusetts Institute of Technology. All rights reserved.

Signature redacted

Signature of Author.....

Department of Brain and Cognitive Sciences

May 4, 2018

Signature redacted

Certified by

Laura E. Schulz

Professor of Cognitive Science

Thesis Supervisor

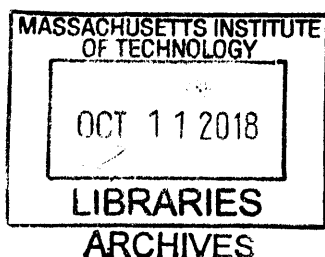
Signature redacted

Accepted by.....

Matthew A. Wilson

Sherman Fairchild Professor of Neuroscience and Picower Scholar

Director of Graduate Education for Brain and Cognitive Sciences



Emotion as information: Inferring the unobserved causes of others' emotional expressions

by

Yang Wu

Submitted to the Department of Brain and Cognitive Sciences on May 4, 2018
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in
Cognitive Science

ABSTRACT

Research in the domain of cognitive science has tended to neglect emotions. In my thesis, I take several steps to fill this gap by looking at people's representation of emotions, and its connection to other representations typically studied in cognitive science. I argue that people have an intuitive theory of emotion that is causally intertwined with their understanding of the physical and social world broadly. This intuitive theory allows us to use observed emotional cues as a window, to recover unobserved information about the world. I study these abilities in both adults and children, to gain insight into the most fundamental representations supporting such abilities. I also use computational models to capture the hierarchical, causal structure of this intuitive theory of emotion.

In Study 1, I show that infants as young as 12-17 months can discriminate diverse within-valence emotional expressions elicited by funny, exciting, adorable, delicious, and sympathetic events, and map them onto their probable causes. In Study 2.1, I present that preschoolers can recover rich mental state information from observed emotional expressions. When the valence of someone's face changes between anticipated and actual outcomes, children by five gain insight into what she wants and believes about the world. Study 2.2 bridges theory of mind research, accounts of emotion attribution, and formal modeling, to provide a formal account of how people jointly infer beliefs and desires from emotional expressions. Study 3 tests children's understanding of social display rules. By middle childhood, children can use one person's emotional expressions regulated by a social context to infer the mental states of another.

Altogether, these findings suggest that emotional cues provide a valuable entrée into the unseen world. Not only adults, but also children, can use observed emotional expressions to infer their external causes and the internal mental states of other people. Although this intuitive theory of emotion may not necessarily mirror the actual processes of how emotions are generated, it supports rational inferences much of time, and it may be formed early in development. I see this work as bridging gaps across disciplines and helping advance the cognitive science of emotion understanding.

Thesis Supervisor: Laura E. Schulz

Title: Professor of Cognitive Science

Acknowledgements

I thank my advisors Dr. Laura Schulz and Dr. Josh Tenenbaum. As anyone who knows them may notice, neither of them had ever worked on emotion and neither of them would necessarily identify themselves as emotion researchers. But what is cool and amazing about them is that they respect students' interests, they support students to develop their own research programs, and they are capable of doing that more than you could imagine. I still remember that when I first came to grad school, I explored my research interests and I got interested in emotion. Although it was a little crazy for me to tell my advisors, two great cognitive scientists, that I wanted to work on emotion, it was (probably) crazier for them to have said yes, organized a reading group about emotion, and spent time learning everything together with me...

Along the way, we have been having very ambitious goals. Sometimes things worked; sometimes things did not work. Sometimes we knew how to proceed; sometimes we did not know. Our emotional responses to these have been rich and complex: excited, worried, perplexed, delighted, thrilled... And no simple emotion word can capture my feelings now, when I am writing this page, but I sincerely thank Laura and Josh for taking the journey with me—it has been unforgettable and fun.

I also thank my committee members Dr. Rebecca Saxe and Dr. Liz Spelke. They are great scientists and they have been giving me the most helpful feedback about my work, which has deeply influenced and shaped this research program.

Thanks to my collaborators Dr. Paul Muentener and Dr. Chris Baker. Paul was a postdoc in Laura's lab when I began grad school, who is now an assistant professor at Tufts University. He taught me some of the biggest and smallest things for successfully conducting developmental research, including how to design a baby experiment, how to critically think about developmental data, and... how to get a baby's attention! Chris was a postdoc in Josh's lab, who is now a co-founder and chief scientist at iSee AI. He collaborated with me on my first modeling project and taught me patiently a great deal about how to build a model.

I thank former and current members of the Early Childhood Cognition Lab and the Computational Cognitive Science Group. They have given me lots of support and fun. I thank my parents, who have given me unconditioned love, and my husband, Jun Li, who loves me as I am.

Last but not least, thanks to my funding support, including Leventhal, Stark, and Henry E. Singleton Fellowships, and the Center for Brains, Minds and Machines (funded by NSF STC Award CCF-1231216).

Table of Contents

Chapter 1 General Introduction.....	8
1.1 INTRODUCTION	8
1.2 PREVIOUS LITERATURE	9
1.2.1 Scientific theories of emotion	9
1.2.2 Emotion recognition.....	12
1.2.3 People’s ability to infer and predict emotions	14
1.3 THE PRESENT RESEARCH: INFERRING CAUSES OF EMOTIONS	17
Chapter 2 Study 1 Inferring External Causes of Emotional Expressions	19
2.1 ABSTRACT.....	19
2.2 SIGNIFICANCE.....	20
2.3 INTRODUCTION	21
2.4 EXPERIMENTS 1-3.....	24
2.4.1 Eliciting cause stimuli and emotional vocalizations.....	25
2.4.2 Experiment 1: 2 to 4-year-olds and adults	26
2.4.3 Experiments 2-3: 12-23-month-olds	27
2.5 EXPERIMENTS 4-5: 12-17-MONTH-OLDS	29
2.6 GENERAL DISCUSSION	32
2.7 MATERIALS AND METHODS.....	35
Chapter 3 Study 2.1 Inferring Beliefs and Desires From Emotional Expressions.....	40
3.1 ABSTRACT.....	40
3.2 INTRODUCTION	41
3.3 EXPERIMENT 1	46
3.3.1 Method	46
3.3.2 Results and discussion	50
3.4 EXPERIMENT 2: REPLICATION.....	52
3.4.1 Method	52
3.4.2 Results and discussion	53
3.5 EXPERIMENT 3	56
3.5.1 Method	57
3.5.2 Results and discussion	58
3.6 GENERAL DISCUSSION	59
Chapter 4 Study 2.2 Inferring Beliefs and Desires From Emotional Expressions: A	
Computational Model.....	66
4.1 ABSTRACT.....	66
4.2 INTRODUCTION	67
4.3 COMPUTATIONAL MODEL	74
4.4 BEHAVIORAL EXPERIMENTS	79
4.4.1 Experiment 1	79
4.4.2 Experiment 2a	88
4.4.3 Experiment 2b.....	90

4.4.4 Experiment 3	97
4.4.5 Experiment 4	102
4.5 COMPARISON WITH OTHER MODELS	106
4.5.1 No-Emotion Model	107
4.5.2 No-Action Model	107
4.5.3 Event-Features Model	109
4.6 GENERAL DISCUSSION	110
Chapter 5 Inferring Recursive Mental States from Emotional Expressions.....	115
5.1 ABSTRACT.....	115
5.2 INTRODUCTION	116
5.3 EXPERIMENT 1	120
5.3.1 Method	120
5.3.2 Results and discussion	124
5.4 EXPERIMENT 2	126
5.4.1 Method	126
5.4.2 Results and discussion	126
5.5 EXPERIMENT 3	127
5.5.1 Method	127
5.5.2 Results and discussion	128
5.6 EXPERIMENT 4	129
5.6.1 Method	129
5.6.2 Results and discussion	130
5.7 EXPERIMENT 5	131
5.7.1 Method	131
5.7.2 Results and discussion	132
5.8 GENERAL DISCUSSION	132
Chapter 6 General Conclusions.....	136
6.1 SUMMARY AND FUTURE DIRECTIONS.....	136
6.1.1 Study 1 Inferring External Causes of Emotional Expressions.....	136
6.1.2 Study 2.1: Inferring Beliefs and Desires From Emotional Expressions	138
6.1.3 Study 2.2: Inferring Beliefs and Desires From Emotional Expressions: A Computational Model	139
6.1.4 Study 3: Inferring Recursive Mental States From Emotional Expressions	140
6.2 BROADER FUTURE DIRECTIONS	142
Appendix I. Study 1 Supporting Information Appendix	146
Appendix II. Study 2.2 Supporting Information	164
References.....	188

Chapter 1 General Introduction



1.1 INTRODUCTION

If you look at the images above, you can recognize many aspects of each scene (e.g., people, hands, sofas, and shirts). Strikingly however, you may find that you can also recover information that is *not in the scene* at all: a sports event in Scene 1, a horror movie in Scene 2, and a baby in Scene 3. You may also get a sense of the *underlying mental states* of these people. Most of the sports fans in Scene 1, for instance, seem to be expecting a desirable outcome with different levels of confidence, while a guy in the back (wearing a maroon shirt) seems to be favoring a different outcome.

What are the representations that support these inferences? While most research in the domain of cognitive science has tended to neglect emotions, in my thesis, I take several steps to fill this gap by looking at people's representation of emotions, and its connection to some other representational structures that are typically studied in cognitive science. Specifically, I argue that people have an intuitive theory of emotion that is causally and structurally intertwined with their understanding of the physical and social world broadly. This intuitive theory allows them to

use observed emotional cues as information, to recover unobserved information about the world that is otherwise underdetermined.

I provide evidence in three studies, as reported in the following chapters. These studies span across ages but my main focus is on infancy and childhood, in order to gain insight into the most fundamental concepts that are shared by even young children. My work also uses methods across disciplines. I have used developmental paradigms including looking time and manual search tasks to investigate the minds of infants and children. I have also built computational models to capture the hierarchical structure of people's conceptual knowledge of emotion, and validated the quantitative predictions of these models using behavioral experiments with human adults.

Such research is distinct from a large body of prior work on emotion. In this introductory chapter, I will first review several lines of work that relate to my research, emphasizing on the connections and differences between previous research and my work. I will then give an overview of the research reported in this thesis.

1.2 PREVIOUS LITERATURE

1.2.1 Scientific theories of emotion

A premise of my work is that humans possess intuitive knowledge of emotions. Similar to other domains of high-level cognition, this knowledge (at least in its mature form as possessed by human adults) is theory-like, which can be used to draw rich, abstract and flexible inferences about the world. The term, *the intuitive theory of emotion*, has been proposed recently to describe such knowledge, and has become a topic of interest in many disciplines including developmental psychology, neuroscience, and computational cognitive science (Ong, Zaki, & Goodman, 2015; Saxe & Houlihan, 2017; Skerry, 2015; Skerry & Saxe, 2015; Skerry & Spelke, 2014; Wu, Baker,

Tenenbaum, Schulz, 2014). However, it is often confused with some *scientific* theories of emotion, and a lot of prior work on emotion did not draw a clear distinction between the two. Here I will review some related scientific theories of emotion and then discuss the connections and differences between intuitive and scientific theories of emotion.

Historically, emotion researchers have proposed a number of scientific theories of emotion to characterize what emotions are and how they are elicited and differentiated within individuals (e.g., Arnold, 1960; Barrett, 2006a; Ekman, 1992; Frijda, 1986; James, 1890; Lazarus, 1966, 1991; Leventhal, 1980, 1984; Lyons, 1980; Nussbaum, 1990; Oatley & Johnson-Laird, 1987; Ortony, Clore, & Collins, 1988; Roseman, Antoniou, & Jose, 1996; Russell, 2003; Schachter, 1964; Scherer, 1984; Smith & Ellsworth, 1985; Solomon, 1976). This set of theories are particularly relevant in the context of this thesis but they, at least in their original forms, attempt to describe the nature of emotions and how they are generated, rather than how *laypeople* think about these questions. An analogy here is that the scientific theories of physics attempt to use mathematical models as well as abstractions of physical objects and systems to explain and predict natural phenomena. Laypeople also have a “theory” of how those natural phenomena work but this intuitive theory differs, in many ways, from scientific theories of physics.

Scientific theories of emotion have proposed some components that are involved in the emotion elicitation and differentiation processes, including eliciting stimuli, physiological responses, cognitive appraisals, subjective experiences, and action tendency. However, different theories vary both in the types of components involved and the order of these components to occur in an emotional episode. For example, an early theory (e.g., James, 1890) does not have the component of cognitive appraisals and postulates one-to-one mappings between physiological responses and subjective experiences (i.e., feelings). However, other theories,

appraisal theories in particular (e.g., Arnold, 1960; Frijda, 1986; Lazarus, 1966, 1991; Oatley & Johnson-Laird, 1987; Ortony, Clore, & Collins, 1988; Roseman, Antoniou, & Jose, 1996; Scherer, 1984; Smith & Ellsworth, 1985; see Moors, Ellsworth, Scherer, & Frijda, 2013 for review), place cognitive appraisals at the central of the emotion elicitation and differentiation processes. Some theories posit that physiological responses precede cognitive appraisals (e.g., we tremble and then we try to interpret it; Schachter, 1964) but appraisal theories place cognitive appraisals at the very beginning of an emotional episode, prior to physiological responses (e.g., we interpret a situation as frightening and then we tremble). Different theories also differ in the extent to which they think these processes are automatic (i.e., uncontrolled, unconscious, and efficient). Some theories (Schachter, 1964; Lyons, 1980; Nussbaum, 1990; Solomon, 1976) assume that they are not automatic while others (Arnold, 1960; Barrett, 2006a; Ekman, 1992; Frijda, 1986; Lazarus, 1966, 1991; Oatley & Johnson-Laird, 1987; Ortony, Clore, & Collins, 1988; Roseman, Antoniou, & Jose, 1996; Russell, 2003; Scherer, 1984; Smith & Ellsworth, 1985) suggest that they can be automatic.

Although some of these scientific theories tap on the extent to which they can be used to characterize laypeople's intuitive theory of emotion, the main focus of these theories has been on the nature of emotions and their elicitation and differentiation processes. Here, my focus is on people's intuitive theory of emotion instead. I argue that laypeople, including both adults and children, have an intuitive understanding of emotions and how they are generated. Although this intuitive theory may not necessarily mirror the nature of emotions, like intuitive theories in many other domains (e.g., physics, biology, and theory of mind), it may govern people's ability to draw fast and accurate inferences about emotions much of time in their daily lives. Importantly, because people rely on their intuitive theory of emotion to infer, predict, and intervene in others'

emotions (Denham et al., 2003; Frederickson, Petrides, & Simmonds, 2012; Skerry & Saxe, 2015; Vaish, Carpenter, & Tomasello, 2009) and atypical development of such conceptual knowledge relates to social difficulties or mental disorders (e.g. Adolphs, Sears, & Piven, 2001; Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Harms, Martin, & Wallace, 2010; Hobson, 1986; Amminger et al., 2011; Gold et al., 2012; de Wied, Gispen-de Wied, & van Boxtel, 2010), it is crucial to understand the form and content of this intuitive theory of emotion, and its development over infancy and childhood. These are the goals of my thesis.

1.2.2 Emotion recognition

Historically, there has also been a large body of work looking at how people recognize what others are feeling based on their emotional expressions. This line of work has primarily been driven by the long-standing debate about the structure and universality of emotions (Barrett, 2006b; Ekman, 1992; Izard, 1971; Panksepp, 1992; Russell, 1980; Russell & Bullock, 1986a, 1986b; Tomkins, 1962; Watson, Wiese, Vaidya, & Tellegen, 1999). According to the *basic* emotion theory, the human mind has been endowed with a small set of five to seven discrete emotions that we have an innate, core understanding of (Tomkins, 1962; Panksepp, 1992; Ekman, 1992; Izard, 1971). Each of these basic emotions is universally expressed and recognized, and has its own evolutionary purpose and physiological basis. More complex emotions arise from combinations of these basic emotions or are culturally influenced and constructed (Du, Tao, & Martinez, 2014; Ekman & Cordaro, 2011). According to the *dimensional* emotion theory, however, emotions arise from more fundamental dimensions such as valence and arousal (Barrett, 2006b; Russell, 1980; Russell & Bullock, 1986a, 1986b; Watson, Wiese, Vaidya, & Tellegen, 1999). The two dimensions are associated with distinct neural systems and are the building blocks of emotional life. On this view, discrete emotions are not

given in nature, but are social-cultural constructs. Much of the evidence supporting either theory comes from studies looking at people's ability to categorize or label emotional facial expressions (e.g., Carroll & Russell, 1996; Du, Tao, & Martinez, 2014; Ekman, 1992; Ekman & Oster, 1979; Elfenbein, Beaupre, Levesque, & Hess, 2007; Izard, 1971; Posamentier & Abdi, 2003; Russell, 1980; Russell & Bullock, 1986a, 1986b; Widen, 2014; Widen & Russell, 2008a, 2010). There is also an increasing number of studies using similar tasks but looking at emotional expressions in other modalities, including emotional vocalizations (Bachorowski & Owren, 2003; Sauter, Eisner, Ekman, & Scott, 2010; Scherer, 2003) and body posture (Atkinson, Dittrich, Gemmell, & Young, 2004; Aviezer, Trope, & Todorov, 2012; Dael, Mortillaro, & Scherer, 2012; de Gelder, 2006; de Gelder, de Borst, & Watson, 2015; Martinez, Falvello, Aviezer, & Todorov, 2016; Meeren, van Heijnsbergen, & Gelder, 2005). However, there is still no simple consensus between researchers supporting these theories.

My research also looks at people's understanding of emotional displays. However, instead of investigating the ability to identify the underlying emotional *categories* (e.g., happy, sad, or angry) of emotional expressions, I focus on people's ability to infer the probable *causes* (e.g., external causes such as snakes and internal causes such as beliefs and desires) of observed emotional cues. In my research, I take as a premise that at least within a well-specified context and shared cultural knowledge, people can probabilistically infer some emotional content from emotional displays. However, I am not particularly interested in the inference from emotional expressions to their categories or labels, but interested in the inference from emotional expressions to the causes that generate these expressions. Critically, as shown by some of my work (see Study 1), the underlying category or label of an emotional expression is not particularly important or necessary for inferring the causes of an emotional display; some

emotional expressions such as the positive emotional responses to adorable babies or to delicious food do not correspond to any simple English emotion word but people, including young children, can make such causal inferences (see Study 1).

1.2.3 People's ability to infer and predict emotions

Compared to research on emotion recognition, relatively fewer studies have looked at people's ability to infer and predict emotions from their eliciting conditions, and many of these studies come from the developmental literature. Such literature has investigated children's ability to infer and predict an agent's emotions from her goals/desires and beliefs. For example, using looking time measures, infants as young as ten months expect an agent to express negative emotions when she fails to achieve her goal rather than when she completes it (Skerry & Spelke, 2014); by twenty months, toddlers expect an agent to be surprised by an outcome if she previously had a false belief about it (Scott, 2017). By two, children can explicitly predict that an agent will be happy if her desires are fulfilled and sad if her desires are thwarted (e.g., Wellman & Wooley, 1990; Yuill, 1984). Between four and six, children become increasingly likely to explicitly predict that someone will be surprised if her beliefs are violated, and someone will feel happy if she falsely believes that her desires are fulfilled (Hadwin & Perner, 1991; Harris, Johnson, Hutton, Andrews, & Cooke, 1989; Wellman & Banerjee, 1991). Other studies suggest that preschoolers also know some "scripts" connecting familiar emotions and events (e.g., getting a puppy and happiness; dropping an ice cream cone and sadness; Russell, 1990; Saarni & Harris, 1991; Widen, Pochedly, & Russell, 2015; Widen & Russell, 2004; 2010).

In the adult literature, some studies have used self-report measures to test how adults would attribute their own emotions when they recalled their past emotional experiences (e.g., Scherer & Meuleman, 2013) or given hypothesized, imagery events (e.g., Smith & Lazarus,

1993). Those studies attempt to provide evidence for scientific theories of emotion (appraisal theories in particular): how real emotional experiences are generated by eliciting events. However, the self-report method has been considered as a flaw or weakness in those studies because it is controversial the degree to which participants were emotionally engaged while performing those tasks. In those studies, participants may have simply used their conceptual knowledge of emotion to make emotion attributions, and if this is the case, those studies shed light on people's intuitive theory of emotion rather than scientific theories of emotion. Other studies have tested adults' conceptual knowledge of emotion more directly (e.g., Fontaine, Scherer, Roesch & Ellsworth, 2007; Skerry & Saxe, 2015). They asked participants to predict what emotion someone else would feel given eliciting events (e.g., being called into the boss' office after learning that the company is planning massive layoffs; Skerry & Saxe, 2015), or directly rate the appraisal features (e.g., caused by chance, confirmed expectations, and treated unjustly) represented by a set of emotion words (e.g., love, jealousy, and irritation; Fontaine, Scherer, Roesch & Ellsworth, 2007). All these studies suggest that adults' emotion attribution given eliciting events can be captured or explained by a number of appraisal features of those events including whether they are desired, expected, familiar, fair and controllable (Fontaine, Poortinga, Setiadi & Markam, 2002; Fontaine, Scherer, Roesch & Ellsworth, 2007; Skerry & Saxe, 2015; Scherer & Meuleman, 2013).

My work differs from the above research in three ways. First and foremost, rather than testing children and adults' ability to infer and predict emotions from beliefs, desires, or eliciting events, I am interested in an "inverse inference" problem: how people use observed emotional expressions to reversely infer the unobserved causes of those emotional expressions. The inverse inference problem is important because in the real world, we do not usually know why other

people feel that way but we often get to see the emotional expressions on their faces or in their body language. This is also the reason why mindreading is a hard inference problem: in many situations the only information that is available is what can be directly observed in a scene, and both mental states and past histories have to be inferred given sparse observations. However, it is precisely in such cases emotional expressions, as an important source of observation, may provide a valuable entrée into the unseen world.

Second, similar to other high-level cognition, the inference from observed emotional cues to their causes may rely on probabilistic reasoning over a richly structured causal model of how emotions are generated but previous research has not had a relatively comprehensive generative model of that. In particular, the adult work reviewed above has largely taken appraisal features as a flat, unstructured representation, which ignores the temporal and causal structures of emotional events. Although this approach has been productive and can (to a certain degree) capture people's emotion attribution, it is unlikely that people's representation of emotional events has no temporal or causal structures at all. Other work has proposed a number of computational models of emotion (e.g., Adam, Herzig, & Longin, 2009; El-Nasr, Yen, & Ioerger, 2000; Gebhard, 2005; Gratch & Marsella, 2004; Marinier III, Laird, & Lewis, 2009; Ortony, Clore, & Collins, 1988; Steunebrink, Dastani, Meyer, 2012; see also Marsella, Gratch, & Petta, 2010 for review); however, these models attempt to formalize scientific theory of emotion rather than laypeople's intuitive theory of emotion, and most of them have not been empirically tested with human behavioral experiments. A recent study (Ong, Zaki, & Goodman, 2015) has modeled people's intuitive theory of emotion. This study looked at people's representation of the relationships between simple event outcomes (i.e., the outcomes of bets on a Roulette wheel) and emotions; however, as suggested by the developmental literature, even young children

understand that people's emotional responses to events do not simply rely on the outcomes of these events but also on their mental representations (e.g., beliefs and desires) of these outcomes. Thus in Study 2.2 of my thesis, I use a similar approach as Ong et al (2015) but build a generative model that integrates people's representations of someone's emotions not only with event outcomes but also with her beliefs, desires, and actions. I validate this model quantitatively with adult behavioral experiments.

Third, the appraisal features used in the studies reviewed above are either handpicked or derived from previous literature. Therefore, those features are neither optimized nor exhaustive. Research in this thesis begins to formalize some of the most fundamental appraisal features (i.e., beliefs and desires) with a computational model (Study 2.2). It also explores some additional appraisal features that are missing in prior work but emerge early in people's intuitive knowledge of emotion (Study 1).

1.3 THE PRESENT RESEARCH: INFERRING CAUSES OF EMOTIONS

This thesis reports three studies looking at adults and children's ability to recover unobserved causes of observed emotional expressions. We use a developmental approach in Studies 1, 2.1 and 3 because we are interested in the most fundamental representations that emerge early in development. In Study 2.2, we take a small step toward modeling people's conceptual knowledge of emotion. We formalize the inferences investigated in Study 2.1 with a computational model, to show (preliminarily) that at least some of the conceptual knowledge of emotion can be characterized by richly structured causal models.

All three studies look at emotional cues that have received minimal attention in the domains of both developmental and cognitive sciences. Specifically, Study 1 looks at one to four-year-olds and adults' ability to identify probable causes of diverse positive emotional

vocalizations elicited by funny, exciting, adorable, delicious and sympathetic events. Studies 2.1 and 2.2 look at children and adults' ability to jointly infer someone's beliefs and desires from the dynamics of her emotional responses between anticipated and observed outcomes. Study 3 looks at children's understanding of someone's changing emotional expressions between social and nonsocial contexts. In particular, we are interested in whether they can use such cues to recover not only the desire of the person expressing emotions and that of her conversational partner.

Overall, I aim to show in this thesis that observed emotional cues provide a valuable entrée into the unobserved world. Not only adults, but also children, can use others' emotional expressions to recover information that is otherwise underdetermined in a given context. Such inferences are supported by an abstract, intuitive understanding of how emotions are generated. Although this knowledge may not necessarily mirror the actual processes of how emotions are elicited and differentiated, it supports rational inferences much of time in our daily life, and it may be formed early in development.

Chapter 2 Study 1 Inferring External Causes of Emotional Expressions

This chapter is based on Wu, Muentener, & Schulz. (2017). One- to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proceedings of the National Academy of Sciences*.

2.1 ABSTRACT

The ability to understand why others feel the way they do is critical to human relationships. Here we show that emotion understanding in early childhood is more sophisticated than previously believed, extending well beyond the ability to distinguish basic emotions or draw different inferences from positively and negatively valenced emotions. In a forced-choice task, two- to four-year-olds successfully identified probable causes of five distinct positive emotional vocalizations elicited by what adults would consider funny, delicious, exciting, sympathetic, and adorable stimuli (Experiment 1). Similar results obtained in a preferential looking paradigm with 12-23-month-olds, a direct replication with 18-23-month-olds (Experiment 2), and a simplified design with 12-17-month-olds (Experiment 3; pre-registered). Moreover, 12-17-month-olds selectively explored given improbable causes of different positive emotional reactions (Experiments 4 and 5; pre-registered). The results suggest that by the second year of life, children make sophisticated and subtle distinctions among a wide range of positive emotions and reason about the probable causes of others' emotional reactions. These abilities may play a critical role in developing theory of mind, social cognition, and early relationships.

2.2 SIGNIFICANCE

We find that very young children make fine-grained distinctions among positive emotional expressions and connect diverse emotional vocalizations to their probable eliciting causes.

Moreover, when infants see emotional reactions that are improbable given observed causes, they actively search for hidden causes. The results suggest that early emotion understanding is not limited to discriminating a few basic emotions or contrasts across valence; rather, young children's understanding of others' emotional reactions is nuanced and causal. The findings have implications for research on the neural and cognitive bases of emotion reasoning, as well as investigations of early social relationships.

2.3 INTRODUCTION

Emotions, in my experience, aren't covered by single words. I don't believe in "sadness," "joy," or "regret" ... I'd like to show how "intimations of mortality brought on by aging family members" connects with "the hatred of mirrors that begins in middle age." I'd like to have a word for "the sadness inspired by failing restaurants" as well as for "the excitement of getting a room with a minibar."

— Jeffrey Eugenides, *Middlesex* (2002)

Few abilities are more fundamental to human relationships than our ability to understand why other people feel the way that they do. Insight into the causes of people's emotional reactions allows us to empathize with, predict, and intervene on others' experiences of the world. Unsurprisingly therefore, the conceptual and developmental bases of emotion perception and understanding have been topics of recent interest in a wide range of disciplines, including psychology, anthropology, neuroscience, and computational cognitive science (e.g., Widen, 2016; Beatty, 2013; Barrett, Wilson-Mendenhall, Barsalou, 2014; Skerry & Saxe, 2015; Ong, Zaki, Goodman, 2015; Wu, Baker, Tenenbaum, & Schulz, 2017). Reasoning about emotion has also been a central goal of recent efforts in artificial intelligence (e.g., Minsky, 2007; Scheutz, 2004). However, much remains to be learned even about how young humans understand others' emotional reactions.

As adults, our understanding of emotion is sufficiently sophisticated that English-speakers can appreciate the distinction between "sadness" and "regret"; or "joy" and "excitement." To the degree that we make such distinctions, we represent not only the meaning of emotion words, but also the causes and contexts that elicit them and the expressions and

vocalizations that accompany them. Nonetheless, as the quote by Eugenides suggests, even the abundance of English emotion words may be insufficient to capture the fine-grained relationship between events in the world and our emotional responses. By the same token however, emotion words may not be necessary to such fine-grained representations; even pre-verbal children may represent not only *what* other people feel, but *why*. Here we ask to what extent very young children make fine-grained, within-valence, distinctions among emotions, and both infer and search for probable causes of others' emotional reactions.¹

Newborns respond differently to different emotional expressions within hours of birth (Field, Woodson, Greenberg, & Cohen, 1982). By seven months, infants represent emotional expressions cross-modally and distinguish emotional expressions within valence (e.g., matching happy faces to happy voices and interested faces to interested ones; Walker-Andrews, 1997; Soken & Pick, 1999; Soderstrom, Reimchen, Sauter, & Morgan, 2017). This early sensitivity might reflect only a low-level ability to distinguish characteristic features of emotional expressions rather than any understanding of emotion per se (e.g., Caron, Caron, & Myers, 1985). However, by the end of the first year, infants connect emotional expressions to goal-directed actions. Ten-month-olds look longer when an agent expresses a negative (versus positive) emotional reaction to achieving a goal (Skerry & Spelke, 2014) and 12-month-olds approach or retreat from ambiguous stimuli depending on whether the parent expresses a positive or negative emotion (e.g., Sorce, Emde, Campos, & Klinnert, 1985). Two and three-year-olds explicitly predict that an agent will express positive emotions when her desires are fulfilled and

¹ Our focus here is on children's intuitive theory of emotions rather than scientific theories of what causes or constitutes an emotion. Thus for simplicity we refer to "emotions" throughout although we recognize that scientific theories make important distinctions between for instance, emotion and affect (e.g., Russell & Barrett, 1999).

negative ones when her desires are thwarted (Wellman & Woolley, 1990) and can guess whether someone is looking at something desirable or undesirable based on whether she reacts positively or negatively (Wellman, Phillips, & Rodriguez, 2000).

Nonetheless, some work suggests that infants and toddlers often fail to connect others' emotional reactions to specific events in the world. Nine-month-olds use novel words, but not differently valenced emotional reactions to distinguish object kinds (Xu, 2002), and 14-month-olds use the direction of someone's eye gaze, but not her emotional expressions to predict the target of her reach (Vaish & Woodward, 2010). Similarly, if an experimenter frowns at a food a child likes but smiles at a food she dislikes, 14-month-olds fail to use the valenced reactions to infer that the experimenter's preferences differ from the child's own (Repacholi & Gopnik, 1997). The interpretation of such failures is ambiguous: object labels and direction of gaze may be more reliable than emotional expressions as cues to object individuation and direction of reach; similarly, infants may understand that emotions indicate preferences while resisting the idea that others' preferences differ from their own. Critically however, whether implying precocious or protracted emotion understanding, prior work on early emotion understanding has focused almost exclusively on children's ability to distinguish a few basic emotions or draw different inferences from positively and negatively valenced emotional expressions. Thus it is unclear to what extent young children make more fine-grained distinctions.

Indeed, some researchers have proposed that children initially categorize emotions only as "feeling good" and "feeling bad" (see 2 for review and discussion). Children struggle with explicitly labeling and sorting emotional expressions well into middle childhood, and are often more successful at identifying emotion labels given information about the cause and behavioral consequences of the emotion than given the emotional expression itself (Widen & Russell, 2004;

2010). Consistent with this, preschoolers know numerous “scripts” connecting familiar emotions and events (e.g., getting a puppy and happiness; dropping an ice cream cone and sadness; Barden, Zelko, Duncan, & Masters, 1980; Stein & Levine, 1989; Stein & Trabasso, 1992).

Insofar as individuals’ emotional reactions depend on their appraisal of events (Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony, Clore, & Collins, 1990), such relationships hold only in probability. However, children’s later developing ability to integrate their understanding of emotion with information about others’ beliefs and desires (see Wellman, 2014; Harris, 2016 for reviews) might be supported by an earlier ability to connect emotional reactions to their probable causes just as causal knowledge supports categorization and conceptual enrichment in other domains (Carey, 2009; Gopnik & Wellman, 2012). Indeed, arguably, it would be surprising if infants were entirely insensitive to such predictive relationships given their general abilities at statistical learning (e.g., Saffran, Johnson, Aslin, & Newport, 1999) and the sophistication of early social cognition overall (see Hamlin, 2013 for review). To the degree that particular kinds of events are reliably associated with particular kinds of emotional reactions within a given cultural context (Lutz, 1982), young children might learn relationships between eliciting events and emotional responses just as they learn probabilistic relationships in other domains (Bergelson & Swingley, 2012; Teglas, Vul, Girotto, Gonzalez, Tenenbaum, & Bonatti, 2011). In this way, children might make nuanced distinctions among emotions well before they learn the words – if any – corresponding to such distinctions.

2.4 EXPERIMENTS 1-3

To look at young children’s understanding of relatively fine-grained relationships between eliciting events and others’ emotional reactions, we present participants with generative causes of five distinct positive emotional vocalizations and ask whether children can link the

vocalization with the probable eliciting cause. We focus on positive emotions because little is known about children's ability to make these kinds of discriminations (i.e., none of the contrasts tested here are represented in distinctions among basic emotions like happiness, sadness, anger, fear, and disgust).

2.4.1 Eliciting cause stimuli and emotional vocalizations

Two female adults blind to the study design were asked to vocalize a non-verbal emotional response to images from five categories: Funny (children making silly faces), Delicious (desserts), Exciting (light-up toys), Adorable (cute babies) and Sympathetic (crying babies). Four individual pictures were chosen from each of the categories, resulting in a set of 20 different images. See Supporting Information (SI) Appendix, Fig. S1, for representative images. One vocalization was chosen for each image. (Audio files here: https://osf.io/an57k/?view_only=def5e66600b0441482c10763541e3ac2²) The categories were chosen semi-arbitrarily, constrained by the criteria that the images had to be recognizable to young children, and elicit distinct positive emotional reactions from adults. With respect to crying babies, note that although the image is negative, the adult response was positive and consoling.

² We cannot do justice here (but see Scherer, 1994 for review) to the interesting question of when spontaneous, emotional responses to stimuli become paralinguistic or entirely lexicalized parts of speech (e.g., the involuntary cry of pain to “ouch!”; the gasp of surprise to “oh!”); however, as Scherer notes, there may be points on the continuum where no clear distinction can be made. Given the conditions under which they were elicited, we believe it is reasonable to treat the current stimuli as intentional, communicative (albeit non-verbal) affect bursts. However, for the current purposes, nothing rests on drawing a sharp distinction between involuntary and voluntary exclamations.

2.4.2 Experiment 1: 2 to 4-year-olds and adults

On each trial one image from each of two different categories was randomly selected and presented on different sides of a screen. The vocalization elicited by one of the images was randomly selected and played. Each image was seen on exactly two test trials, once as the target and once as the distractor; each vocal expression was played on only a single test trial. Children were told the sound was made by a doll (Sally) who sat facing the screen and were asked “Which picture do you think Sally is looking at?”

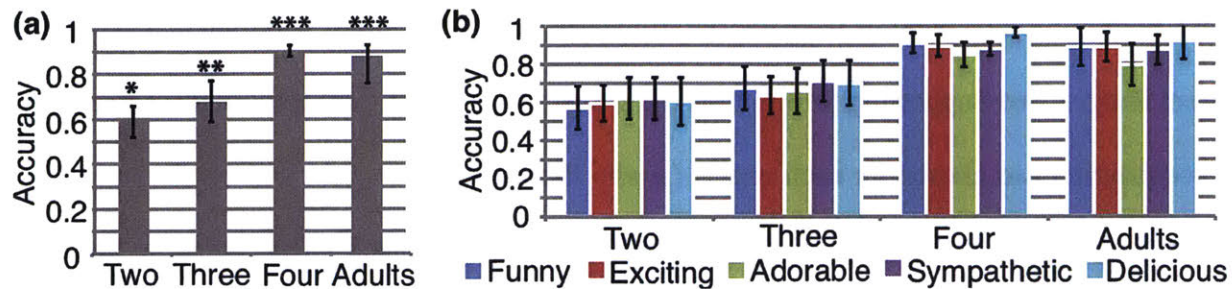


Fig. 1. Results of Experiment 1. (a) Accuracy by age group. (b) Accuracy by age group and the category of the eliciting cause. Error bars indicate 95% confidence intervals. * $p < .05$; ** $p < .01$; *** $p < .001$.

Accuracy was calculated as the number of correct responses over the total number of trials completed. Overall, children successfully matched the vocalizations to their probable eliciting causes ($M = .73$, $SD = .192$, 95% $CI [.67, .78]$, $t(47) = 8.18$, $p < .001$, $d = 1.18$; one sample t test, two-tailed). A mixed-effects model showed no main effect of the category of the eliciting cause ($F(4, 188) = 0.93$, $p = .449$) but a main effect of age ($F(1, 46) = 32.77$, $p < .001$). (See SI Appendix 1.1 and Tables S1-S2 for more information about the model.) Post-hoc analyses found that children in every age bin succeeded (collapsing across categories, two-year-olds: $M = .60$, $SD = .143$, 95% $CI [.52, .66]$, $t(15) = 2.75$, $p = .015$, $d = .69$; three-year-olds: $M = .68$, $SD = .194$, 95%

$CI [.58, .77]$, $t(15)=3.64$, $p=.002$, $d=.91$; four-year-olds: $M=.90$, $SD=.055$, 95% $CI [.88, .93]$, $t(15)=29.59$, $p<.001$, $d=7.40$). A group of adults also succeeded ($M=.88$, $SD=.153$, 95% $CI [.76, .93]$, $t(15)=9.82$, $p<.001$, $d=2.45$). Two and three-year-olds performed similarly to each other ($p=.466$; Tukey's test); neither age group reached adult-like performance ($ps<.001$). By contrast, four-year-olds differed from younger children ($ps<.001$) and were indistinguishable from adults ($p=.950$). See Fig. 1.

2.4.3 Experiments 2-3: 12-23-month-olds

Given that even two-year-olds selected the target picture above chance, we asked whether younger children ($N = 32$) might succeed at a non-verbal version of the task. The materials were identical to those in Experiment 1. On each trial, two images were presented on the screen. A vocalization corresponding to one image was played for 4 seconds, followed by a 3 second pause; then the other vocalization was played. See Fig. 2(a).

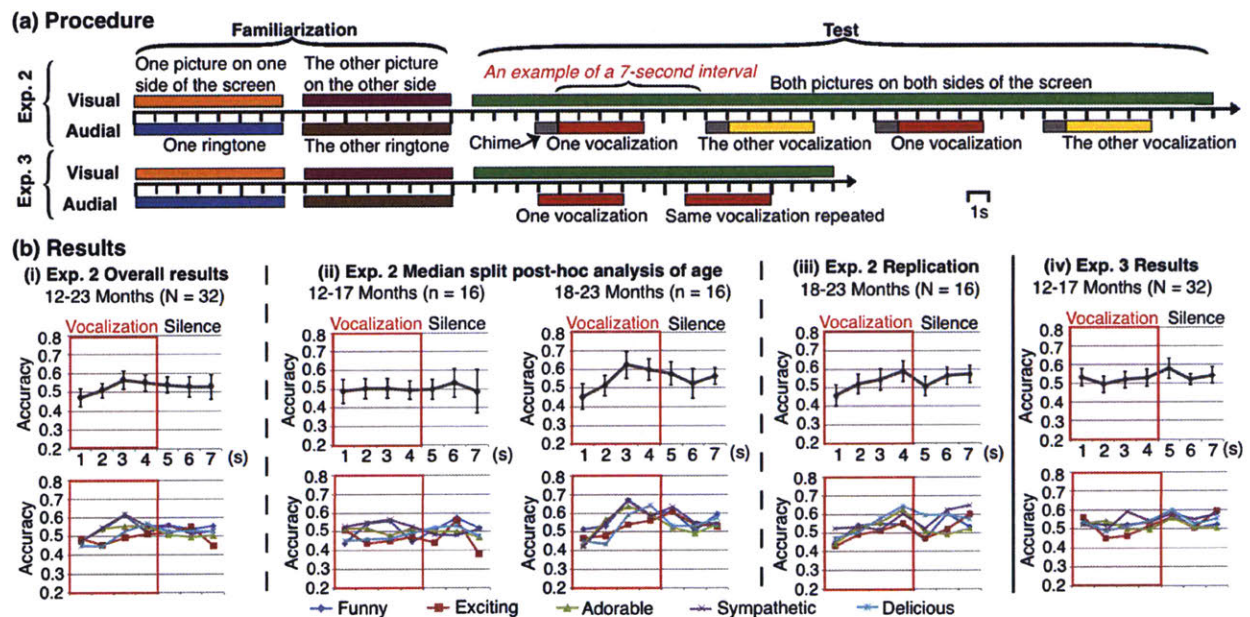


Fig. 2. Procedure and results of Experiments 2 and 3. (a) Procedure of each trial of the preferential-looking task. (b) Mean proportion of accurate looking to the corresponding picture over total looking to both pictures during the 7-second intervals. Upper row: accuracy collapsing across categories; lower row: accuracy by category. Error bars indicate 95% confident intervals.

Overall, children preferentially looked at the picture corresponding to the vocalization ($M=.53$, $SD=.070$, 95% CI [.50, .55], $t(31)=2.05$, $p=.049$, $d=.36$). A mixed-effects model, with a 7 (Second) by 5 (Category) by 32 (Subject) data matrix showed no effect of Age ($F(1, 30)=1.79$, $p=.191$) or Category ($F(4, 980)=1.39$, $p=.235$), but a main effect of Time ($F(6, 980)=3.34$, $p=.003$), consistent with children shifting their gaze towards the correct picture over each 7-second interval. See Fig. 2(b)(i). We also found a significant interaction between Age and Time ($F(6, 980)=2.84$, $p=.010$). (See SI Appendix 1.2 and Tables S3-S4.) As an exploratory analysis we performed a median split and looked separately at the 12-17-month-olds and the 18-23-month-olds. The 12-17-month-olds performed at chance ($M=.50$, $SD=.075$, 95% CI [.47, .54], $t(15)=.17$, $p=.866$, $d=.04$), and there was neither a main effect of Category ($F(4, 486)=1.02$, $p=.396$) nor Time ($F(6, 486)=0.74$, $p=.621$). In contrast, the 18-23-month-olds successfully matched the vocalization to the corresponding picture ($M=.55$, $SD=.059$, 95% CI [.52, .57], $t(15)=3.23$, $p=.006$, $d=.81$). There was no main effect of Category ($F(4, 490)=0.69$, $p=.596$) but a significant main effect of Time ($F(6, 490)=6.55$, $p<.001$). See Fig. 2(b)(ii). See also SI Appendix 1.2 and Table S5.

Children's average looking time at the target was only slightly above chance. This is perhaps unsurprising given a number of factors: the fine-grained nature of the distinctions, that the relationships between eliciting causes and reactions hold only in probability, that children had to move their gaze to the target over the 7-second interval (see Figure 2(b)(ii)), and that the visual stimuli were not matched for salience; in particular, children sometimes observed an object and an agent on the screen simultaneously. However, because the effect was subtle and the age split was post-hoc, we replicated the experiment with a separate group of 18-23-month-olds ($N=16$). As in the initial sample, participants preferentially looked at the target picture ($M=.53$, $SD=.045$, 95%

$CI [.51, .56]$, $t(15)=3.04$, $p=.008$, $d=.76$); there was no main effect of Category ($F(4, 527)=2.02$, $p=.090$) but a significant main effect of Time ($F(6, 527)=6.96$, $p<.001$), consistent with children moving their gaze towards the target. See Fig. 2(b)(iii). See also SI Appendix 1.2 and Table S6. Across the initial sample and the replication, a preference for the target across trials was observed in most of the 18-23-month-olds tested (27 of 32).

As noted, in Experiment 2 the vocalizations alternated on each trial so children had to switch from looking at one picture to looking at the other. These task demands may have overwhelmed the younger children so in Experiment 3 (pre-registered here: https://osf.io/m3u67/?view_only=3da43a84fd004f4095ac65ae298c567c), we tested 12-17-month-olds ($N=32$) using a simpler design in which only a single vocalization was played repeatedly on each trial. See Fig. 2(a). Infants looked at the matched picture above chance across trials ($M=.53$, $SD=.055$, 95% $CI [.51, .55]$, $t(31)=3.14$, $p=.004$, $d=.55$). The mixed-effects model showed no main effect of Age ($F(1, 30)=.03$, $p=.858$) or Category ($F(4, 1052)=.50$, $p=.736$), but a main effect of Time ($F(6, 1052)=2.54$, $p=.019$), consistent with infants shifting their looking towards the target picture. See Fig. 2(b)(iv). See also SI Appendix 1.3 and Tables S7-S8.

2.5 EXPERIMENTS 4-5: 12-17-MONTH-OLDS

To validate the previous results with converging measures, and also look at the extent to which infants might actively search for causes of others' emotional reactions, in Experiments 4 and 5, we tested 12-17-month-olds using a manual search task (adapted from Feigenson & Ccarey, 2003; Xu, Cote, & Baker, 2005). See Fig. 3(a). The experimenter peeked through a peep hole in top of a box and made one of two vocalizations (Experiment 4: "Aww!" or "Mmm!"; Experiment 5: "Aww!" or "Whoa!"). Infants were encouraged to reach into a felt slit on the side of the box. They retrieved a toy that was either Congruent or Incongruent with the vocalization

(Experiment 4: half the infants retrieved a stuffed animal; half retrieved a toy fruit; Experiment 5: half retrieved a stuffed animal; half retrieved a toy car). The experimenter took the retrieved toy away and looked down for 10 seconds. We coded whether infants reached into the box again, and how long they searched. A new box was introduced for a second trial. Infants who had retrieved a Congruent toy on the first test trial retrieved an Incongruent toy on the second test trial and vice versa. We were interested in whether infants would search longer on the Incongruent than Congruent Trials. Because the effect of congruency was the primary question of interest (rather than the effect of the particular emotion contrast tested in each experiment), we report the results of each individual experiment and then a summary analysis of the effect of congruency across both experiments.

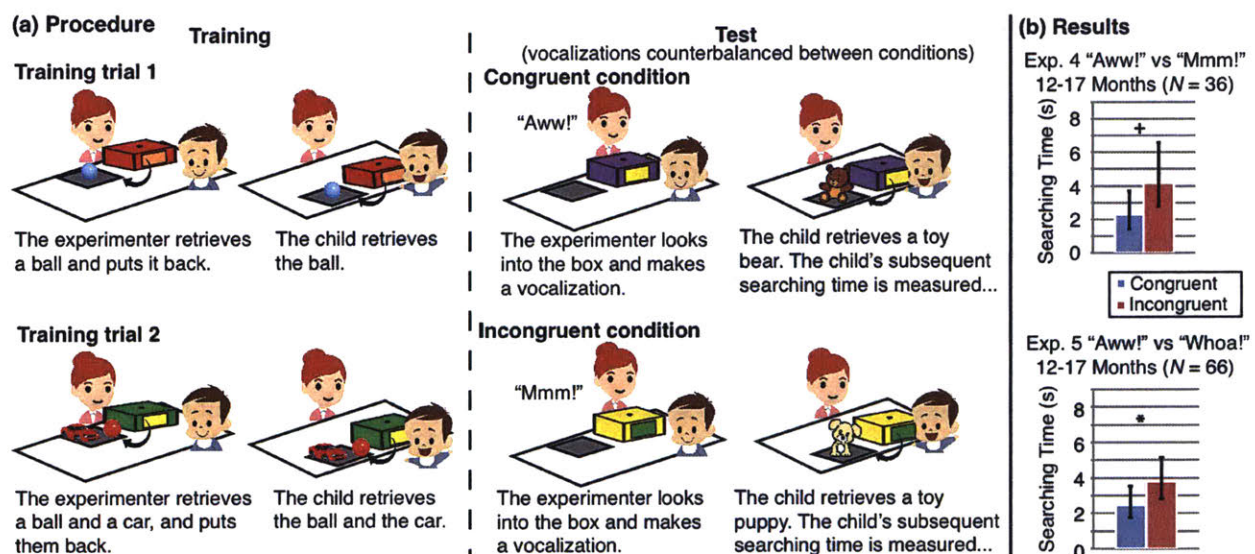


Fig. 3. Procedure and results of Experiments 4 and 5. (a) An example of the procedure of the manual search task. (b) Infants' searching time in the congruent and incongruent conditions. Error bars indicate 95% confident intervals. + $p < .10$; * $p < .05$.

Experiment 4 Results

One analysis was pre-registered: a permutation test on the search time (https://osf.io/9qwcp/?view_only=25bb71fc775748f3a6cab34cf6734dae). There was a trend for infants to search longer in the Incongruent than the Congruent condition (Incongruent: $M=4.18s$, $SD=5.600$, 95% CI [2.76, 6.61]; Congruent: $M=2.29s$, $SD=3.436$, 95% CI [1.40, 3.76]; $T=35.34$, $p=.053$). See Fig. 3(b).

Experiment 5 Results

Four analyses were pre-registered: 1) a mixed effects model of the raw data, 2) a permutation test of the raw data, 3) a permutation test of the proportional searching, and 4) the non-parametric McNemar's test of the number of children searching in each condition (https://osf.io/a4yed/?view_only=79b03b38f6eb47518107f5851b8770ff; see also SI Appendix 2). The mixed effect model revealed no main effect of Category ($F(1, 63)=.20$, $p=.656$) but a main effect of Age ($F(1, 63)=4.27$, $p=.043$), suggesting that older infants searched longer overall than younger ones, and a significant main effect of Congruency ($F(1, 65)=4.47$, $p=.038$). The interaction between Age and Congruency did not survive model selection (see SI Appendix 1.4 and Tables S9-S10) thus no further age analyses were conducted. Infants searched longer in the Incongruent ($M=3.82s$, $SD=4.818$, 95% CI [2.81, 5.17]) than Congruent condition ($M=2.48s$, $SD=3.557$, 95% CI [1.77, 3.52]; $T=67.25$, $p=.039$; permutation test). See Fig. 3(b). Neither the proportional search time nor the number of infants who searched at all differed by condition.

Summary analysis

A meta-analysis (McShane & Bockenholt, 2017) across both experiments, found that infants searched longer in the Incongruent than Congruent conditions (effect: 1.51, 95% CI [.48, 2.54]; I^2 : .00, 95% CI [.00 13.76]). They also spent proportionally more time searching the

Incongruent than Congruent box (effect: .18, 95% *CI* [.06, .29]; F : 31.93, 95% *CI* [.00 92.92]). Finally, they were more likely to search again after retrieving the first toy given the Incongruent than Congruent box (effect: .18, 95% *CI* [.01, .36]; F : 58.13, 95% *CI* [.00 88.07]). See SI Appendix, Fig. S2.

Note that in the incongruent condition of Experiments 4 and 5, the probable eliciting cause was never observed. Thus the results cannot be due to infants merely associating the stimulus and the emotional reaction, or generating their own first person response to the eliciting cause. Rather, the results suggest that infants represent probable causes of others' emotional reactions, and actively search for unobserved causes when observed candidate causes are implausible.

2.6 GENERAL DISCUSSION

Across five experiments, we found that very young children make nuanced distinctions among positive emotional vocalizations and connect them to probable eliciting causes. The results suggest that others' emotional reactions provide a rich source of data for early social cognition, allowing children to recover the focus of others' attention in cases otherwise underdetermined by the context (Experiments 1-3) and to search for plausible causes of others' emotional reactions, even when the causes are not in the scene at all (Experiments 4 and 5).

As noted, the vast majority of previous work on early emotion understanding has focused on children's understanding of a few basic emotions, or on distinctions between positively and negatively valenced emotions (see Widen, 2016 for review). Similarly, influential accounts of emotional experience in both adults and children have often focused exclusively on dimensions of valence and arousal (e.g., Kensinger, 2004; Kuppens, Tuerlinckx, Russell, & Barrett, 2013; Posner, Russell, & Peterson, 2005). Our results go beyond previous work in suggesting that very

young children make nuanced distinctions among positive emotional reactions and have a causal understanding of emotion: they recognize that events in the world generate characteristic emotional responses. These findings have a number of interdisciplinary implications, raising questions about how the findings might generalize to other socio-cultural contexts and other species, constraining hypotheses about the neural and computational bases of early emotion understanding, and suggesting new targets for infant-inspired artificial intelligence systems.

In representing the relationship between emotional vocalizations and probable eliciting causes, what are children representing? One possibility is that infants deploy their general ability to match meaningful auditory and visual stimuli (e.g., Bergelson & Swingley, 2012; Smith & Yu, 2008) to connect emotional vocalizations with eliciting events, and to search for plausible elicitors when they are otherwise unobserved, but without representing emotional content per se. On this account, children might either learn predictive relationships between otherwise arbitrary classes of stimuli or they might represent the vocalizations as having non-emotional content. For instance, they might assume the vocalizations identify rather than react to the targets (i.e., like object labels). We think this is unlikely however, given that neither natural nor artifact kinds capture the distinctions infants made in this study (e.g., grouping light-up toys and toy cars together on the one hand, and stuffed animals and babies on the other.) Alternatively, children might treat the vocalizations adjectivally. However, this too seems unlikely given that sensitivity to modifiers is rare in the second year of life and that the vocalizations were uttered in isolation not in noun phrases (“Aww...” not “the aww bunny”; Mintz & Gleitman, 2002).

Note however, that even considering only the five positive emotions distinguished here, many other eliciting events were possible (e.g., cooing over pets, clucking over skinned knees, ahing over athletic events, etc.). Given myriad possible combinations, inferring abstract

relations may simplify a difficult learning problem and indeed, be easier than learning individual pairwise mappings (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Thus another explanation, and one we favor, is that, infants represent at the very least, proto-emotion concepts. Given that infants engage in sophisticated social cognition in many domains (e.g., distinguishing pro- and anti-social others, in-group and out-group members and the equitable and inequitable distribution of resources; Hamlin, 2013; Powell & Spelke, 2013; Schmidt & Sommerville, 2011), it is at least conceivable that infants also represent the fact that some stimuli elicit affection, others excitement, and others sympathy etc. However, the degree to which infants' early representations have relatively rich information content even in infancy, or serve primarily as placeholders to bootstrap the development of richer intuitive theories later on (e.g., Carey, 2009), remains a question for further research.

As noted at the outset, these data bear upon the development of children's intuitive theory of emotions and are orthogonal to debates over what emotion is and how it is generated (see also Skerry & Saxe, 2015; Ong, Zaki, & Goodman, 2015; Wu, Baker, Tenenbaum, & Schulz, 2017). That adults may be capable of nuanced distinctions among emotions does not, in itself, invalidate the possibility that there are a small set of innate, evolutionarily specified, universal emotions ("basic emotions"; Ekman, 1992), that emotions can be characterized primarily on dimensions of valence and arousal (Russell, 2003), that emotions depend on individuals' appraisal of events (Moors, Ellsworth, Scherer, & Frijda, 2013), or that emotions arise, as other concepts do, from cultural interactions that cannot be reduced to any set of physiological and neural responses (Barrett, Wilson-Mendenhall, & Barsalou, 2014). Finding that young children are sensitive to some of the same distinctions as adults similarly does not resolve these disputes one way or the other. Nor are these findings in tension with the finding that it may take many years for children

to learn explicit emotion categories (see Widen, 2016 for review). Indeed, arguably, the very richness of infants' early representations may make it particularly challenging for infants to isolate the specific emotion concepts reified within any given culture. These results do however, suggest that early in development, infants make remarkably fine-grained distinctions among emotions, and represent causal relationships between events and emotional reactions in a manner that could support later conceptual enrichment.

Finally, we note that even in adult judgment there may be some dispute about the degree to which the response to each of the eliciting causes investigated here qualifies as an emotional reaction per se. People may say they feel “excited” on seeing a toy car or light-up toy or “amused” when they see silly faces, however there is no simple English emotion word that captures the response to seeing a cute baby (endeared? affectionate?), a crying baby (sympathetic? tender?), or delicious food (delighted? anticipatory?). We believe this may speak more to the impoverished nature of English emotion labels than to the absence of emotional responses to such stimuli. As our opening quotation illustrates, there are myriad emotion words in English but even these are not exhaustive. The fact that cultures vary in both the number and kind of emotional concepts they label (Lutz, 1982), suggests that we may be capable of experiencing more than any given language can say. The current results however, suggest that at least some of the subtleties and richness of our emotional responses to the world are accessible, even in infancy.

2.7 MATERIALS AND METHODS

Participants. Child participants were recruited from a children's museum; adults were recruited on Amazon Mechanical Turk. Experiment 1 included 48 children (mean: 3.4 years, range: 2.0-4.8) and 16 adults. Experiment 2 included 32 12-23-month-olds (mean: 17.8 months, range: 12.2-23.3).

An additional 16 18-23-month-olds (mean: 21.2 months, range: 18.1-23.9) were recruited for a replication study. Experiment 3 included 32 12-17-month-olds (mean: 14.8 months, range: 12.1-17.8). Experiment 4 included 36 12-17-month-olds (mean: 14.8 months, range: 12.0-17.9). The sample size was determined by a power analysis, using the effect size found in a pilot study (see SI Appendix 3) and setting $\alpha=.05$ and $\text{power}=.80$. In Experiment 5, we increased our power to .90, and ran a power analysis based on the effect size found in 15-17-month-olds in Experiment 4, but allowing power to test for the age difference revealed in an exploratory analysis (see SI Appendix 4). This resulted in a sample size of 33 12-14-month-olds (mean: 13.7 months, range: 12.2-14.8) and 33 15-17-month-olds (mean: 16.3 months, range 15.0-17.6). See SI Appendix 5 for exclusion criteria. Parents provided informed consent; the MIT Institutional Review Board approved the research. Data and R code for analyzing the data can be found here: https://osf.io/ru2t4/?view_only=54d9cf5b3a1141729a4e5b3d0a1e01a6.

Materials. In Experiment 1, images were presented on a 15-inch laptop; vocalizations were played on a speaker. A doll was placed on the speaker. A warm-up trial used a picture of a beautiful beach and a picture of a dying flower. Training vocalizations were elicited as for test stimuli. In Experiments 2 and 3, images were presented on a monitor (93cm x 56cm); vocalizations were played on a speaker. Two 7-second ringtones were used for familiarization; a chime preceded the presentation of the vocalizations in Experiment 2 only. A multi-colored pinwheel was used as an attention getter. In Experiment 4, four different colored cardboard boxes (27cm × 26cm × 11cm) were used. A 4 cm diameter hole was cut in the top of each box allowing a partial view of the interior; a 19 cm × 8 cm opening was cut on the side of each box. The opening was covered by two pieces of felt. Velcro on the table and the bottom of the boxes was used to standardize the placement of the boxes. Two boxes were used in the training phase

to teach infants that there could be either one (a 5cm blue ball) or two (a 7cm red ball, and a 11cm x 5cm x 5cm toy car) objects inside. The other two boxes were used in the test phase. For half the participants, a stuffed animal was inside each box (a bear in one and a puppy in the other; each approximately 14cm × 9cm × 7cm). For the remaining participants, a toy fruit was inside each box (an 8cm orange in one and a 19cm banana in the other). A black tray was used to hold the items retrieved from the boxes. A timer was used in the test phase. The same materials were used in Experiment 5 except that one training box contained a blue ball while the other contained a red ball and a toy banana. For half the children, both test boxes contained cute stuffed animals (a bear in one and a puppy in the other); for the remaining children, both test boxes contained toy cars (one red and one blue).

Procedure. In Experiment 1, the experimenter introduced the doll, saying, “Hi, this is Sally! Today we’ll play a game with Sally!” The experimenter placed the doll on the speaker, facing the laptop screen. A practice trial (see SI Appendix 6.1) preceded the test trials. On each test trial, the experimenter pushed a button on the keyboard to trigger the presentation of two pictures and said: “Here are two new pictures, and Sally makes this sound.” Then she pushed a button on the keyboard to trigger the vocalization. She asked the child: “Which picture do you think Sally is looking at?” There were a total of 20 trials presented in a random order.

Adults were tested online. They were told that the vocalization on each trial was someone’s response to one of the pictures and their task was to guess which picture the person was looking at. We generated a randomly ordered set of ten picture pairs. Half the adults were given the vocalization corresponding to one picture in each pair; the other half were given the vocalization corresponding to the other picture. Adults had ten trials rather than twenty.

In Experiment 2, the child's parent sat in a chair with her eyes closed. The child sat on the parent's lap, approximately 63 cm in front of the screen. The experimenter could see the child on a camera but was blind to the visual stimuli throughout. At the beginning of each trial, the attention getter was displayed. When the child looked at the screen, the experimenter pressed a button to initiate the familiarization phase. The computer randomly selected one image (17cm x 12cm) from one category and presented it on one side of the screen (left/right counterbalanced), accompanied by one of the two ringtones. After 7 seconds, the image disappeared. Then the computer randomly selected another image (17cm x 12cm) from a different category and presented it on the other side of the screen, accompanied by the other ringtone. After 7 seconds, the image disappeared. For the test phase, both pictures were presented simultaneously. A chime was played to attract the child's attention and the vocalization corresponding to either the left or right picture (randomized) was played for 4 seconds followed by 3 seconds of silence. The chime was played again, and the vocalization corresponding to the other picture was played for 4 seconds followed by 3 seconds of silence. This was repeated. Then the computer moved on to the next trial. See Fig. 2(a).

In Experiment 3, the procedure was identical to Experiment 2 except that during the test trial, a single vocalization (corresponding to one image, chosen at random) was played for 4 seconds, followed by a 3 second pause. This was then repeated. See Fig. 2(a).

In Experiment 4, the experimenter played with the child with some warm-up toys and then initiated the training phase (see SI Appendix 6.2). After the training phase, she introduced the test box containing either a toy bear or a toy fruit, counter-balanced across participants. The experimenter said: "Here is another box. Let me take a look." She looked into the box and said either "Aww!" (as if seeing something adorable) or "Mmm!" (as if seeing something yummy)

counter-balanced across participants. She looked at the child, looked back into the box, and then repeated the vocalization. She repeated this a third time. Then the experimenter affixed the box to the table with the felt opening facing the child, and encouraged the child to reach in the box, retrieve the toy, and put the toy in the tray. Once the child did, the experimenter removed the toy, set the timer for 10 seconds, and looked down at her lap. After 10 seconds, the experimenter looked up. If the child was still searching, the experimenter looked down for another 10 seconds. She repeated this until the child was not searching when she looked up. (The box was always empty at this point but it was hard for the infant to discover this given the size of the box, peep hole and felt opening.) The experimenter then moved on to the second test box. This box contained a toy similar to the one in the previous box (i.e., a puppy if the previous one was a bear; a banana if the previous one was an orange) but the experimenter made the other vocalization (i.e., “Mmm!” if she had said “Aww!” or “Aww!” if she had said “Mmm!”). Thus within participants, the object the child retrieved was congruent with the vocalization on one trial and incongruent on the other (order counterbalanced). In Experiment 5, the procedure was identical except that the “Mmm!” was replaced with a “Whoa!” Children’s looking and searching behavior were all coded offline from video clips. See SI Appendix 7 for details.

Chapter 3 Study 2.1 Inferring Beliefs and Desires From Emotional Expressions

This chapter is based on Wu & Schulz. (2017). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*.

3.1 ABSTRACT

Researchers have long been interested in the relation between emotion understanding and theory of mind. This study investigates a cue to mental states that has rarely been investigated: the dynamics of valenced emotional expressions. When the valence of a character's facial expression was stable between an expected and observed outcome, children ($N = 122$; $M = 5.0$ years) recovered the character's desires but did not consistently recover her beliefs. When the valence changed, older, but not younger children recovered both the characters' beliefs and desires. By contrast, adults jointly recovered agents' beliefs and desires in all conditions. These results suggest that the ability to infer mental states from the dynamics of emotional expressions develops gradually through early and middle childhood.

Keywords: emotion understanding; theory of mind; mental state inference; preschoolers

3.2 INTRODUCTION

Researchers have long noted correlations between the development of children's belief-desire psychology and their understanding of emotions (e.g., Bartsch & Estes, 1996; de Rosnay, Fink, & Begeer, Slaughter, & Peterson, 2014; Harwood & Farrar, 2006; Hughes & Dunn, 1998; LaBounty, Wellman, Olson, Lagattuta, & Liu, 2008; Wellman, 2014; Widen, 2013; Widen & Russell, 2008), and proposed that children construct an intuitive theory of mind in which beliefs, desires, and emotions are causally linked (e.g., Harris, Johnson, Hutton, Andrews, & Cooke, 1989; Thompson & Lagattuta, 2006; Lagattuta, Wellman, & Flavell, 1997; Lagattuta & Wellman, 2001; see Harris, 2008 and Wellman, 2014 for discussion and review). Here we ask to what extent this intuitive theory allows children to recover others' beliefs and desires from their emotional reactions to events.

Given the extensive history of work on theory of mind and emotion, some justification is required for posing this as an unanswered question. Note however, that much of this research has looked at children's ability to use knowledge of others' beliefs and desires to predict their emotions. Here we are interested in the inverse problem: children's ability to use emotional expressions to recover other mental states. To follow, we briefly summarize past research and then ask whether children can use someone's emotional reactions to anticipating and observing an outcome to jointly infer her beliefs and desires.

Human beings are sensitive to others' emotional expressions from birth (Field, Woodson, Greenberg, & Cohen, 1982). Infants show different patterns of behavior in response to happy, fearful, sad, and angry faces and voices (Field et al., 1982; Haviland & Lelwica, 1987; Montague, Walker-Andrews, 2001), represent emotions cross-modally (Walker-Andrews, 1997) and discriminate expressions even within valence (e.g., matching happy faces to happy voices

and interested faces to interested voices; Soken & Pick, 1999; see also Flom & Bahrick, 2007; Haviland & Lelwica, 1987; Hoehl & Striano, 2008; Soderstrom, Reimchen, Sauter, & Morgan, 2015). Older infants check their parents' faces given ambiguous stimuli and approach or retreat depending on the valence of the parents' facial expression (e.g., Hornik & Gunnar, 1988; Moses, Baldwin, Rosicky, & Tidball, 2001; Mumme & Fernald, 2003; Sorce, Emde, Campos, & Klinnert, 1985; Walden & Ogan, 1988).

Infants might respond differentially to emotional expressions without understanding emotions *per se*; however recent research suggests that even infants relate emotional expressions to goal-directed actions. As early as eight months, infants expect agents to express positive rather than negative emotions when they achieve their goals (Skerry & Spelke, 2014). By two, children explicitly predict that an agent will be happy if her desires are fulfilled and sad if they are thwarted (e.g., Wellman & Wooley, 1990; Yuill, 1984). In two-year-olds, these emotion concepts may be relatively undifferentiated, distinguishing primarily between positive and negative valences (see e.g., Widen & Russell, 2008; 2010); nonetheless, children appropriately map goal-fulfillment onto positive emotions and goal-failure to negative ones.

More mixed findings obtain for belief inferences. As young as three, children seem to experience suspenseful emotions in response to false belief scenarios; they are more likely to furrow their brow and bite their lips when they see that someone is about to act consistent with a false (versus true) belief (Moll, Kane, & McGowan, 2016). Between four and six, children become increasingly likely to predict that someone else will feel surprised if her beliefs are violated, and feel happy if she falsely believes that an action will fulfill her desires (Hadwin & Perner, 1991; Harris, et al., 1989; Wellman & Banerjee, 1991). However, considerable research suggests that children's ability to attribute the emotional reactions generated by true and false

beliefs lags behind their ability to explicitly represent the beliefs themselves (e.g., de Rosnay, Pons, Harris, & Morell, 2004; Hadwin & Perner, 1991; Harris, et al., 1989; Pons, Harris, & de Rosnay, 2004; Ruffman & Keenan, 1996; Wellman & Bartsch, 1988). Thus four- and five-year-olds may know that Red Riding Hood falsely believes her grandmother is in bed, and nonetheless conclude that Red Riding Hood is frightened (Bradmetz & Schneider, 1999).

Fewer studies have looked at children's ability to reason backwards from emotional reactions to desires and beliefs, and again the strongest evidence is for inferences about desires. By eighteen months, infants can use an agent's verbal cues ("Yummy!" versus "Yucky!") together with her emotional expressions to decide if she wants a food different from what the child herself wants (e.g., broccoli rather than goldfish crackers; Repacholi & Gopnik, 1997). Similarly, two- and three-year-olds can use someone's emotional reaction to infer whether she is looking at desirable crackers or undesirable broccoli (Wellman, Philips & Rodriguez, 2000). By preschool, children map happy and sad emotional reactions onto familiar desirable or undesirable events (e.g., getting a puppy or dropping an ice cream cone; Denham, Zoller, & Couchoud, 1994; Fabes, Eisenberg, McCormick & Wilson, 1988; Gnepp, McKee & Domanic, 1987; Harris, Olthof, Terwogt, & Hardman, 1987; Widen & Russell, 2010).

However, children younger than six rarely refer to agents' beliefs in explaining others' emotional reactions (Rieffe, Terwogt & Cowan, 2005). One exception is that four-year-olds mention beliefs in explaining fearful reactions (e.g., saying "She thought it was a ghost" if a character is scared at hearing a noise) and atypical ones (saying "She thought it would be something else" if a character is sad upon opening a present; Rieffe, et al., 2005). Four-year-olds also spontaneously refer to beliefs in explaining surprise or curiosity, and do so more often given these "epistemic" emotions than happy or sad ones (Wellman & Banerjee, 1991).

As noted however, by preschool, children have learned scripts relating familiar events and emotions (e.g, Fabes, et al., 1988; Gnepp, et al., 1987; Harris, et al., 1987; Widen & Russell, 2010). Children might link fear and ghosts, or a disappointing gift with sadness (Rieff, et al., 2005) without necessarily reasoning about the relation between emotions and beliefs more broadly. Similarly, children might guess that an agent did not know about, or expect unusual or mysterious events (“she didn’t think there would be a giraffe”; “she didn’t know what was in the box”; Wellman & Banerjee, 1991), because the events are atypical or mysterious rather than because they reason about the beliefs underlying emotional reactions generally.

Perhaps the strongest support for the idea that children infer the thoughts underlying emotional responses comes from work showing that children invoke others’ thoughts about the past to explain their emotions in the present (e.g., Harris, Guz, Lipian & Man-shu, 1985; Lagattuta, et al, 1997; Lagattuta & Wellman, 2001; Taylor & Harris, 1983). Four-year-olds recognize that people respond more intensely to recent events than past events; by six, children recognize that people’s responses to events depends on whether they remember them, and that people are happier remembering positive experiences and forgetting negative ones (e.g., Harris, et al., 1985; Taylor & Harris, 1983). Children also understand that individuals’ particular histories can lead to idiosyncratic emotional responses. If a girl’s doll is broken by a clown, children predict that she will be sad on seeing another clown and explain her sadness by saying she is thinking about her doll (Lagattuta, et al., 1997; Lagattuta & Wellman, 2001).

Findings like these suggest that children understand that thoughts affect feelings. However, they do not address the question of whether children can use someone’s emotional reactions to recover the content of her otherwise unknown thoughts. Theory of mind is a hard inference problem because it often involves situations where the only information available is

that which can be gleaned from the environment and observed behavior. In these cases, an observer has no more access to others' past history of emotional experiences than she does to their beliefs and desires. However, it is precisely in such contexts that others' emotional reactions might be an especially valuable cue to their mental states.

Collectively therefore, these findings leave open the question of the degree to which children can use their understanding of emotions to gain insight into others' minds. It may be difficult or impossible (in the absence of extensive prior knowledge) to recover the representations driving the changes in emotional expressions when these changes are unrelated to external, observable events (e.g., as when someone looks happy at one moment and sad the next simply because they first think about a joyous event and then an unhappy one). However, it may be possible to recover others' mental states when these changes are probabilistically associated with external events. Here we focus on agent's emotional expressions in response to testimony about, and observation of, an event. When an agent's mental states are otherwise undetermined by her actions and the context, can children compare her emotional reactions to an expected and observed outcome to jointly recover her desires and beliefs?

Previous work on inferring beliefs from emotions has found both successes and failures between ages four and six (e.g., Hadwin & Perner, 1991; Harris et al., 1989; Rieffe, et al., 2005; Wellman & Banerjee, 1991), thus we focus our investigation on the same age range. In contrast to most previous work, we focus 1) specifically on "backwards inferences", from emotional reactions to beliefs and desires; 2) on children's ability to recover both beliefs and desires simultaneously; 3) on inferences about ordinary (rather than unusual, surprising) events, and 4) on a cue to agents' mental states that has rarely been investigated: the dynamics of valenced facial expressions. We predict that children may be able to use an agent's emotional reactions to

anticipated and observed outcomes to recover both her beliefs and desires. Specifically, given the emotional reaction to the observed outcome, children may be able to infer the agent's desires; given the presence or absence of a change in valence between anticipated and observed outcomes, children may be able to infer the content of the agent's initial beliefs.

Finally, because the question of interest here is whether children can infer mental states from the dynamics of emotional reactions to expected and observed outcomes, we will elide a question that has been the focus of many previous investigations: how children draw inferences from emotional reactions to the emotions themselves (e.g., Ekman & Oster, 1979; Gross & Ballif, 1991; Izard, 1994; Widen & Russell, 2008, 2010). We take as a premise that, within a well-specified context and shared cultural knowledge, four- and five-year-olds can use prototypical happy and sad facial expressions to infer happiness and sadness (see e.g., Widen, 2013). Our question is whether children can use information in others' emotional responses to anticipated and observed outcomes to jointly infer their beliefs and desires.

3.3 EXPERIMENT 1

3.3.1 Method

Participants

Thirty-two children ($M = 5.0$ years; range: 4.1-5.9; 50% girls) were recruited from an urban children's museum between April and July 2014. While most of the children were white and middle class, a range of ethnicities and socioeconomic backgrounds reflecting the diversity of the local population (47% European American, 24% African American, 9% Asian, 17% Latino, 4% two or more races) and the museum population (29% of museum attendees receive free or discounted admission) were represented throughout.

To follow and throughout, we treat age as a continuous variable and then perform post-hoc analyses in age bins (four-year-olds and five-year-olds) in order to enable comparison with the previous literature on theory of mind which has largely treated age groups categorically (see Wellman, 2014 for review). To ensure a balanced distribution across ages, children were recruited in age bins consisting of 16 four-year-olds ($M = 4.4$ years; range: 4.1-4.8; 56% girls) and 16 five-year-olds ($M = 5.5$ years; range: 5.0-5.9; 44% girls).

Materials

Each child saw four illustrated stories (see Figure 1), two presenting Valence Stable conditions (Happy-Happy and Sad-Sad) and two presenting Valence Change conditions (Happy-Sad and Sad-Happy). Because we used canonical happy and sad faces and research suggests that by four, children can interpret these as such (see Widen, 2013 for review) expressions coded by the Facial Action Coding System (Ekman & Friesen, 1976) were not critical here; the facial expressions used here were from istock photos (<http://www.istockphoto.com/>). The mapping between stories and conditions, the order of conditions, and the expected-actual contents of the containers were counterbalanced across participants, resulting in a total of 16 storybooks. A different agent and a different bunny (indicated by its color) were used in each story.

Procedure

Children were tested individually; all sessions were videotaped. Children were asked check questions to encourage them to follow along. Incorrect responses were corrected throughout. (Collapsing data from all three experiments in the study, children's accuracies on the six check questions were .74, .73, .80, .99, .99, and 1.00 respectively. Incorrect responses consisted primarily of guessing: e.g., when asked "Do we know what is in the container?" or "Do we know what Sally wants to eat" children who answered incorrectly would say "an apple" or "a

banana” rather than say “no”. No children were excluded on the bases of incorrect responses however none of the results change if only children who answered all six check questions correctly are included.)

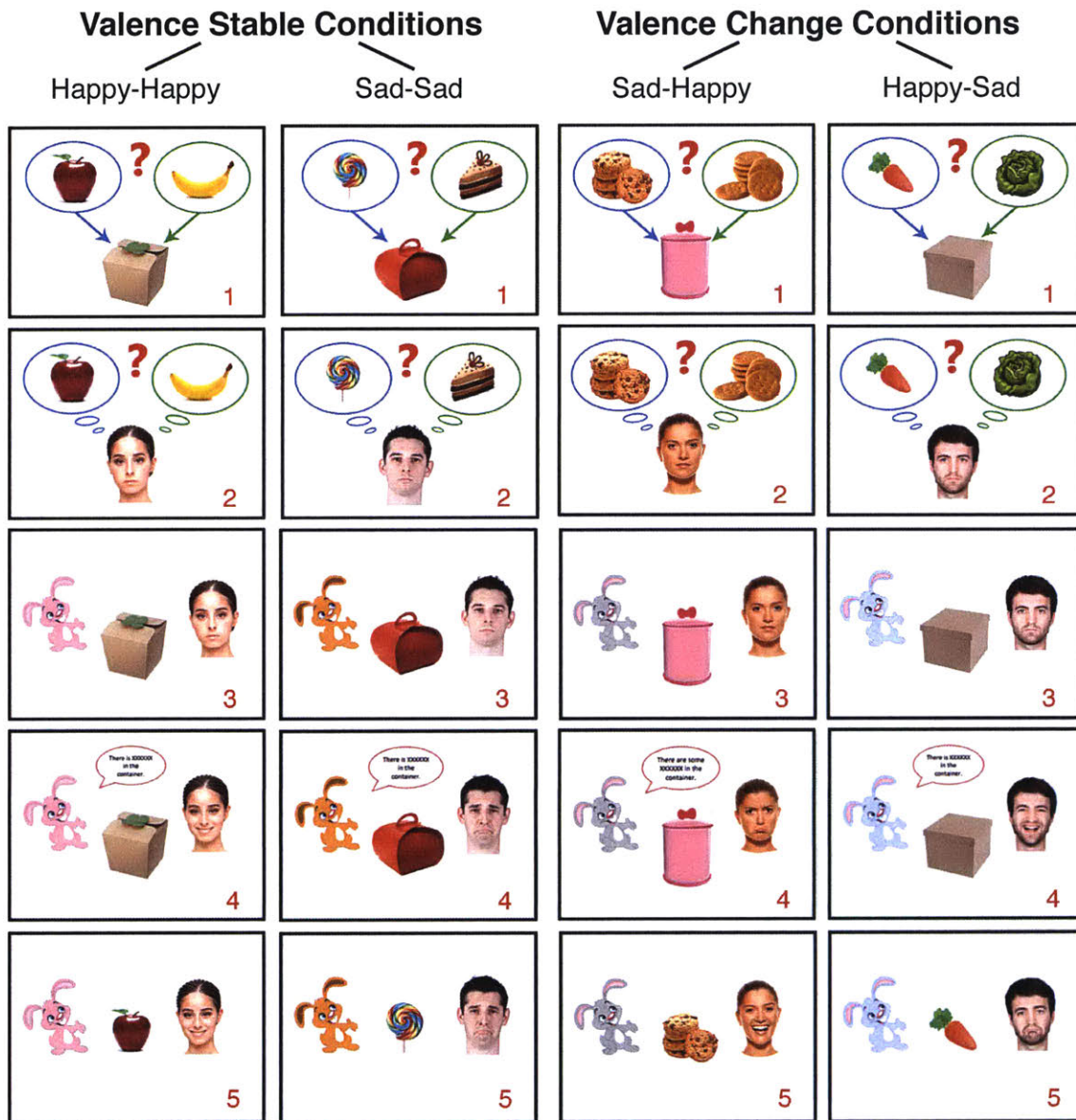


Figure 1 Examples of picture stimuli used in the Happy-Happy, Sad-Sad, Sad-Happy, and Happy-Sad conditions respectively. The mapping between stories and conditions, the order of conditions, and the expected-actual contents of the containers was counterbalanced across participants. See text for details.

Each story was read consecutively, as follows (using the apple-banana story as an example). The experimenter placed Picture 1 on the table and said, “This is a container. Sometimes there is an apple inside and sometimes there is a banana. But before we open it, we don’t know what’s inside.” She introduced Picture 2 and said, “This is Sally. She wants something to eat. She might want an apple, or she might want a banana; but she hasn’t told us.” Children were asked (Check questions 1 and 2): “Do we know what’s in the container?” and, “Do we know what Sally wants to eat?” The experimenter introduced Picture 3 and said, “This is the pink bunny. The bunny wants to help. So the bunny says: ‘Hi, Sally! Let me tell you what’s in the container!’ But, the bunny could be right or wrong.” Children were asked (Check question 3): “Is the bunny always right?” The experimenter introduced Picture 4, with the facial expression appropriate to the condition and said, “The bunny tells Sally what he thinks is in the container. But he doesn’t tell us. It’s a secret. After hearing the bunny’s secret, Sally’s response is this ...” Children were asked (Check question 4): “Is she happy or sad?” The experimenter introduced Picture 5 and said, “Then the bunny opens the container and takes out what’s inside.” Children were asked (Check question 5): “What’s inside?” Pointing to Picture 5, the experimenter said, “After seeing what’s actually in the container, Sally’s response is this ...” Children were asked (Check question 6): “Is she happy or sad?”

Finally, children were asked two test questions, in fixed order. The experimenter pointed to Picture 5 and asked (Desire question): “Based on this response, what did Sally want to eat today, an apple or a banana?” Then, she pointed to Picture 4 and asked (Belief question): “Before the container was opened, but after hearing the bunny’s secret, what did Sally think was inside: an apple or a banana?”

Coding

The first author coded all the responses to the two test questions offline from videotape. Seventy-five percent of these responses were recoded by an independent coder blind to hypotheses and conditions; there was 100% agreement on children's responses. Children's responses to the Desire question were coded as "Actual content" if they chose the content in the container and "Alternative" if they chose the other food. Children's responses to the Belief question were coded as "True belief" if they chose the content in the container, and "False belief" if they chose the other food. Two of 256 responses could not be classified (i.e., "both" and "she did not know"); these were coded as missing values. (Categorizing these as wrong responses instead of missing values does not change the results here, or in the following experiments.)

3.3.2 Results and discussion

As discussed (see Participants), we first analyze the effect of age as a continuous variable and follow-up with post-hoc analyses by age group to enable comparison with the previous literature. Generalized Estimating Equations were used for all analyses except as indicated. Generalized Estimating Equations are comparable to Repeated Measures ANOVA but unlike Repeated Measures ANOVA they are appropriate for categorical outcomes (as in the forced choice binary responses here) and robust to missing cells (see Coding). See Figure 2 for Results.

Desires: There was a main effect of age ($b = -3.07$, $SE = 1.07$; $\chi^2(1, N = 128) = 8.20$, $p = .004$) and of the valence of the second facial expression ($b = 33.49$, $SE = 4.88$; $\chi^2(1, N = 128) = 12913.87$, $p < .001$) on children's responses; the interaction was also significant ($b = 3.07$, $SE = 1.09$; $\chi^2(1, N = 128) = 7.92$, $p = .005$). Five-year-olds performed at ceiling; all children inferred that the agent wanted the food in the container when the final expression was happy (Happy-Happy, Sad-Happy) and the alternative when the expression was negative (Sad-Sad, Happy-Sad).

Four-year-olds performed at ceiling when the expression was positive but had slightly more difficulty when the expression was negative (Sad-Sad: 12/16, $p = .077$, 95% CI [.48, .93]; Happy-Sad: 13/16, $p = .021$, 95% CI [.54, .96]; comparisons to chance by binomial test throughout).

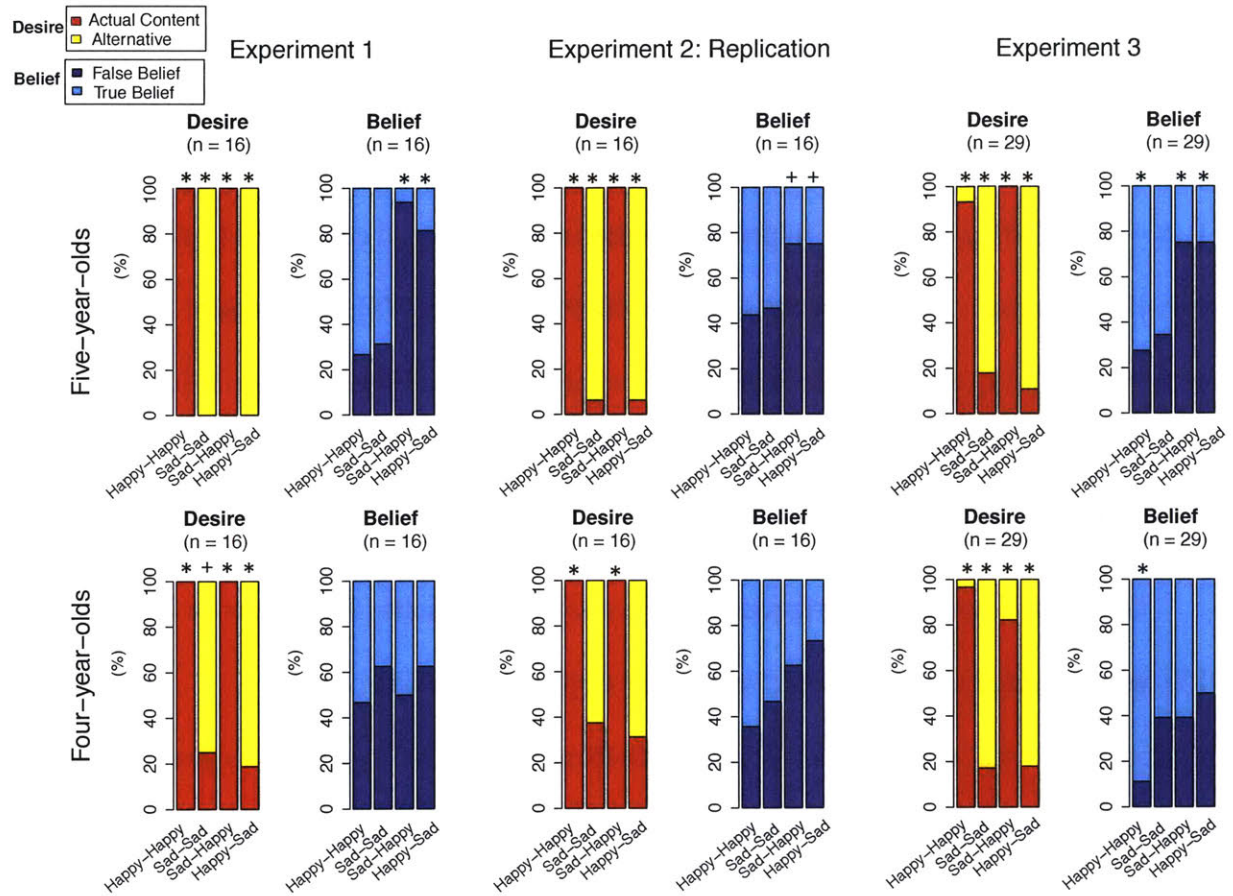


Figure 2 Five- and four-year-olds' inferences about desires and beliefs in Experiments 1-3. * $p < .05$; + $p < .10$.

Beliefs: There was no main effect of age ($b = .60$, $SE = .42$; $\chi^2(1, N = 126) = 1.09$, $p = .297$) but there was a main effect of the presence or absence of a valence change ($b = 9.67$, $SE = 3.62$; $\chi^2(1, N = 126) = 11.13$, $p < .001$) and an interaction ($b = -2.24$, $SE = .76$; $\chi^2(1, N = 126) = 8.82$, $p = .003$).

For five-year-olds, the presence or absence of a valence change affected their responses ($b = -2.84, SE = .67; \chi^2(1, N = 63) = 18.20, p < .001$). Five-year-olds inferred false beliefs in the Valence Change conditions (Sad-Happy: 15/16, $p < .001$, 95% CI [.70, 1.00]; Happy-Sad: 13/16, $p = .021$, 95% CI [.54, .96]) but did not infer true beliefs in the Valence Stable conditions (Happy-Happy: 11/15, $p = .119$, 95% CI [.45, .92]; Sad-Sad: 11/16, $p = .210$, 95% CI [.41, .89]).

By contrast four-year-olds had difficulty recovering the agent's beliefs in both conditions. The effect of the presence or absence of a valence change was not significant ($b = -.06, SE = .51; \chi^2(1, N = 63) = 0.01, p = .910$). Children performed at chance in all conditions (Valence Change, false belief responses: Sad-Happy: 8/16, $p = 1.000$, 95% CI [.25, .75]; Happy-Sad: 10/16, $p = .455$, 95% CI [.35, .85]; Valence Stable, true belief responses: Happy-Happy: 8/15, $p = 1.000$, 95% CI [.27, .79]; Sad-Sad: 6/16, $p = .454$, 95% CI [.15, .65]).

3.4 EXPERIMENT 2: REPLICATION

Given concern about the reproducibility of scientific results (Open Science Collaboration, 2015), we conducted a pre-registered replication (Open Science Framework: https://osf.io/e36vm/?view_only=643ddd21a06a49ca987105b1069b487a). The pre-registered analyses were the Generalized Estimated Equations by age group, predicting that the presence or absence of a valence change would affect belief inferences in five but not four-year-olds (https://osf.io/kx97a/?view_only=5590ef3ed01548b4b963ba83667fdd63). We also tested adults to clarify children's responses to the Valence Stable conditions.

3.4.1 Method

Participants

Thirty-two children ($M = 4.9$ years; range: 4.0-5.9; 41% girls) were recruited from the museum between August and September 2015. To ensure a balanced distribution across ages,

children were recruited in age bins consisting of 16 four-year-olds ($M = 4.4$ years; range: 4.0-4.9; 31% girls) and 16 five-year-olds ($M = 5.4$ years; range: 5.0-5.9; 50% girls).

Sixty-five adults were recruited on Amazon Mechanical Turk. Participation was restricted to individuals with IP addresses from the USA and with HIT approval rate of 95% or higher. Because Amazon Mechanical Turk workers are paid by the task, there is an incentive to rush through tasks. To ensure that online participants were attentive, check questions for the adults were used as exclusion criteria. Four participants were excluded for answered one or more check questions incorrectly. (See Materials and Procedure.)

Materials and procedure

The materials and procedure for the children were identical to the initial experiment. Adults were tested online using the same materials except that all materials were written, only three of the check questions were asked (“Is the bunny always right?” and “Is Sally happy or sad?” when Sally responded to the bunny’s secret and to the actual contents in the container).

Coding

Coding was identical to the initial experiment. Seventy-five percent of responses were recoded by an independent coder blind to hypotheses and conditions; there was 99% agreement on children’s responses. Four of 256 responses could not be classified (e.g., “lollipop and cake” and “I don’t know”); as in Experiment 1, these responses were coded as missing values.

3.4.2 Results and discussion

Desires: For children, there was a main effect of age ($b = -2.16$, $SE = .70$; $\chi^2(1, N = 128) = 9.70$, $p = .002$) and the valence of the final expression ($b = 36.85$, $SE = 3.28$; $\chi^2(1, N = 128) = 19652.07$, $p < .001$); the interaction was also significant ($b = 2.16$, $SE = .72$; $\chi^2(1, N = 128) = 8.93$, $p = .003$). Five-year-olds performed near ceiling (Happy-Happy: 16/16, $p < .001$, 95% CI

[.97, .1.00]; Sad-Happy: 16/16, $p < .001$, 95% CI [.97, .1.00]; Sad-Sad: 15/16, $p < .001$, 95% CI [.70, 1.00]; Happy-Sad: 15/16, $p < .001$, 95% CI [.70, 1.00]). Four-year-olds performed at ceiling when the final expression was positive but again had more difficulty when the expression was negative (Happy-Happy: 16/16, $p < .001$, 95% CI [.97, .1.00]; Sad-Happy: 16/16, $p < .001$, 95% CI [.97, .1.00]; Sad-Sad: 10/16, $p = .455$, 95% CI [.35, .85]; Happy-Sad: 11/16, $p = .210$, 95% CI [.41, .89]).

Beliefs: There was a main effect of the presence or absence of a valence change ($b = 1.81$, $SE = 3.22$; $\chi^2(1, N = 124) = 8.95$, $p < .003$). Neither the main effect of age ($b = .26$, $SE = .42$; $\chi^2(1, N = 124) = .01$, $p = .941$) nor an interaction between age and a valence change ($b = -.60$, $SE = .66$; $\chi^2(1, N = 124) = .84$, $p = .360$) was significant.

Again, five-year-olds' inferences were affected by the presence or absence of a valence change ($b = -1.29$, $SE = .55$; $\chi^2(1, N = 63) = 5.63$, $p = .018$). In the Valence Change conditions, children tended to infer false beliefs (Sad-Happy: 12/16, $p = .077$, 95% CI [.48, .93]; Happy-Sad: 12/16, $p = .077$, 95% CI [.48, .93]); children did not infer true beliefs in the Valence Stable conditions (Happy-Happy: 9/16, $p = .804$, 95% CI [.30, .80]; Sad-Sad: 8/15, $p = 1.000$, 95% CI [.27, .79]).

Also as predicted, the effect of valence change was not significant in four-year-olds although there was a trend on replication ($b = -1.00$, $SE = .53$; $\chi^2(1, N = 61) = 3.53$, $p = .060$). Four-year-olds did not perform above chance in any condition (Sad-Happy: 10/16, $p = .455$, 95% CI [.35, .85]; Happy-Sad: 11/16, $p = .210$, 95% CI [.41, .89]; Happy-Happy: 9/14, $p = .424$, [.35, .87]; Sad-Sad: 8/15, $p = 1.000$, 95% CI [.27, .79]). See Figure 2.

Aggregating the data from the initial experiment and replication suggests that children's chance performance on the belief inferences is unlikely to be due to the experiments being

underpowered: overall, there was a main effect of the presence or absence of a valence change in five-year-olds ($b = -1.99$, $SE = .41$; $\chi^2(1, N = 126) = 23.2$, $p < .001$) but not in four-year-olds ($b = -.51$, $SE = .36$; $\chi^2(1, N = 124) = 1.98$, $p = .160$). Five-year-olds in the Valence Change conditions inferred false beliefs (Sad-Happy: 27/32, $p < .001$, 95% CI [.67, .95]; Happy-Sad: 25/32, $p = .002$, 95% CI [.60, .91]) but chose at chance in the Valence Stable conditions (Happy-Happy: 20/31, $p = .150$, 95% CI [.45, .81]; Sad-Sad: 19/31, $p = .281$, 95% CI [.42, .78]); four-year-olds chose at chance in all conditions (Sad-Happy: 18/32, $p = .600$, 95% CI [.38, .74]; Happy-Sad: 21/32, $p = .110$, 95% CI [.47, .81]; Happy-Happy: 17/29, $p = .458$, 95% CI [.39, .77]; Sad-Sad: 14/31, $p = .720$, 95% CI [.27, .64]). Aggregating the data also allows us to see whether four-year-olds' failures to infer false beliefs in the Valence Change conditions might be due to task switching demands. However, those four-year-olds who saw a Valence Change story on the first trial were no more likely to succeed on the first trial than overall (10/17, $p = .629$, 95% CI [.33, .82]).

For comparison, adults recovered agents' desires near ceiling (Happy-Happy: 61/61, $p < .001$, 95% CI [.94, 1.00]; Sad-Happy: 60/61, $p < .001$, 95% CI [.91, 1.00]; Sad-Sad: 61/61, $p < .001$, 95% CI [.94, 1.00]; Happy-Sad: 61/61, $p < .001$, 95% CI [.94, 1.00]). Comparing the adults with the five-year-olds, there was no main effect of age ($b = 1.52$, $SE = .39$; $\chi^2(1, N = 370) = 2.9$, $p = .090$) but there was a main effect of the presence or absence of a valence change ($b = 2.73$, $SE = 1.51$; $\chi^2(1, N = 370) = 133.5$, $p < .001$) and an interaction ($b = -2.36$, $SE = .59$; $\chi^2(1, N = 370) = 15.8$, $p < .001$). Post-hoc analyses found that the presence or absence of a valence change affected adults' belief inferences ($b = -4.36$, $SE = .43$; $\chi^2(1, N = 244) = 105.00$, $p < .001$). In the Valence Change conditions, adults and five-year-olds both inferred false beliefs (111/122 adults versus 52/64 five-year-olds; $\chi^2(1, N = 186) = 2.83$, $p = .093$, 95% CI [-.02, .22], two-

sample test for equality of proportions with continuity correction; adults, Sad-Happy: 59/61, $p < .001$, 95% CI [.89, 1.00]; Happy-Sad: 52/61, $p < .001$, 95% CI [.74, .93]). However, unlike children, adults inferred true beliefs in the Valence Stable conditions (108/122 adults versus 39/62 five-year-olds; $\chi^2(1, N = 184) = 15.20$, $p < .001$, 95% CI [.11, .40]; adults, Happy-Happy: 60/61, $p < .001$, 95% CI [.91, 1.00]; Sad-Sad: 48/61, $p < .001$, 95% CI [.66, .88]).

3.5 EXPERIMENT 3

There are two ways in which Experiments 1 and 2 may have underestimated the children. First, although we said that the agent (e.g., Sally) might want an apple or might want a banana we failed to specify that the two options were mutually exclusive, thus some of the children may have decided that she liked or disliked both. Second, because the questions were always asked in a fixed order—desire first and belief second—some of the children, and especially the youngest ones, may have had difficulty answering the second question. If so four-year-olds' apparent failure to infer beliefs may have been due to an overall degradation in their attention rather than a specific difficulty with belief questions. In Experiment 3, we address both these concerns by clarifying that the options are mutually exclusive and by always asking the belief question first. (We did this rather than counterbalance order both because the belief question was of primary interest given that other studies have shown that children at this age can infer agents' desires from emotional expressions and because the belief question is presumably harder given children's near ceiling performance on the desire question.) Finally, the previous results enabled us to estimate the effect size so we ran a power analysis to ensure that we had sufficient power to detect above chance performance (if present). The power analysis indicated that for a power of 0.80, a two-tailed alpha less than .05, and an effect size of Cohen's $h = 0.52$ (a conservative effect size calculated based on the replication data in the Valence Change conditions), we should

test 29 four-year-olds and 29 five-year-olds. We pre-registered this experiment on Open Science Framework (https://osf.io/nduwg/?view_only=5a9590c9183e4a01b1a4f3bcb45d6639).

3.5.1 Method

Participants

Fifty-eight children ($M = 5.0$ years; range: 4.0-5.9; 50% girls) were recruited from the museum between March and April 2016. To ensure a balanced distribution across ages, children were recruited in age bins consisting of 29 four-year-olds ($M = 4.5$ years; range: 4.0-4.9; 55% girls) and 29 five-year-olds ($M = 5.5$ years; range: 5.0-5.9; 45% girls).

Materials and procedure

The materials and procedure for the children were identical to the initial experiment with two exceptions. First, when the experimenter introduced the agent's possible desires, she specified that the two candidate desires were mutually exclusive. Using the apple-banana story as an example, when the experimenter placed Picture 2 on the table, she said, "This is Sally. She *either* likes apples *or* she likes bananas but she *doesn't* like both. Today she wants something to eat. So she might want an apple, or she might want a banana but we don't know *which one* she wants." Second, throughout we asked the belief questions first and the desire questions second.

Coding

Coding was identical to the initial experiment. Seventy-five percent of responses were recoded by an independent coder blind to hypotheses and conditions; there was 99% agreement on children's responses. Eight of 464 responses could not be classified (e.g., "She doesn't know" and "I don't know"); as in previous experiments, these responses were coded as missing values. Additionally, two responses were dropped because of sibling interference and two were dropped because of experimental error.

3.5.2 Results and discussion

Desires: Although the desire question was asked second in this experiment, both five- and four-year-olds had no difficulty recovering the agents' desires. There was a main effect of the valence of the final expression ($b = -4.90, SE = 3.59; \chi^2(1, N = 227) = 89.91, p < .001$). The main effect of age was not significant ($b = -.61, SE = .45; \chi^2(1, N = 227) = .00, p = .947$), and although the interaction between age and the valence of the final expression was significant ($b = 1.90, SE = .76; \chi^2(1, N = 227) = 6.26, p = .012$) both five and four-year-olds performed near ceiling in each of the four conditions (five-year-olds: Happy-Happy: 27/29, $p < .001$, 95% CI [.77, .99]; Sad-Happy: 29/29, $p < .001$, 95% CI [.88, 1.00]; Sad-Sad: 23/28, $p < .001$, 95% CI [.63, .94]; Happy-Sad: 24/27, $p < .001$, 95% CI [.71, .98]; four-year-olds: Happy-Happy: 28/29, $p < .001$, 95% CI [.82, 1.00]; Sad-Happy: 23/28, $p < .001$, 95% CI [.63, .94]; Sad-Sad: 24/29, $p < .001$, 95% CI [.64, .94]; Happy-Sad: 23/28, $p < .001$, 95% CI [.63, .94]).

Beliefs: Overall we replicated children's performance. Specifically, there was a main effect of age ($b = -.23, SE = .32; \chi^2(1, N = 225) = 14.06, p < .001$) and the presence or absence of a valence change ($b = 5.22, SE = 2.43; \chi^2(1, N = 225) = 21.80, p < .001$) on children's belief inferences; the interaction was also significant ($b = -1.34, SE = .49; \chi^2(1, N = 225) = 7.43, p = .006$).

Further analyses showed that five-year-olds' inferences were affected by the presence or absence of a valence change ($b = -1.90, SE = .42; \chi^2(1, N = 114) = 20.50, p < .001$). In the Valence Change conditions, they inferred false beliefs (Sad-Happy: 21/28, $p = .013$, 95% CI [.55, .89]; Happy-Sad: 21/28, $p = .013$, 95% CI [.55, .89]); in the Valence Stable conditions, children inferred true beliefs in Happy-Happy condition but not in the Sad-Sad condition (Happy-Happy: 21/29, $p = .024$, 95% CI [.53, .87]; Sad-Sad: 19/29, $p = .136$, 95% CI [.46, .82]).

The effect of the presence or absence of a valence change was also significant in four-year-olds ($b = -.86$, $SE = .41$; $\chi^2(1, N = 111) = 4.39$, $p = .036$) but it was driven by children's unpredicted success in Happy-Happy condition. In the Valence Change conditions, four-year-olds did not infer false beliefs (Sad-Happy: 11/28, $p = .345$, 95% CI [.22, .59]; Happy-Sad: 14/28, $p = 1.000$, 95% CI [.31, .69]); in the Valence Stable conditions, they inferred true beliefs in the Happy-Happy condition (24/27, $p < .001$, 95% CI [.71, .98]) but not in the Sad-Sad condition (17/28, $p = .345$, 95% CI [.41, .79]). See Figure 2.

In general, the results of Experiment 3 replicate the previous studies, suggesting that children's ability to recover both beliefs and desires from changes in the valence of agents' emotional reactions develops between four and five. The success of both four- and five-year-olds at recovering the agent's beliefs in the condition when she was happy both in anticipating and observing the results was unexpected, and inconsistent with the results of the previous studies. Note however, that in this condition, children could succeed simply by reporting the item actually observed in the container throughout. It is interesting that children made this response only in Experiment 3 and not in the previous studies, however, four-year-olds' failure to recover beliefs in all of the remaining conditions suggest that their isolated success in this condition is unlikely to indicate a genuine ability to recover beliefs from the dynamics of emotional expressions. Similarly, five-year-olds' chance performance in the Sad-Sad condition suggests that, at best, their ability to recover beliefs from stable emotional expressions is fragile.

3.6 GENERAL DISCUSSION

These results suggest that by age five, children can use changes in the valence of an agent's emotional reaction to recover both her beliefs and desires in contexts where both are unknown and the agent's actions are not differentially informative. Four-year-olds used the

emotional reactions to recover the agent's desires but did not use the valence change to infer the agent's beliefs. Moreover, neither age group reliably treated a stable valence as informative about the agent's beliefs. When someone looked happy or sad about both an expected or observed outcome, adults inferred that she expected the outcome. By contrast, children gave inconsistent responses across studies to happy expressions and consistently chose at chance in response to sad expressions.

To our knowledge, this is the first study looking at how the dynamics of facial expressions inform children's theory of mind. Previous research suggests that children selectively invoke agent's beliefs in response to surprised, curious, or frightened responses to unusual or mysterious events (Rieffe et al., 2005; Wellman & Banerjee, 1991). Here however, and in contrast to studies where the emotional stakes have been relatively high (at least from a child's perspective – lost bunnies, dead turtles, growling dogs, etc. Lagattuta, et al., 1997; Pons et al., 2004; Widen & Russell, 2010) five-year-olds inferred agents' beliefs given only happy and sad expressions and entirely ordinary events (e.g., finding fruit in a container). These results suggest that children treat changes in emotional valence as informative even when the emotions themselves are of relatively little import.

In some respects, the current task resembles the classic unexpected contents tasks (Hogrefe, Wimmer & Perner, 1986; Perner, Leekam, & Wimmer, 1987) in that children have to reassess an initial belief about the contents of a container based on subsequent evidence. However, in the unexpected contents task, the initial belief is explicitly cued by the container itself (e.g., a "Smarties" container); the question is whether children continue to access this belief when it is subsequently proven false. By contrast, in the current task, the container provides no cues to its contents: children must simultaneously infer both the content and epistemic status of

the agent's initial belief. Thus the current study is not merely an affective version of an unexpected contents task. Rather, it tests children's ability to use emotional expressions to infer the content of mental states that are otherwise under-determined by the agent's actions and the context.

Like classic false belief tasks however, this study arguably makes high demands on information processing abilities distinct from theory of mind. (See Baillargeon, Scott, & He, 2010; Perner & Lang, 1999; Sodian, 2011; Wellman, 2014 for diverse perspectives and reviews.) In light of this, five-year-olds' successes in the valence change conditions may be particularly convincing, and four-year-olds' failure less so. In particular, four-year-olds were more successful in recovering the agent's desires than beliefs. This might be because children could infer the agent's desires in our task by attending only to a single emotional reaction (the agent's response to the observed outcome) whereas inferring beliefs required children to compare two emotional reactions (the agent's response to both the anticipated and observed outcome). Although we cannot rule out the possibility that the younger children's failure may be due to performance deficits, our results are consistent with a number of studies suggesting that children represent desires earlier and more robustly than beliefs. (See Wellman & Liu, 2004 for meta-analysis.) One possibility is that the processing demands are bound up with conceptual development insofar as the relative subtlety of emotional cues to beliefs versus desires may contribute to desires being represented earlier and more robustly than beliefs.

Additionally, our results are consistent with a large body of work suggesting that the integration of emotion understanding and belief-desire psychology undergoes substantial development between four and six (e.g., Bradmetz & Schneider, 1999; de Rosnay, et al., 2004; Hadwin & Perner, 1991; Harris, 2008; Harris, et al., 1989; Lagattuta & Wellman, 2001; Pons, et

al., 2004; Ruffman & Keenan, 1996; Wellman & Bartsch, 1988; Wellman & Liu, 2004; Widen & Russell, 2008). In particular, a number of studies suggest that children's ability to link mental representations of past events to current emotions improves over the preschool years. Thus for instance, although four-year-olds understand that someone's emotional reaction to an event may wane over time, not until six do children understand that this depends on whether the person remembers or forgets the event (Harris, et al., 1985). Similarly, children becoming increasingly able to understand what kinds of cues about past events might trigger memories that could affect someone's emotions in the present (Lagatutta, et al., 1997). In the current study, children did not have to link a mental representation of a past event to a current emotion; rather they had to use both current and past emotional expressions to infer past mental representations (the character's earlier beliefs). However, improvements in children's overall ability to represent the causal relations between mental representations and emotion over time might contribute to the developmental change in children's performance between four and six on this task.

Given that five-year-olds recovered the agent's false beliefs from the changing valence, why, unlike adults, did they fail to attribute true beliefs consistently when the agent had the same reaction to expecting and observing an outcome? Some work suggests that children's ability to understand subtle aspects of emotion (e.g., discrepancies between true and expressed emotion, or mixed, ambivalent emotions) undergoes protracted development (Pons, et al., 2004). A failure to change expressions between an observed and expected outcome is *prima facie* a very subtle cue to agent's mental states: children's ability to draw inferences from such subtle cues might continue to develop through middle childhood. A related possibility is that, the stable emotional expressions may have been less salient than the changing ones; thus children might have attended less to the emotional reactions overall in the same valence conditions. Additionally, we

note that some studies suggest that when children begin to pass explicit false belief tasks they over-attribute false beliefs when they should not (see Hedger & Fabricius, 2011 for a review). The current results might reflect an overall increase in children's willingness to infer false beliefs rather than true ones.

A final possibility is that five-year-olds might have been more likely than adults to allow for the possibility that the agent liked, or disliked, both outcomes equally (even when, as in Experiment 3, the experimenter specified that the preferences were mutually exclusive). More than adults, children may have resisted learning the relatively arbitrary rule that the agent liked only one of the two items in each category given real world experience that children who like (or dislike) fruit, desserts, snacks, or vegetables tend to apply that preference to the category as a whole. However, if the agent did like or dislike both items equally, her emotional response would be uninformative because it would be identical whether her expectations were fulfilled or violated. To the degree that children assume the agent had no preference between outcomes, they might reasonably treat the agent's stable valence as uninformative about her beliefs. In this respect, the task of jointly inferring an agent's beliefs and desires may be more difficult than inferring someone's beliefs when her desires are fully specified.

Finally, our study assumes that the agent in each story trusted the Bunny's testimony and thus formed a belief about what was in the container and emotionally responded to this belief. This is consistent with other studies suggesting a strong default assumption that testimony should be trusted (e.g., Jaswal, Croft, Setia & Cole, 2010). However, across trials, bunnies were unreliable agents -- half the bunnies were correct about the contents and half were not -- and previous research suggests that four- and five-year-olds track the reliability of informants (see e.g., Corriveau & Harris, 2009; Jaswal & Neely, 2006; Koenig, Clement, & Harris, 2004). In this

study, children did not have sufficient information to form reliability judgments given that the color of the bunny changed on each trial (suggesting that there was a different informant each time), and no individual agent had prior knowledge about that bunny's reliability. Nonetheless, we cannot rule out the possibility that children might have inferred that the bunny's testimony led the actor to represent (and emotionally react to) the contents the bunny identified without specifically developing an expectation that those contents were in the container. That is, children might have inferred that the agent formed a mental representation consisting of a "thought" about the contents rather than a "belief" about them and answered the belief question on this basis. If true, this changes our account of the results only modestly: rather than suggesting that children used the changing valence between the expected and observed outcome to infer the agent's beliefs and desires it would suggest that children used the changing valence between the reported and observed outcome to infer the agents' thoughts and desires.

The mental state inferences here were challenging insofar as the input was impoverished: the character showed only two expressions and did not engage in any goal-directed actions. However, the hypothesis space here was restricted to two alternatives, children had continuous access to the agent's emotional reaction to both the expected and actual outcomes, and the emotional reactions were highlighted. In the real world, emotional reactions are transient and typically go unremarked. Future research might investigate children's ability to recover agent's beliefs and desires in contexts where the emotional reactions unfold in time, and where both the hypotheses and emotional reactions are more complex than those used here. Future research might also look at older children to see when children's performance converges with adults'.

The current results however, suggest that by age five, children's intuitive theory of mind begins to support mental state inferences from others' emotional reactions. Extending previous

work, children not only understand that thinking affects feeling (e.g., Harris, et al., 1985; Lagattuta et al., 1997; Lagattuta & Wellman, 2001; Taylor & Harris, 1983), they can use others' feelings to infer otherwise unknown thoughts. When children see someone's face change from sadness to happiness, or from happiness to sadness, they gain insight not only into how the person feels, but what she wants and believes about the world.

Chapter 4 Study 2.2 Inferring Beliefs and Desires From Emotional Expressions: A Computational Model

This chapter is based on Wu, Baker, Tenenbaum, & Schulz. (2017). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*.

4.1 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 ($r_s \geq .95$), 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

4.2 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 et al., 2017; Baker, Saxe & Tenenbaum, 2009; 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 have 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' 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 agent's goals and their emotions within the first year of life. Thus for instance, eight-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 two, 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 four and six 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, four- and five-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 other's 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 fourteen-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, fourteen 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 eighteen-months, toddlers succeed (Repacholi & Gopnik, 1997). By two, children can use an agent's emotional reaction to say explicitly whether she is looking at something she does or does not want (Wellman, Philips & Rodriguez, 2000).

Such inferences refer to others' desires; inferences about others' beliefs undergo more protracted development. Even children as old as six rarely refer to others' beliefs in explaining their emotional reactions (Rieffe, Terwogt & Cowan, 2005). The exceptions are that four and five-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 & Flavell, 1997; Lagattuta & Wellman, 2001; Taylor & Harris, 1983). Thus for instance, between four and six, children expect people to feel more intensely about recent events than past ones, and 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, four and five-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 five-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, 1990; 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, 1990; Scherer & Meuleman, 2013; Zaki, Bolger & Oschner, 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 emotion reaction to events (see e.g., Skerry & Saxe, 2015; Scherer, Schorr & Johnstone, 2001; Scherer & Meuleman, 2013). 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 & Fernandez-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., Darwin, 1872/1965; Ekman & Friesen, 1971; Lee & Anderson, 2016; Matsumoto & Willingham, 2009; Elfenbein, Beaupré, Lévesque & Hess, 2007; Meeren, van Heijnsbergen & de Gelder, 2005; Carroll & Russell, 1996). 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.

4.3 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; Baker et al., 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, et al., 2015; Shafto et al., 2012; Shafto 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 (*Reaction₀*) 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 (*Reaction₁*) 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 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), *Reaction₀* and all arrows connected with it drop out.

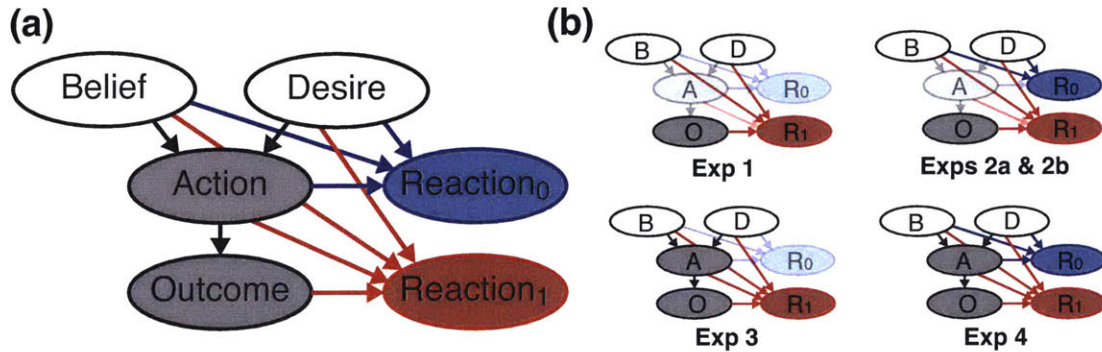


Fig. 1 (a) Template for Bayesian network models of people's intuitive theory of emotional responses and its integration with theory of mind. Arrows indicate hypothesized causal relationships between mental states, actions, outcomes, and emotional reactions. This generative model starts with people's 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*. *Reaction₀* 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). *Reaction₁* 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 sub-networks 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, *Reaction₀* (as well as arrows directly connected with it) drops out (Experiments 1 and 3).

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 (*R₁*) 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). Experiment 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 Equation 1. We compute

the normalized products of the forward distributions according to Equation 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 Equation 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 (Ong et al., 2015), 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. 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. The current study focuses on how we might use emotional reactions, even to identical events, to recover these distinct combinations of beliefs and desires.

4.4 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 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 2b; and photographs versus movies in Experiments 3 and 3 Supplementary.

4.4.1 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.

4.4.1.1 Method

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 SI Text 1.1 for the complete scenario.)

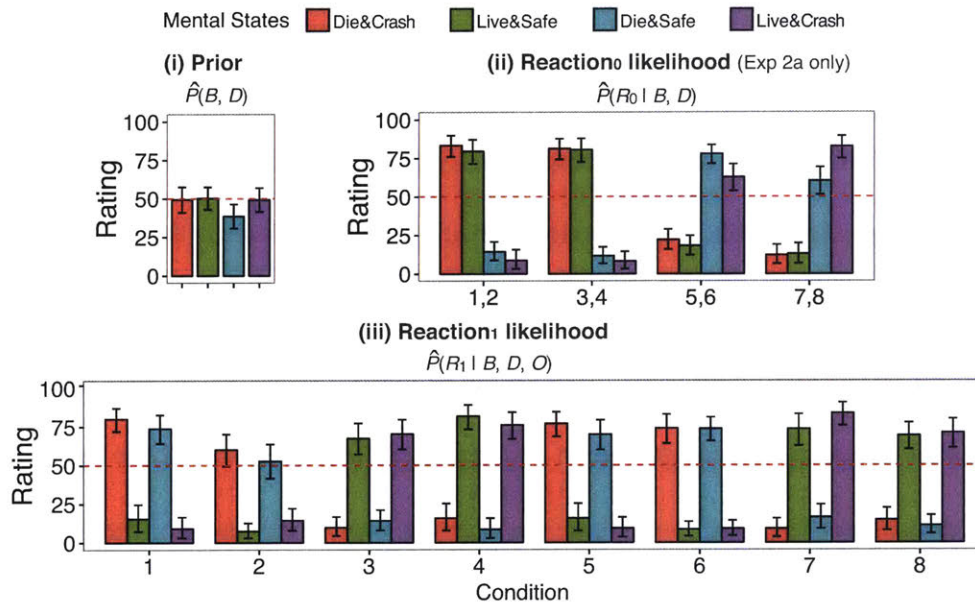
The eight possible combinations of Grace's belief, desire, and the outcome yield Conditions 1-8 of the experiment. See Fig. 2(a). To generate Grace's emotional reaction in each condition, we used a facial morphing software to create photograph stimuli. Consistent with the developmental literature (e.g., MacLaren & Olson, 1993; Hadwin & Perner, 1991; Repacholi & Gopnik, 1997; Skerry & Spelke, 2014; Stein & Levine, 1989; Wellman & Banerjee, 1991; Wellman & Woolley, 1990; Yuill, 1984)³, 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 SI Text 2.1.1 and Table S1 for more details.

³ 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).

(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)



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

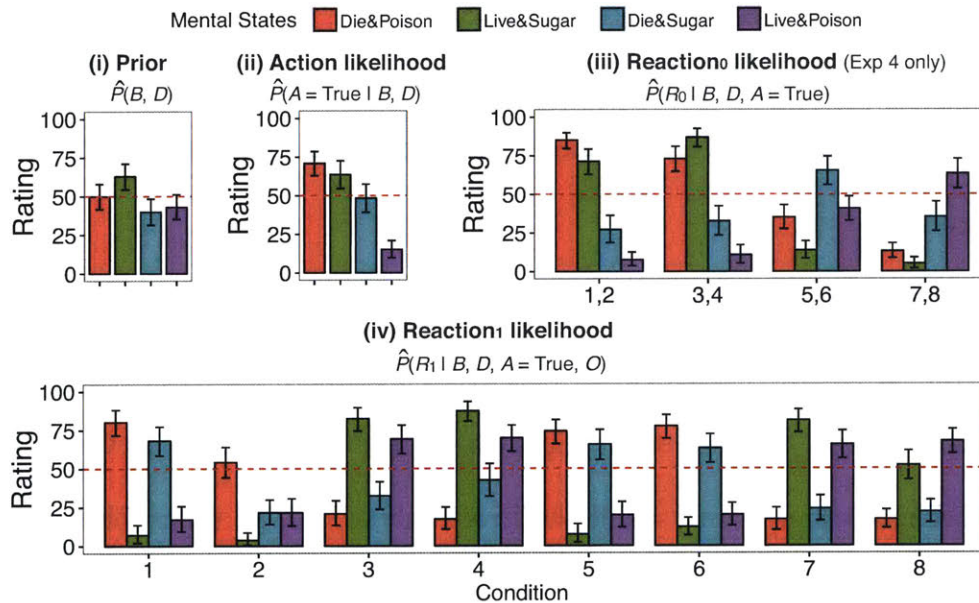


Fig. 2 (a) Design of Experiments 1, 2a, 3 and 4. The beliefs *Crash* and *Safe* refer to the plane-crash scenario while *Poison* and *Sugar* refer to the chemical-factory scenario; (b) Given the plane-crash scenario, participants' model calibration judgments on an un-normalized 0-100 scale for (i) the 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.)

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 pre-set the sample size for each group of participants at $n = 60$, sufficient for 97% power assuming a medium effect size (Cohen's $d = .50$). 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 SI 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: (1) Grace wants John to die and believes John was on the flight that crashed (Die&Crash), (2) Grace wants John to live and believes John was on a safe flight (Live&Safe), (3) Grace wants John to die and

believes John was on a safe flight (Die&Safe), and (4) 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 Equation 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 SI Text 3 for details.

4.4.1.2 Results and discussion

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 SI 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 first emotional expression. 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 SI Text 4.4 for detailed analyses.)

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 ($F(7, 1561) = .18, p = .989$) 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 8 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 $ps > .953$). 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.

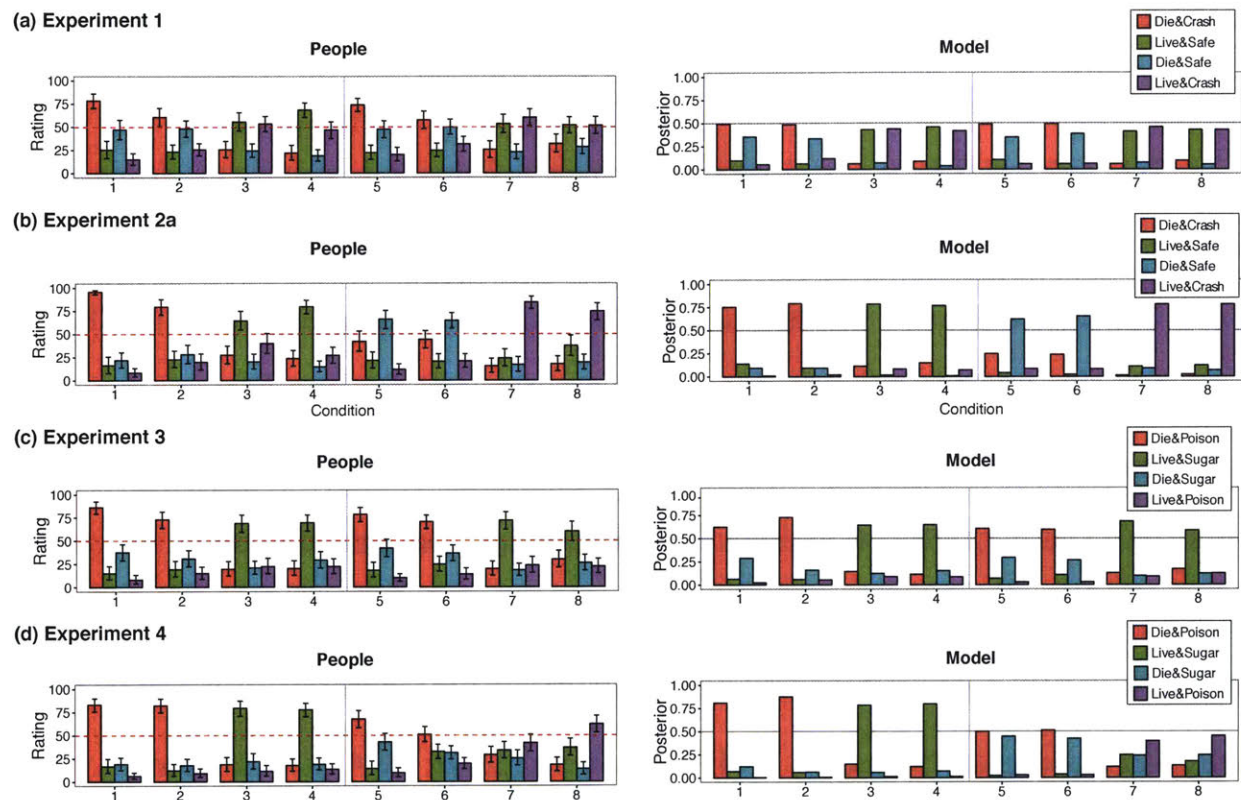


Fig. 3 People's mental state inferences on an un-normalized 0-100 scale and model predictions in Experiments 1, 2a, 3 and 4. Error bars indicate 95% confidence intervals.

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 $ps = 1.000$; mental states with the incorrect desire were rated significantly below 50: all $ts < -3.67$, all $ps < .018$).

The model predictions were generated according to Equation 1 (omitting the *Action* and *Reaction*₀ term; see SI Text 5.1), using the independent raters' 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, and 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.

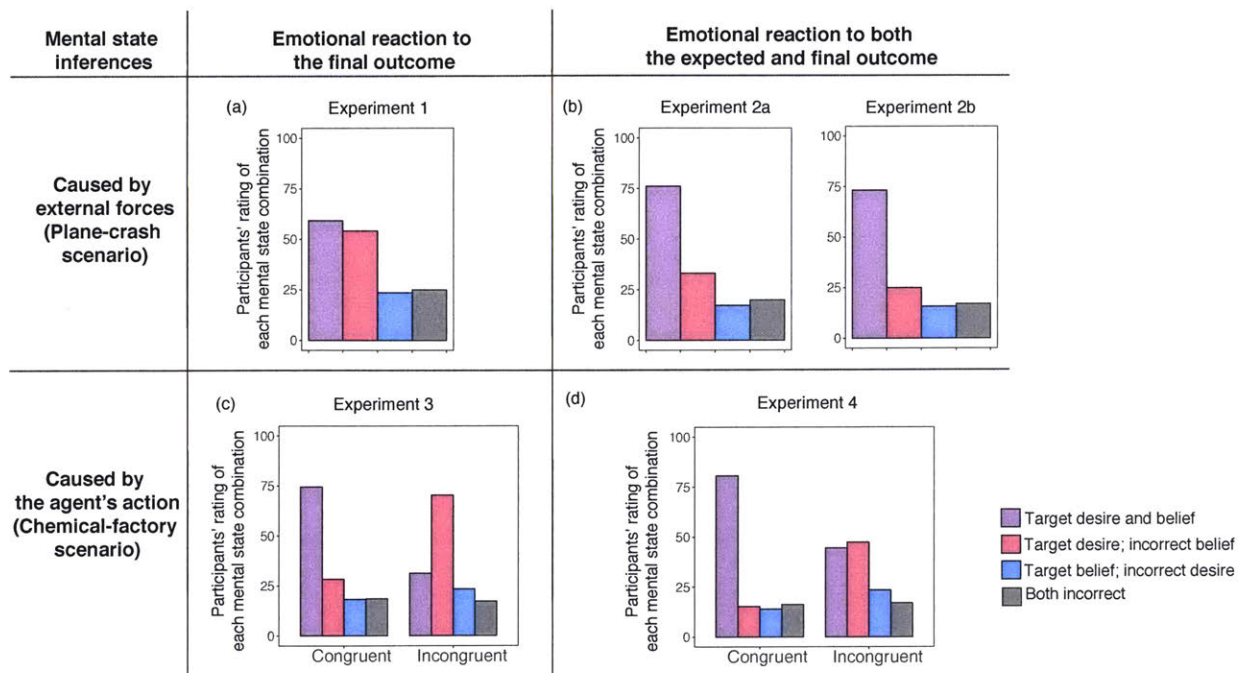


Fig. 4 Participants' mental state inferences averaged across conditions. In each plot, the first bar (purple) indicates the average rating of the target combination of desires and beliefs used to generate the facial expressions. The following three bars indicate the average ratings of each of the three non-target combinations. The pink bar indicates the target desire but incorrect belief; the blue bar indicates the target belief but incorrect desire; the grey 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.

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

4.4.2 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).

4.4.2.1 Method

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 SI Text 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 SI Text 3.

4.4.2.2 Results and discussion

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 SI Text 4.3 for detailed analyses).

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 Equation 1, with $Reaction_0$ and $Reaction_1$ terms but no $Action$ term; see SI 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. 3(b)); 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.

4.4.3 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 Supplementary) 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 (SI 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, 1990). 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 replicate 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.

4.4.3.1 Method

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 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.

Participants and procedures

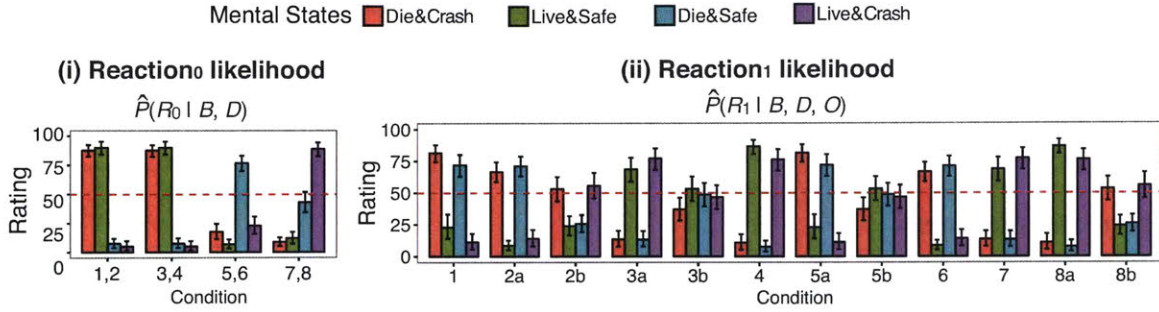
To calibrate the model, we measured people's judgments on the likelihood of the new set of stimuli. Participants ($n = 58$) rated each of the four facial expressions responding to the expected outcome ($Reaction_0$) given Grace's belief and desire, $P(R_0|B,D)$. A separate set of participants ($n = 58$)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

Experiment 2b

(a) Design

Desire&Belief	Die&Crash			Live&Safe			Die&Safe			Live&Crash		
Reaction ₀												
Outcome	Die	Live	Live	Die	Die	Live	Die	Die	Live	Die	Live	Live
Reaction ₁												
Condition	1	2a	2b	3a	3b	4	5a	5b	6	7	8a	8b

(b) Likelihoods of facial reactions



(c) Mental state inferences

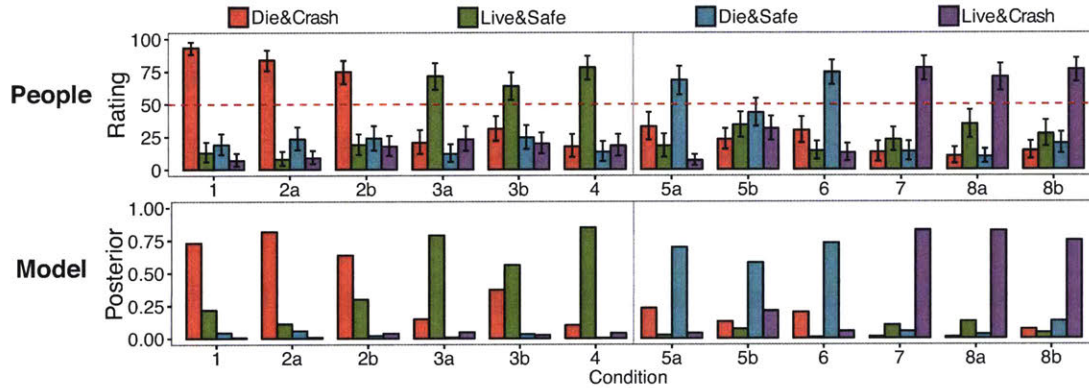


Fig. 5 Design and results of Experiment 2b. (a) Design. (b) Participants' model calibration judgments on an un-normalized 0-100 scale for the conditional likelihoods of (i) *Reaction*₀ and (ii) *Reaction*₁. (c) Participants' mental state inferences on an un-normalized 0-100 scale and model predictions.

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, $P(B, D | O, R_0, R_1)$.

4.4.3.2 Results and discussion

Model calibration

For *Reaction*₀ and the valenced *Reaction*₁, the estimated likelihoods were similar to those found in Experiments 1 and 2a (Fig. 4(b), SI Texts 4.3 and 4.4). 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 SI Texts 4.3 and 4.4 for detailed analyses.)

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 ($F(3, 2490) = 498.35, p < .001$) 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 $ts < -3.83$, all $ps < .016$). 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 $ps > .141$; mental states Die&Crash and Live&Crash were rated significantly below 50: both $ts < -3.91$, both $ps < .013$).

These behavioral responses were also predicted by our model. Model predictions were generated according to Equation 1, with *Reaction*₀ and *Reaction*₁ terms but no *Action* term; see SI 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 other's desires from their emotional reaction to events, but 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). The current study additionally highlights the role of a presence or absence of a change of 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., 2017; Baker et al., 2009) without considering emotions, or has looked at the relationship between 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.

4.4.4 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.

4.4.4.1 Method

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 also two possible outcomes: John either lives or dies after drinking the coffee. (See SI 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=b3cb225cdbdc498caa900e7431322fda) 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 SI Text 2.2); we refer to this as Experiment 3 Supplementary.

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.

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 SI Text 3 for details.

4.4.4.2 Results and discussion

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 SI 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) (Fig. 2(c)(ii)). (See SI Text 4.1-4.2 for details.) 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 SI Text 4.4.

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 Equation 1 (but omitting the $Reaction_0$ term, see SI 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 Supplementary). The behavioral results replicated those from the photograph stimuli in all respects (see SI 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 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.

4.4.5 Experiment 4

Experiment 4 is identical to Experiment 3 except that (as in Experiments 2a and 2b) 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.

4.4.5.1 Method

Design and materials

We used the same chemical-factory scenario as in Experiment 3 and the same photograph stimuli used in Experiment 2a.

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 re-used 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 SI Text 3.

4.4.5.2 Results and discussion

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 SI Text 4.3 for detailed analyses).

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 reactions) higher than the other three mental states 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 SI 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 agent's emotional reactions and their actions.

4.5 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.

4.5.1 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, 2b, 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.

4.5.2 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).

4.5.3 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 SI Texts 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 $r = 0.957$ (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.

4.6 GENERAL DISCUSSION

The current results suggest both the sophistication and limitations of people's ability to recover mental states from observed emotional reactions. On the 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 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 the current 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 combining emotions and basic emotions (Experiment 2a vs. 2b) or photographs and movies as cues to mental states (Experiment 3 vs. Experiment 3 Supplementary). Intuitively, richer sources of information about 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 Equation 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 (Ernst & Banks, 2002; Körding & Wolpert, 2004; Weiss, Simoncelli & Adelson, 2002; Battaglia & Schrater, 2007; Beierholm, Quartz & Shams, 2009). The principles of Bayesian inference have also been proposed as a potential unifying framework for cue integration in social cognition (Zaki, 2013; Wolpert, Doya & Kawato, 2003). 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 Equation 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.

Chapter 5 Inferring Recursive Mental States from Emotional Expressions

This chapter is based on Wu & Schulz. (in revision). Understanding social display rules: Using one person's emotional expressions to infer the desires of another.

5.1 ABSTRACT

This study investigates children's ability to use social display rules to infer others' desires. Children ($N = 211$; $M = 8.3$ years) saw a protagonist express one emotional reaction to an event in front of her social partner (Social Context), and a different expression behind her partner's back (Nonsocial Context). Although the protagonist expressed two contradictory emotions (and the social partner expressed no emotions at all), children successfully inferred both the protagonist's and social partner's desires (Experiments 1-2). Experiments 3-5 ruled out alternative explanations of the results. These results suggest that children can use changing emotional expressions between social and non-social contexts to recover not only the desire of the person displaying the emotions but also that of her audience.

Keywords: emotional expression; social display rules; theory of mind

5.2 INTRODUCTION

Young children can use emotional expressions to draw inferences about both external events in the world (e.g., Berman, Chambers, & Graham, 2010; Feinman, Roberts, Hsieh, Sawyer, & Swanson, 1992; Wu, Muentener, & Schulz, 2017), and others' internal mental states (e.g., Repacholi & Gopnik, 1997; Rieffe, Terwogt, & Cowan, 2005; Wellman, Phillips & Rodriguez, 2000; Wu & Schulz, 2017). However, because people sometimes go to great lengths to disguise their true feelings, emotional expressions can be misleading. Someone who is speaking in front of a large audience may pretend to be calm, even if she is nervous. A polite child who receives an undesirable gift may pretend to be happy even if she is disappointed. As we will review, a relatively large body of research has looked at children's understanding of real and apparent emotional expressions in the context of social display rules. Here however, we consider a feature of social display rules that has been largely overlooked in prior work: masked emotions may reveal as much about an individual's intended audience as they conceal about the individual herself. For instance, when someone congratulates a friend in public but fumes in private, we learn not only that this person's true feelings about the event are negative, but also that her friend's feelings are probably positive. Thus, given evidence about someone's feelings in both social and non-social contexts, an observer might therefore recover information both about the individual's mental states, and that of the society she keeps.

This kind of inference is non-trivial: it requires tracking someone's emotional expressions across social and non-social contexts and reasoning recursively about what the protagonist thinks her social partner thinks. To our knowledge, despite abundant work on emotion understanding and theory of mind in early childhood (see Wellman, 2014 for review), no one has yet looked at whether children can use emotional expressions in social and non-social contexts to infer not

only the mental states of the person expressing the emotions but also of their intended audience. That is our goal here.

First however, as noted, considerable work has looked at children's understanding of social display rules. Much of this work has focused on children's ability to regulate their own behavior appropriately. Such studies have found that young children can often modulate both their verbal and nonverbal responses in social contexts for the sake of politeness and to protect others' feelings (Cole, 1986; Saarni, 1984; Talwar & Lee, 2002; Talwar, Murphy, & Lee, 2007; Xu, Bao, Fu, Talwar, & Lee, 2010). If for instance, an experimenter has lipstick on her nose and asks a child how she looks, children as young as three lie and tell her that she looks okay (Talwar & Lee, 2002). By three and four, children (at least in the laboratory) inhibit their negative emotional responses to an undesirable gift in front of a gift giver (Cole, 1986). Children are more likely to lie for pro-social purposes than for self-protective purposes with age (Xu, Bao, Fu, Talwar, & Lee, 2010), and some studies suggest that girls are better than boys at regulating their verbal and nonverbal behaviors (Cole, 1986; Davis, 1995; Saarni, 1984).

Between ages three and ten, children are also increasingly able to understand others' masked emotions in social contexts. When predicting a recipient's response to an undesirable gift, children invoke both verbal display rules (e.g., judging that the recipient will tell a white lie) and facial display rules (e.g., judging that she will express happiness rather than disappointment; Broomfield, Robinson, & Robinson, 2002; see also Gnepp & Hess, 1986). Children appear to understand verbal display rules earlier than facial display rules (Broomfield, Robinson, & Robinson, 2002), and are better at understanding display rules for pro-social purposes than for self-protective purposes (Gnepp & Hess, 1986; but see Misailidi, 2006). These abilities may be influenced by family emotional climates. For example, negative expressiveness in a family

environment correlates positively with children's understanding of self-protective display rules and negatively with their understanding of pro-social display rules (Jones, Abbey, & Cumberland, 1998). Additionally, some researchers (Banerjee, 2002; Banerjee & Yuill, 1999a, 1999b; Naito & Seki, 2009) have argued that the understanding of social display rules relies on an ability to represent second-order mental state information. In support of this, children's performance on a second-order false belief task predicts their understanding of self-protective (Banerjee & Yuill, 1999b; Naito & Seki, 2009) and pro-social display rules (Naito & Seki, 2009).

However, much of this literature has used tasks with very rich contextual information (Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman & Liu, 2004; Naito & Seki, 2009; Gross & Harris, 1988). This is especially true for studies involving very young children. For example, in Banerjee's study (1997), preschoolers were read stories including an eliciting event (e.g., "Michelle is sleeping over at her cousin's house but she forgot her favorite teddy bear at home"), an agent's mental state (i.e., "Michelle is really sad that she forgot her teddy bear"), the agent's intention to hide her true feeling (i.e., "Michelle doesn't want her cousin to see how sad she is"), and a reason for hiding that feeling (i.e., "because her cousin will call her a baby"). Children were then asked about what the agent really feels and what expression she will display on her face. In such contexts, children successfully identify both the real emotion and the facial expression she will exhibit but their success may depend heavily on the detailed contextual information available in the stories.

Consistent with this concern, studies using less informative contexts have found that an understanding of masked emotion and social display rules emerges much later in development (Broomfield, Robinson, & Robinson, 2002; Gnepp & Hess, 1986; Jones, Abbey, & Cumberland,

1998). For instance, Gnepp & Hess (1986) provided children (first, third, fifth, and tenth graders) with an eliciting event and an agent's mental state but did not explicitly mention the agent's intention to hide her feelings nor any reason for her doing so. Nearly half of the first and third graders failed to use verbal display rules. Even adolescents (who successfully predicted the use of verbal display rules) frequently failed to predict that the agents would try to regulate their facial expressions. However, with less information in the stories, there is more uncertainty about whether the protagonist intended to be polite or not; given this uncertainty, children may have preferred to report the emotional expression that directly mapped onto the protagonist's true mental state.

Thus, there remains some ambiguity about what children understand, and when, about masked emotions. Rich detailed scenarios may overestimate children's ability to understand social display rules, while ambiguous scenarios may be open to interpretations that do not involve social display rules at all. Also critically for the present purposes, previous work does not ask whether children can use emotional expressions to recover information not only about the person displaying the emotion, but also about the person who is the intended audience of the emotion.

To see to what extent children can use emotional expressions in social and non-social contexts to recover both the protagonist's true feelings and those of her social partner, we introduce children to simple stories in which one of two teams wins a game. An observer of the game displays one of two emotional reactions (happy or sad) in front of a social partner and the contrasting emotional expression (sad or happy) behind the social partner's back. We ask children both what the person expressing the emotion wants and what her partner wants. Critically, not only do children not see the social partner's face, they have no other source of

information about his emotions or desires: the only way they can infer the social partner's desires is by using the protagonist's display of an emotion in his presence.

Since abundant work suggests that even infants and toddlers understand that someone whose desires are fulfilled will be happy and that someone whose desires are thwarted will be sad (see e.g., Skerry & Spelke, 2014; Stein & Levine, 1989; Wellman & Woolley, 1990; Yuill, 1984), we took it for granted that by middle childhood, children could make this inference. The critical question was whether children could recover each participant's true desires given that one person (henceforth the Protagonist) displayed contradictory emotions in the social and non-social contexts, and the other person (henceforth the Social Partner) never displayed any emotion at all. Given that without considerable scaffolding, children only appear to understand masked emotions relatively late in development (e.g., Broomfield et al., 2002; Gnepp & Hess, 1986; Jones et al., 1998), in Experiment 1 we test seven- to ten-year-olds. In Experiment 2, we reduce the task demands and test seven- to eight-year-olds. Experiments 3 to 5 investigate heuristics children might use to succeed on the task that do not rely on mental state understanding.

5.3 EXPERIMENT 1

5.3.1 Method

Participants

Ninety-two children between ages seven and ten ($M = 8.9$ years; range: 7.0-10.9; 54% girls) were recruited from an urban children's museum between January and November 2017 in the United States. While most of the children were white and middle class, a range of ethnicities and socioeconomic backgrounds reflecting the diversity of the local population (47% European American, 24% African American, 9% Asian, 17% Latino, 4% two or more races) and the

museum population (29% of museum attendees receive free or discounted admission) were represented throughout.

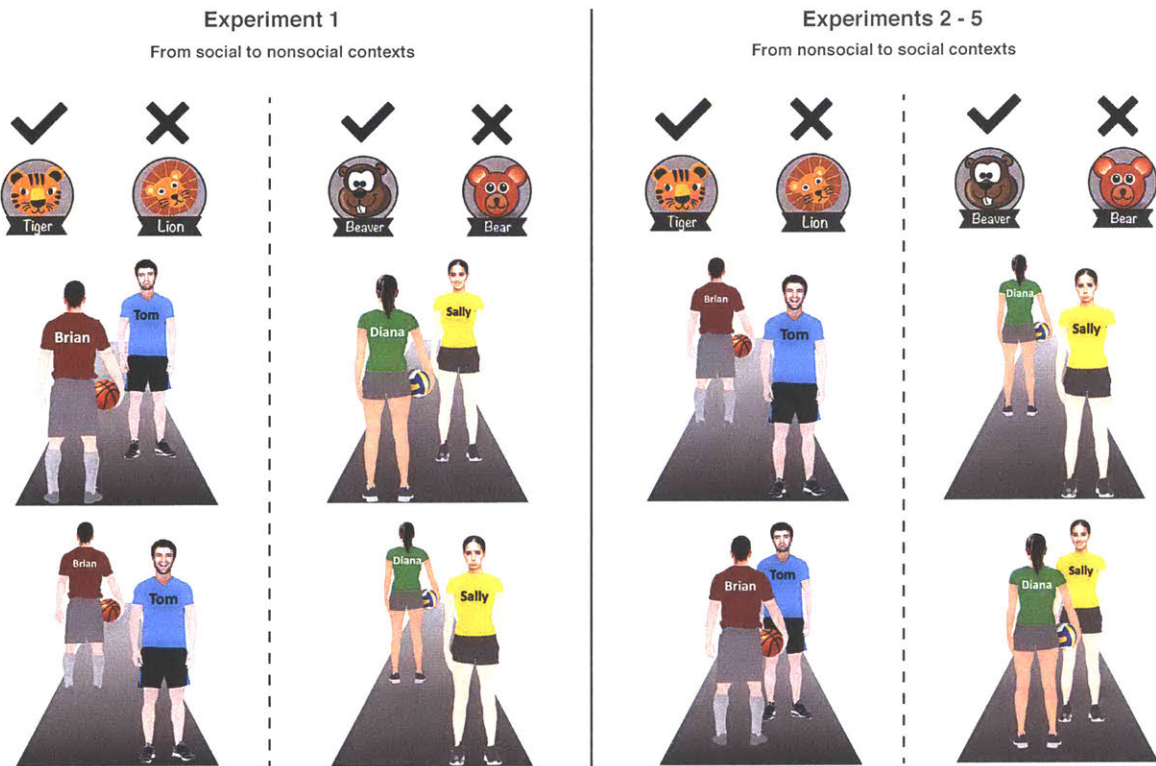


Figure 1 Example materials used in Experiments 1-5 (corresponding to the last three pictures described in Procedure).

Materials

Each child saw two illustrated stories. Different agents and games were used in each storybook (Tom, Brian, and basketball in one story and Sally, Diana, and volleyball in the other). One story presented the Happy-Sad condition (i.e., [Tom/Sally] was happy in front of [Brian/Diana] but sad behind [Brian/Diana]’s back) and the other presented the Sad-Happy condition (i.e., [Tom/Sally] was sad in front of [Brian/Diana] but happy behind [Brian/Diana]’s back). See Figure 1 Experiment 1 for examples. The particular story (i.e., basketball or volleyball) used in each condition (i.e., Happy-Sad or Sad-Happy), and the order of the two

conditions were counterbalanced across participants. The facial expressions were from iStock photos (<http://www.istockphoto.com/>).

Procedure

Children were tested individually; all sessions were videotaped. Each story was read consecutively, as follows (using the basketball-game story as an example). Children were asked check questions to encourage them to follow along. Incorrect responses were corrected throughout. Children had little difficulty with the check questions. (Collapsing data from all five experiments in the study, children's accuracies on the four check questions were .92, .95, .98, .99, respectively). Check questions were used only to maintain children's attention and were not used as inclusion criteria.

The experimenter placed the first picture on the table and said, "There is a basketball game today. It's the Tiger team against the Lion team." She introduced the next picture and said, "This is Tom. Tom is a basketball fan. He loves watching basketball games. He goes to watch the game. He is either a fan of the Tiger team, or the Lion team, but we don't know which one." Children were asked (Check question 1): "Do we know which team Tom is a fan of?" The experimenter introduced the third picture and said, "This is Brian. Brian was Tom's friend when they were little, but now they don't get to see each other very much. Brian became a basketball player. He plays in the game. He either plays for the Tiger team or the Lion team, but we don't know which one." Children were asked (Check question 2): "Do we know which team Brian plays for?" The experimenter introduced the fourth picture and said, "The results of the game were that the Tiger team won, and the Lion team lost." Then the experimenter introduced the fifth picture and said, "After the game, Brian ran back to the locker room. Tom was passing by and saw Brian. It was a very noisy and crowded room and they didn't have a chance to talk.

However, in front of Brian, when Tom came passing by, Tom made a face like this.” Children were asked (Check question 3): “Did Tom look happy or sad?” The experimenter introduced the sixth picture and said, “However, behind Brian’s back, as soon as Brian passed by and couldn’t see Tom, Tom made another face.” Children were asked (Check question 4): “Did Tom look happy or sad?” See Figure 1, Experiment 1 for the last three pictures. To match the contexts on surface details, both characters were present in both the social and non-social contexts; the difference was only that in the social context, they were facing each other, and in the nonsocial context, they were facing away from each other.

Finally, the experimenter asked two test questions. The first question was about the protagonist (Protagonist Question): “Now I am going to ask you some questions. In front of Brian, Tom looked [happy/sad] but behind Brian’s back, Tom looked [sad/happy]. Do you think Tom is a fan of the Tiger team or the Lion team?” The experimenter then asked the other test question (Social Partner Question): “Does Brian play for the Tiger team or the Lion team?” We asked about the team affiliation (rather than the direct question: “Who did Tom/Brian want to win?”) because it required children to reason about the character’s desires but seemed more natural in this context than asking children them to invoke a character’s desire for a counterfactual event. The experimenter coded children’s responses to the two test questions offline from videotapes. All these responses were recoded by an independent coder blind to conditions; there was 99% agreement on children’s responses.

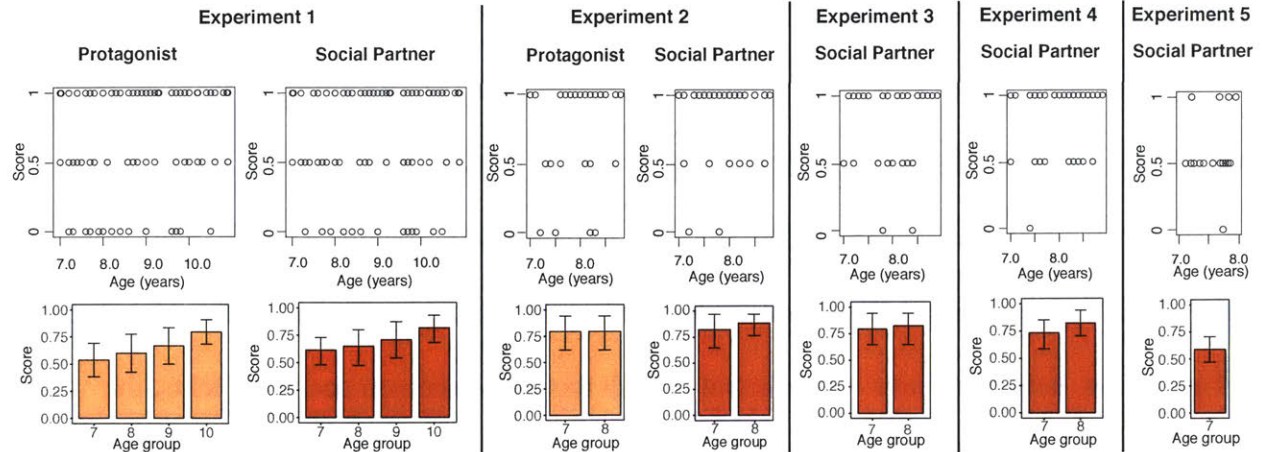


Figure 2 Results of Experiments 1-5. The top row shows individual children's performance as a function of age, and the bottom row shows children's performance averaged by age group. Error bars indicate 95% confidence intervals.

5.3.2 Results and discussion

We scored children's responses separately for the protagonist and the social partner. Children received one point for answering a question correctly and none for answering it incorrectly. Children's scores were then averaged across the two stories. Children successfully recovered both the protagonist's desire ($M = .65$, $SD = .40$, $z = 3.35$, $p = .001$; Exact Wilcoxon-Pratt Signed-Rank Test) and the social partner's desire ($M = .70$, $SD = .37$, $z = 4.50$, $p < .001$), and both abilities improved with age (protagonist's desire: $\beta = .38$, $SE = .17$, $z = 2.25$, $p = 0.024$; social partner's desire: $\beta = .35$, $SE = .17$, $z = 2.08$, $p = 0.038$; Ordinal Logistic Regression). See Figure 2.

Because of the significant age effect, we did a post-hoc, exploratory analysis by age group. Seven-year-olds ($n = 26$) and eight-year-olds ($n = 20$) did not perform above chance on either question (seven-year-olds: protagonist, $M = .54$, $SD = .40$, $z = .50$, $p = .804$; social partner, $M = .62$, $SD = .33$, $z = 1.73$, $p = .146$; eight-year-olds: protagonist, $M = .60$, $SD = .42$, $z = 1.07$, $p = .424$; social partner, $M = .65$, $SD = .40$, $z = 1.60$, $p = .180$). Nine-year-olds ($n = 24$) performed

above chance on the social partner question ($M = .71$, $SD = .41$, $z = 2.24$, $p = .041$) but not the protagonist question ($M = .67$, $SD = .43$, $z = 1.79$, $p = .115$). Ten-year-olds ($n = 22$) performed above chance on both (protagonist: $M = .80$, $SD = .30$, $z = 3.36$, $p < .001$; social partner: $M = .82$, $SD = .33$, $z = 3.30$, $p = .001$). See Figure 2.

As noted, many previous studies suggest that by seven and eight, children can predict an agent's real and apparent emotions given relatively rich contextual information (Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman & Liu, 2004; Naito & Seki, 2009; Gross & Harris, 1988; Gnepp & Hess, 1986; Broomfield, Robinson, & Robinson, 2002; Jones, Abbey, & Cumberland, 1998). They can also represent second-order mental state information (Perner & Wimmer, 1985; Sullivan, Zaitchik, & Tager-Flusberg, 1994), which supports the understanding of social display rules (Banerjee & Yuill, 1999b; Naito & Seki, 2009). Thus, it is possible that younger children's difficulties here were due to task demands. In particular, children may have tripped up by the fact that the order of presentation was fixed: the social context always preceded the non-social context, thus the first expression children saw was a masked emotional expression. Only when children saw the second expression, did they have sufficient information to realize that the first expression may have reflected only the protagonist's belief about the social partner, rather than the protagonist's true feelings.

In the next experiment, we reduce these task demands by flipping the order of the social and nonsocial contexts and test only seven and eight-year-olds. Here, children first see the agent's emotional expression in the nonsocial context and then the contrasting valence in the social context. This order does not require children to re-interpret the first emotional expression; additionally, the first expression may provide a basis for children to understand the expression displayed in the social context.

5.4 EXPERIMENT 2

5.4.1 Method

Participants

We calculated the sample size for this and all following experiments using the effect size found in the ten-year-olds of Experiment 1 and setting $\alpha = .05$. The power analysis indicated that a sample size of $n = 17$ per age group sufficed to detect effects with .95 power. Thus, we recruited children between ages seven and eight ($M = 7.9$ years; range: 7.0-8.8; 68% girls) from the children's museum with seventeen children in each age bin (seven-year-olds: $M = 7.5$ years, range: 7.0-7.9, 53% girls; eight-year-olds: $M = 8.3$ years, range: 8.0-8.8, 82% girls).

Materials, procedure and coding

The materials, procedure and coding were identical to Experiment 1 except that we flipped the order of the social and nonsocial contexts. See Figure 1. Specifically, instead of first showing Tom's emotional expression in front of Brian, the experimenter presented Tom's expression behind Brian's back, and said: "After the game, Tom made a face like this. At this moment, Brian was nearby but Tom didn't see him." Children were asked a check question: "Did Tom look happy or sad?" The experimenter then introduced the next picture and said, "However, Tom turned around and saw Brian. Tom made another face." Children were asked another check question: "Did Tom look happy or sad?" Children's responses to the test questions were recoded by a second coder blind to conditions; there was 99% agreement on children's responses.

5.4.2 Results and discussion

We used the same analyses as in Experiment 1. Overall, seven- and eight-year-olds recovered both the protagonist's desire ($M = .79$, $SD = .35$, $z = 3.78$, $p < .001$) and the social partner's desire ($M = .85$, $SD = .29$, $z = 4.54$, $p < .001$). There was no effect of age (protagonist:

$\beta = .41$, $SE = .73$, $z = .55$, $p = .580$; social partner: $\beta = .33$, $SE = .79$, $z = .43$, $p = .671$). Planned analyses revealed that both seven- and eight-year-olds succeeded in both questions (seven-year-olds: protagonist, $M = .79$, $SD = .36$, $z = 2.67$, $p = .013$; social partner, $M = .82$, $SD = .35$, $z = 2.84$, $p = .007$; eight-year-olds: protagonist, $M = .79$, $SD = .36$, $z = 2.67$, $p = .013$; social partner, $M = .88$, $SD = .22$, $z = 3.61$, $p < .001$). See Figure 2.

These results suggest that by age seven, children can use changing emotional expressions between social and nonsocial contexts to recover the desires of both participants in a social exchange. However, because the protagonist and the social partner questions were always asked in a fixed order, and the protagonist and the social partner were on different teams, children may have succeeded on the social partner question simply by flipping their answer to the protagonist question. Thus, in the next experiment, we replicate the design of Experiment 2 but ask only about the social partner.

5.5 EXPERIMENT 3

5.5.1 Method

Participants

As in Experiment 2, thirty-four children ($M = 8.0$ years; range: 7.0-8.9; 53% girls) were recruited from the children's museum. Seventeen of them were seven-year-olds ($M = 7.5$ years; range: 7.0-7.9; 59% girls). The remaining 17 children were eight-year-olds ($M = 8.4$ years; range: 8.0-8.9; 47% girls).

Materials, procedure and coding

The materials, procedure and coding were identical to Experiment 2 except that only the social partner question was asked. Children's responses to the test questions were recoded by a second coder blind to conditions; there was 99% agreement on children's responses.

5.5.2 Results and discussion

We used the same analyses as the proceeding experiments. Again, seven- and eight-year-olds recovered the social partner's desire ($M = .81$, $SD = .30$, $z = 4.20$, $p < .001$). There was no effect of age ($\beta = .41$, $SE = .68$, $z = .61$, $p = .542$). Planned analyses showed that both seven- and eight-year-olds succeeded on the social partner question (seven-year-olds: $M = .79$, $SD = .31$, $z = 2.89$, $p = .006$; eight-year-olds: $M = .82$, $SD = .30$, $z = 3.05$, $p = .003$). See Figure 2.

This experiment replicated Experiment 2 in suggesting that seven-year-olds can use someone's changing emotional expressions between social and nonsocial contexts to recover the desire of her social partner. These results at least partially rule out the possibility that children's success in the social partner question in Experiment 2 was due to their tendency to flip their answer to the protagonist question. However, even though the protagonist question was not asked, children may have inferred the desire of the protagonist and (given the change in valence in the protagonist's emotional expressions) simply concluded that the social partner had the opposing desire. In Experiment 4, we investigate this possibility by making the change of valence irrelevant to the inference. Experiment 4 is identical to Experiment 3 except that the protagonist is not a fan of either team; in the non-social context, she expresses an emotional response to an unrelated event (i.e., getting a new book or losing a favorite book). In this case, children can no longer infer the desire of the social partner by inferring the desire of the protagonist. If children still succeed, it provides stronger evidence that children can selectively use someone's emotional expression in a social context to recover the desire of her intended audience.

5.6 EXPERIMENT 4

5.6.1 Method

Participants

As in Experiments 2 and 3, thirty-four children ($M = 7.9$ years; range: 7.0-8.9; 32% girls) were recruited from the children's museum. Seventeen of them were seven-year-olds ($M = 7.5$ years; range: 7.0-7.9; 29% girls). The remaining 17 children were eight-year-olds ($M = 8.4$ years; range: 8.0-8.9; 35% girls).

Materials, procedure and coding

Experiment 4 was identical to Experiment 3 except as follows. When the experimenter introduced the protagonist, she did not say that he was a sports fan. Instead she said: "This is Tom. Today Tom got a new book he expected for a long time" (Happy-Sad condition) or "This is Tom. Tom lost his favorite book today" (Sad-Happy condition). Tom's facial expression was happy or sad respectively in the two conditions. The experimenter asked a check question: "Did he look happy or sad?"

Minor changes were made in the description of the last three pictures as well. See Figure 1. When the experimenter introduced the results of the game, she said: "The results of the game were that the Tiger team won and the Lion team lost." She also emphasized: "Everyone knew the results" to tell the child that Tom knew the results even though he did not go to watch the game. The experimenter then placed the next picture on the table and said: "After the game, Tom was still [happy/sad] about his [new/lost] book. At this moment, Brian was nearby but Tom did not see him." She placed the final picture on the table and said: "Then Tom turned around and saw Brian. Tom made a different face." The experimenter asked another check question: "Did he look happy or sad?" The materials, procedure and coding were otherwise the same as

Experiment 3. Children's responses to the test questions were recoded by a second coder blind to conditions; there was 99% agreement on children's responses.

5.6.2 Results and discussion

We used the same analyses as the proceeding experiments. Children again successfully recovered the social partner's desire ($M = .78$, $SD = .28$, $z = 4.15$, $p < .001$) and there was no effect of age ($\beta = .75$, $SE = .62$, $z = 1.21$, $p = .228$). Planned analyses revealed that children in both age groups succeeded at the task (seven-year-olds: $M = .74$, $SD = .31$, $z = 2.53$, $p = .021$; eight-year-olds: $M = .82$, $SD = .25$, $z = 3.32$, $p < .001$). See Figure 2.

Thus, together with Experiment 3, Experiment 4 provides converging evidence that children's success in the social partner question cannot be explained by their tendency to choose the desire contrasting with that of the protagonist. Even when the desire of the protagonist was not explicitly referenced (Experiment 3), and when the protagonist's affiliation with the teams was unknown (Experiment 4), children selectively used the protagonist's emotional expression in the social context to infer the desire of her social partner.

In the next experiment, we investigate a different heuristic that children might have used to pass the social partner question: the protagonist's facial expression when the protagonist and the social partner were facing each other. To rule out the possibility that children simply assumed the social partner had the same expression as the protagonist when the two were facing each other, in Experiment 5, we made the protagonist's emotional expressions in both the social and nonsocial contexts irrelevant to the story. Specifically, in the Happy-Sad condition, children were told that the protagonist was happy about her new book in the non-social context and sad in front of her social partner because her new book fell into a puddle and got all muddy; in the Sad-Happy condition, the protagonist was sad about her lost book in the non-social context but happy in

front of her social partner because she found the lost book. If children simply used a face-matching heuristic in the face-to-face context to pass the social partner question they should perform similarly here. However, if children's successes in Experiments 2-4 were based on mental state inferences about social display rules, then children should perform at chance in Experiment 5 where the expressions are not governed by those rules. Because the concern about using a simple face matching heuristic presumably applies more to younger children than older ones, we focused only on seven-year-olds in Experiment 5.

5.7 EXPERIMENT 5

5.7.1 Method

Participants

Seventeen seven-year-olds ($M = 7.6$ years; range: 7.1-7.9; 65% girls) were recruited from the children's museum.

Materials, procedure and coding

The materials, procedure and coding were the same as Experiment 4 with one exception. In the Happy-Sad condition, when the experimenter introduced the last picture, she said: "Then Tom turned around and saw Brian. At the same moment, Tom's new book fell into a puddle and got all muddy!" In the Sad-Happy condition, she said: "Then Tom turned around and saw Brian. At the same moment, Tom saw his book sitting right there on the bench. So his book wasn't lost at all!" The experimenter then asked a check question: "Did he look happy or sad?" Children's responses to the test questions were recoded by a second coder blind to conditions; there was 100% agreement on children's responses.

5.7.2 Results and discussion

Here, where the protagonist was responding to an unrelated event in the social context seven-year-olds did not use the protagonist's emotional expression to infer the desire of the social partner and performed at chance ($M = .59$, $SD = .26$, $z = 1.34$, $p = .375$). There was no effect of age ($\beta = 1.12$, $SE = 1.98$, $z = .56$, $p = .573$). See Figure 2. This suggests that children's ability to recover the desire of the social partner in Experiments 2 to 4 was not simply due to matching the protagonist's expression when the protagonist and social partner were facing each other. Since children did not simply assume the protagonist's expression matched the social partners when they were facing each other but the protagonist was responding to an unrelated event, it is unlikely that children's ability to recover the desire of the social partner in Experiments 2 to 4 was due to a simple heuristic rather than the (behaviorally equivalent but inferentially richer) understanding that the protagonist was displaying the emotion congruent with her social partner's desires.

5.8 GENERAL DISCUSSION

In five experiments, we investigated children's ability to use the information embedded in social display rules to recover others' otherwise under-determined mental states. Children saw an emotional expression when a protagonist was in front of a social partner (Social Context), and a different expression when the protagonist was behind the social partner's back (Nonsocial Context). Children as young as seven were able to use the expression in the nonsocial context to infer the protagonist's desire, and the expression in the social context to infer the social partner's desire.

Our study builds on many previous studies that have looked at children's ability to predict an agent's real and apparent emotions given rich mental state information (e.g., the agent's

desires, true feelings, her intentions, and a motivation to hide her true feelings; Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman & Liu, 2004; Naito & Seki, 2009; Gross & Harris, 1988; Gnepp & Hess, 1986; Broomfield, Robinson, & Robinson, 2002; Jones, Abbey, & Cumberland, 1998). Here by contrast we provided children with very minimal background information, and no direct information about the agent's mental states. Nonetheless, given an agent's two contradictory emotional reactions to an event, children were able to use the agent's expression in a non-social context to recover her true desire and her expression in the social context to recover the desire of a social partner, whose emotional expressions were never observed at all. These results are consistent with a host of studies suggesting that children can recover rich information from observed emotional cues (e.g., Berman, Chambers, & Graham, 2010; Feinman, Roberts, Hsieh, Sawyer, & Swanson, 1992; Wu, Muentener, & Schulz, 2017; Repacholi & Gopnik, 1997; Rieffe, Terwogt, & Cowan, 2005; Wellman, Philips & Rodriguez, 2000; Wu & Schulz, 2017). It goes beyond those studies in suggesting that children can use the information in emotional expressions regulated by social display rules to infer mental states otherwise underdetermined by the context.

Note also that our experiment contained both contexts in which the protagonist's expression in the social context most likely belied her true feelings (Experiments 1-3, where the protagonist displayed the opposite valenced emotion in the non-social context) and contexts in which there was no reason to suppose that the protagonist's expression in the social context was not entirely sincere (Experiment 4, where the opposite valenced emotion in the non-social context referred to an unrelated event). These different contexts capture some of the range of authentic emotions that might underlie any socially displayed expressions: Protagonists may wholeheartedly share the feelings of their social partner, have genuine feelings on the other

person's behalf independent of their own feelings about the matter (e.g., someone may be sincerely happy or sad for someone else even if they feel the opposite way themselves), be entirely neutral, or be at odds with their partner and displaying feigned, inauthentic emotions. Much of the developmental research has investigated the last of these, focusing on children's ability to distinguish real and apparent emotions (Banerjee, 1997; Harris, Donnelly, Guz, Pitt-Watson, 1986; Misailidi, 2006; Josephs, 1994; Wellman & Liu, 2004; Naito & Seki, 2009; Gross & Harris, 1988; Gnepp & Hess, 1986; Broomfield, Robinson, & Robinson, 2002; Jones, Abbey, & Cumberland, 1998). The current research however, suggests that the emotion apparent in social contexts may be informative despite the range of real emotions that could underlie it; although displayed emotions may be ambiguous about the protagonist's own feelings, they may nonetheless be relatively unambiguous about what the protagonist thinks about her audience's beliefs and desires. Consistent with this, researchers have suggested that emotional expressions may serve both as a component of an authentic emotional response and be adapted for communicative purposes (see Shariff & Tracy, 2011 for review).

Although there has been debate on the extent to which reasoning about pro-social display rules requires second-order mental state representation (Banerjee, 2002; Banerjee & Yuill, 1999a, 1999b; Naito & Seki, 2009), the debate pertains largely to contexts in which children might understand the social-display rule because the intended audience's desires are well-established (e.g., as in gift giving scenarios where the giver presumably wants the receiver to like the present). In our task by contrast, the social partner's beliefs, desires, and emotions were unknown throughout. To recover information about the social partner, children had to selectively use the protagonist's emotional expressions to gain insight into the mind of her audience. We believe that this kind of inference does require recursive mental state reasoning. The current

results suggest that the ability to make these inferences is present by middle childhood, consistent with other work on children's ability to entertain second-order beliefs like "John thinks that Mary thinks ..." (see Grueneisen, Wyman, & Tomasello, 2015; Perner & Wimmer, 1985; Talwar, Gordon, & Lee, 2007; though see Sullivan, Zaitchik, & Tager-Flusberg, 1994 for even earlier success on simplified tasks).

Critically, children succeeded here in a very tightly constrained context: there were only two possible outcomes (one of two teams won a game), two possible emotional responses (happy or sad) and two social partners. Moreover, the task design virtually eliminated any memory demands: children did not need to track the changing emotional expressions over time; they were all concurrently displayed in the storybook card format, together with the social context. Future work might look at children's ability to draw comparable inferences when they must track changing emotional dynamics over time and in more complex, multi-participant scenarios. Note however, that although more realistic scenarios may add processing demands and complexity, they may also provide children with richer cues to agents' mental states.

Overall the current results suggest that at least in constrained contexts, children can recover otherwise underdetermined mental states from emotional expressions in social contexts. Intriguingly, the current results also suggest that there is a limit to how much we can hide in hiding our feelings. In disguising our true feelings, we may reveal what we think about what other people want.

Chapter 6 General Conclusions

6.1 SUMMARY AND FUTURE DIRECTIONS

6.1.1 Study 1 Inferring External Causes of Emotional Expressions

Summary: In a series of experiments, I investigated whether children ages 1-4 years and adults could make fine-grained within-valence distinctions among emotional expressions, and connect those expressions to their probable eliciting events. In a forced-choice task, children ages 2-4 years and adults successfully identified probable causes of five distinct positive emotional vocalizations elicited by funny, exciting, delicious, sympathetic and adorable stimuli (Experiment 1). Similar results obtained in a preferential looking task with 12-23-month-olds, a direct replication with 18-23-month-olds (Experiment 2), and a simplified design with 12-17-month-olds (Experiment 3; pre-registered on the Open Science Framework). To validate these results with converging measures, and to look at whether infants might actively search for causes of others' emotional reactions, we tested 12-17-month-olds with a manual search task (Experiments 4 and 5; both pre-registered). An experimenter peeked through a peep hole in top of a box and made one of two vocalizations (Experiment 4: "Aww!", as if seeing something cute, or "Mmm!", as if seeing something delicious; Experiment 5: "Whoa!", as if seeing something exciting, or "Aww!"). We found that 12-17-month-olds were more likely to reach into the box again and search longer after they retrieved a toy incongruent with the vocalization (e.g., retrieving a banana upon hearing "Aww!") than a toy congruent with it (e.g., retrieving a stuffed animal upon hearing "Aww!"). These results suggest that early emotion understanding is much more sophisticated than previously believed, extending well beyond the ability to distinguish a few basic emotions or contrasts across valence. As young as 12-17 months, infants make

nuanced distinctions among positive emotional vocalizations and connect them to their probable eliciting causes.

Future directions: Future research could look at the emergence of such abilities in the first year of life. In our study, we did not find a significant age effect between ages 12 and 17 months. This suggests that some of these abilities may be present even in infants younger than 12 months. Consistent with this speculation, other studies found that infants at the second half of the first year reliably discriminate some canonical emotional expressions and can match them cross-modally (e.g., matching happy and sad faces to happy and sad voices respectively; Walker-Andrews, 1997; Walker-Andrews & Lennon, 1991). Additionally, infants at ten months can connect negative emotions with failed goals rather than completed ones (Skerry & Spelke, 2014), and infants between seven and eighteen months can associate a frightened voice with evolutionarily relevant causes such as snakes (DeLoache & LoBue, 2009). Thus, it is possible that we can find some abilities tested in Study 1 in infants within their first year of life as well.

Additionally, much remains to be known about the full space of early emotion understanding. Previous literature has primarily looked at young children's ability to make distinct inferences from positive and negative emotions, or the ability to discriminate a small number of basic emotions. Our Study 1 takes one step towards discovering young children's understanding of other emotions. Our success suggests that there may be a lot more to be explored. Consistent with this, a recent study (Soderstrom, Reimchen, Sauter, & Morgan, 2017) found that infants show some ability to discriminate the emotional vocalizations of triumph and relief. More work is warranted to give a complete picture of the space of early emotion understanding.

Last, as with infancy research in general, studies on emotion understanding in autism spectrum disorders (ASD) have focused almost exclusively on positive and negative emotions or

a few basic emotions, and the findings of those studies have been mixed and controversial (see Harms, Martin, & Wallace, 2010 for a review). Our work looking at a wider range of emotional expressions in typical development lays the foundation for studying atypical development associated with ASD, which may provide new insights into the symptoms and early diagnosis of ASD.

6.1.2 Study 2.1: Inferring Beliefs and Desires From Emotional Expressions

Summary: Here I tested children ages 4-5 years and adults' ability to recover internal mental states (i.e., beliefs and desires) from observed emotional expressions. Beliefs and desires are important moderators of individuals' emotional responses to an external cause; but in many cases beliefs and desires have to be inferred simultaneously because both are intrinsically hidden. Thus in three experiments, I investigated an emotional cue that has rarely been studied but could inform both beliefs and desires: the dynamics of valenced emotional expressions. We showed participants someone's face retaining the same happy or sad expression, or changing from happy to sad, or from sad to happy, between expected and actual event outcomes. Adults recovered both whether she wanted the outcome to occur and whether she had a true or false belief about the outcome in all cases. Five-year-olds performed like adults given changing emotional expressions; however, they inferred only desires, but not beliefs, given stable valenced expressions. Four-year-olds inferred only desires in all conditions. These results suggest that the ability to recover mental states from emotional expressions develops gradually over childhood, but that dynamic emotional expressions provide crucial information about others' beliefs and desires. When someone's face changes between anticipated and actual outcomes from happy to sad, or from sad to happy, children at least by age five gain insight into not only how she feels, but also what she wants and believes about the world.

Future direction: One limitation of this study is that the emotional information that is at stake here may only be the valence information: whether an emotional response is positive or negative. Future research could look at children's ability to connect mental states with other kinds of emotion contrasts. For example, intuitively, someone's beliefs about responsibility attribution differentiate two within-valence emotional responses: if someone thinks that a bad outcome is caused by herself, she may feel *guilty*; but if she thinks that it is caused by others, she may feel *mad* at other people. Take another example. Someone's knowledge state of an event (i.e., knowledgeable or ignorant) may differentiate whether she will or will not have an emotional response at all. So the contrast here is not between positive and negative emotional responses, or between two within-valence emotional reactions, but is between the presence and absence of an emotional response.

6.1.3 Study 2.2: Inferring Beliefs and Desires From Emotional Expressions: A Computational Model

Summary: Here I developed a formal model that accounts for adults' inferences of beliefs and desires from emotional expressions in diverse contexts and conditions. I began by specifying a probabilistic generative model of how an agent's beliefs and desires about an event might lead to goal-directed actions, and how these beliefs and desires might also generate emotional reactions to the expected and actual outcomes of the event. This forward model forms the core of a Bayesian account, letting us consider how an ideal observer might reason backward from an agent's emotional reactions to the beliefs and desires that generated them. I validated this Bayesian account with four behavioral experiments. I manipulated whether participants saw an agent's emotional reaction only to the actual outcome (Experiments 1 and 3) or emotional reactions to both the anticipated and actual outcomes (Experiments 2 and 4), and whether the

agent merely observed the event outcome (Experiments 1 and 2), or caused it (Experiments 3 and 4). I found that participants recovered the agent's desires throughout, but their ability to recover her beliefs varied depending on the amount of the emotional cues they saw and whether the agent's actions and emotional reactions provided congruent information about her mental states. The Bayesian model captured the behavioral results in all conditions ($r_s = .950$ or above). This suggests that participants integrated multiple sources of information together (i.e., actions, emotional reactions, and context) to jointly infer others' beliefs and desires. This study bridges theory of mind research, accounts of emotion attribution, and formal modeling, to provide a more unified formal model of how we reason about others' emotional responses to events.

Future direction: 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 adults' intuitive theory of emotion, it does not specify the precise mechanism by which people represent these relationships. That is, the functional form between these causal relationships is represented only implicitly in the forward judgments (i.e., subjective judgments of priors and likelihoods) which were elicited directly from human subjects rather than explicitly modeled. It remains an important task for future work to explicitly model how people represent these fine-grained generative relationships.

6.1.4 Study 3: Inferring Recursive Mental States From Emotional Expressions

Summary: People's emotional expressions in a social context are often regulated by social display rules. In this study, I looked at a feature of social display rules that has been largely overlooked: emotional expressions displayed in a social context may disguise an individual's own feelings while being informative about the feelings of her social partner. I presented 7- to 10-year-olds a protagonist's emotional expression (i.e., happy or sad) in front of her social partner (Social

Context), and a different one (i.e., sad or happy) behind her social partner's back (Nonsocial Context). At least around age seven, children were able to use the emotional expression in the nonsocial context to infer the desire of the protagonist. More importantly, they were also able to use the expression in the social context to infer the desire of the social partner, who showed no emotional expressions at all. For instance, given that at a basketball game, the Bears beat the Lions, and someone looked sad on seeing a team player but happy when she was alone, children not only inferred that the person was a fan of the Bears but also that the team player played for the Lions. These findings suggest that at least by middle childhood, children are able to use someone's changing emotional expressions between social and nonsocial contexts to recover not only the desire of the person expressing emotions but also that of her intended audience.

Future directions: This study tests children's understanding of the communicative role of emotional expressions and found such abilities in children at age seven. However, much remains to be known about whether children consider others' emotional expressions as communicative signals (as opposed to intrinsic, authentic emotional responses to events that are relatively immune to social contexts) more broadly. For example, it is possible that children in Studies 1 and 2.1 interpreted the emotional expressions as communicative signals as well. For example, when an adult expressed a "Whoa!" in front of them, they considered this behavior as trying to signal something exciting in the environment, rather than an intrinsic emotional response to an object. If this is the case, children's interpretation of someone's emotional expressions may be closely embedded with and influenced by the social context an emotion is expressed (e.g., whether there is someone observing this emotional expression and what are the observer's beliefs, desires, and knowledge). Consistent with this, researchers (e.g., Chapman, Kim, Susskind, & Anderson, 2009; Eibl-Eibesfeldt, 1989; Ekman, 1992; see Shariff & Tracy,

2011 for review) have proposed that although emotional expressions may be evolved originally to serve adaptive functions, the primary purpose of emotional expressions in contemporary human life may have more to do with their capacity to quickly and nonverbally communicate socially significant information. Thus, one important question for future research is to investigate how infants and children flexibly interpret others' emotional expressions given different social contexts.

Additionally, the mental state inferences here were challenging insofar as the input was impoverished: one character showed two conflicting emotional expressions, the other showed none, and both of them did not engage in any goal-directed actions. However, in this study (and in Study 2.1 as well), the hypothesis space was restricted to two alternatives, children had continuous access to the agent's emotional expressions in both contexts, and the emotional expressions were highlighted. In the real world, emotional expressions are transient and typically go unremarked. Future research could investigate children's ability to recover agents' mental states in contexts where the emotional responses unfold in time, and both the hypotheses and emotional expressions are more complex than those used here.

6.2 BROADER FUTURE DIRECTIONS

This thesis looks at children and adults' ability to recover rich unobserved information from observed emotional expressions. A more fundamental question that underlies this research program is: what are the content and structure of people's intuitive theory of emotion that support such sophisticated inferences? Despite decades of research on emotion, it remains an important question for future work. Critically, one challenge for future research is that although there has been a large body of work looking at a small number of basic emotions or a few core dimensions of emotions (e.g., valence and arousal), the fine-grained nature of this intuitive

theory of emotion has largely been unexplored. Intuitively however, our conceptual knowledge of emotion is so sophisticated that it cannot be simply reduced to understanding a few emotion categories or dimensions. Take for instance this quote by Stephen King (2010): “I recognize *terror* as the finest emotion and so I will try to terrorize the reader. But if I find that I cannot terrify, I will try to *horrify*, and if I find that I cannot horrify, I'll go for the *gross-out*. I'm not proud.” More salubriously, here is the beginning of a list of words for happiness (De Rose, 2005): “*airy, amused, animated, beatific, blissful, blithe, bright, brisk, buoyant, cheerful, cheery, comfortable, contented...*” To the degree that we distinguish these emotions, we represent not only the meaning of these emotion words, but also the causes and contexts that elicit them and the expressions and vocalizations that accompany them. Although the studies in the current thesis take several steps towards uncovering the form and content of this rich intuitive theory of emotion, much remains to be known about how this knowledge is structured and what are the precise mechanisms that compute the causal relationships between different components in the framework.

To answer these questions, multiple approaches can be used. First, I argue that studying infants and children gives us important insight into the content and structure of this intuitive theory of emotion. Much of our commonsense knowledge of the physical and social world is constructed in early childhood. By studying infants and children, we can identify the fundamental concepts and learning principles that are likely to be crucial for understanding human mind and behavior broadly. Additionally, the developmental trajectory of an ability may shed light on the underlying structure of our knowledge: the representations that emerge early in development are more likely to form the core structure and those that emerge late are more likely to build on, enrich, or revise a pre-existing framework. Such developmental approach has been

used to study human cognition in many domains (e.g., Carey, 2009; Gopnik & Wellman, 2012; Spelke, 2017) and may provide an important entrée to understanding people's intuitive theory of emotion as well.

Second, recent progress in the domain of computational cognitive sciences may provide new opportunities to answer these questions. Researchers in that area have used computational models to characterize people's intuitive theories in domains that are closely related to emotions, including how people, as observers, represent an agent's beliefs, desires, and actions (e.g., Baker, Saxe, & Tenenbaum, 2009; 2017), her costs and rewards (see Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016 for a review), the moral permissibility of her behaviors (e.g., Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015), and her reasoning about counterfactual possibilities (e.g., Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014). Although people represent emotions in almost all of these cases, none of these models make reference to emotions at all. However, these models provide important bases for studying people's intuitive theory of emotion. While Study 2.2 in this thesis takes one *small* step towards modeling the link between emotions and theory of mind, future research could look at how people's representations of emotions are causally and structurally intertwined with many other cognitive components as mentioned above, and look at the fine-grained functional forms of these generative causal relationships, to provide a unified formal account of people's intuitive theory of emotion.

Last, as proposed by Marr (1982), the goals of cognitive science include an understanding of the human mind at three distinct, complementary levels of analysis: computational, algorithmic/representational, and implementational. Thus, to fully understand

people's intuitive theory of emotion, future work also needs to look at how the conceptual knowledge of emotion is encoded in the human brain. This is a potentially challenging task for neural scientists both because of the limitation of current methods and tools and because of the complicated nature of emotion understanding, which may involve not only domain-specific knowledge of others' minds, but also domain-general world knowledge, semantic representations, and sensory processing. On the other hand, the intuitive theory of emotion can serve as an interesting and exciting case study for understanding the architecture of the human brain, which may advance both the development of novel tools, and our insight into the human mind and brain.

Appendix I. Study 1 Supporting Information Appendix

1. Mixed-effects models

1.1. Experiment 1

We used a mixed-effects model to analyze the effects of Category (of the eliciting cause) and Age on children's performance. Children's accuracy by Category was calculated as the proportional accuracy on the trials involving each category, either as a target or distractor. Note that this controls for baseline preferences; if a child simply preferred to point to, for instance, the picture of a silly face, she would be correct when the picture was a target but incorrect when it was a distractor. This resulted in a 5 (Category) by 48 (Subject) data matrix. By design, Category (a categorical variable) and Age (a continuous variable) were fixed factors, and Subject was a random factor. We used likelihood ratio tests to decide whether the interaction term should be added to the final model. The final best-fit model did not include the interaction term. See Tables S1 and S2.

1.2. Experiment 2

A mixed-effects model was used to analyze the effects of Time, Category, and Age on participants' performance. As above, each picture appeared as a target half the time and a distractor half the time, controlling for baseline preferences. We first calculated participants' accuracy on each of the seven seconds on each trial, averaged across the four 7-second intervals. (See Fig. 2A.) Then we averaged across trials involving each category. This resulted in a 7 (Second) by 5 (Category) by 32 (Subject) data matrix. Time (a categorical variable), Category (a categorical variable) and Age (a continuous variable) were fixed factors, and Subject was a random factor. We treated Time as a categorical variable rather than a continuous one because the effect of Time on accuracy was nonlinear (see Fig. 2B) and our interest was in whether Time

had an effect on accuracy at all rather than in its specific relationship (e.g., linear, quadratic, or cubic) to accuracy. We used likelihood ratio tests to see whether any interaction term or the autocorrelation structure of order 1 would significantly improve the model. The final best-fit model included the interaction between Time and Age, and the autocorrelation structure of order 1. See Tables S3 and S4.

The post-hoc analysis of age used the same best-fit model except that instead of taking Age as a continuous variable, we did a median split on age and looked at the effects of Time and Category on each of the two age groups. See Table S5. We used the same model to analyze the data replicating the older age group. See Table S6. (In this replication study, we also initially pre-registered an analysis to investigate the possibility of a quadratic curve in 18-23-month-olds' preferential looking over the 7-second interval; however, we elected not to pursue this line of analysis and instead treated Time as a categorical variable.)

1.3. Experiment 3

We analyzed the effects of Time, Category, and Age on participants' performance. We calculated participants' accuracy on each of the seven seconds on each trial, averaged across the two 7-second intervals (see Fig. 2A). Then we averaged across trials involving each category of the eliciting cause, either as a target or a distractor (note that, as above, this controls for baseline preferences). As in the previous experiments, we used likelihood ratio tests to select the final best-fit model. It did not include any interaction term but included the autocorrelation structure of order 1. See Tables S7 and S8.

1.4. Experiment 5

We were interested in the effects of Congruency, Category, and Age on participants' performance. By design, Congruency (a categorical variable), Category (a categorical variable),

and Age (a continuous variable) were fixed factors, and Subject was a random factor. The likelihood ratio tests suggested that no interaction term should be added to the final best-fit model. See Tables S9-S10.

2. Pre-registration for Experiment 5

In addition to the four analyses reported in the paper, we also mistakenly pre-registered an incorrect analysis: a mixed-effects model looking at the effects of Congruency, Category, and Age on infants' proportional searching in the Congruent and Incongruent conditions. However, because each participant's proportional searching in the two conditions always summed to 1, this analysis cannot estimate the main effects of Category and Age and we did not run it.

3. Pilot study for Experiment 4

We conducted a pilot study including 16 12-17-month-olds (M: 15.3 months, range: 12.0-17.6). There was a trend for participants to search longer in the Incongruent than the Congruent condition (Incongruent: M=5.32s, SD=5.487, 95% CI [2.95, 8.13]; Congruent: M=2.38s, SD=2.332, 95% CI [1.34, 3.53]; $Z=1.73$, $p=.083$, permutation test).

4. Exploratory analyses of Experiment 4

Exploratory analyses collapsing the pilot sample ($N=16$) and the main sample ($N=36$) suggested that the effect between conditions was present in 15-17-month-olds but absent in 12-14-month-olds. The older group ($n=24$) searched significantly longer in the Incongruent than the Congruent condition (Incongruent: M=5.74s, SD=5.371, 95% CI [3.90, 8.07]; Congruent: M=2.14s, SD=2.897, 95% CI [1.26, 3.69]; $Z=2.52$, $p=.009$, permutation test). The younger group ($n=28$) did not (Incongruent: M=3.49s, SD=5.557, 95% CI [2.01, 6.45]; Congruent: M=2.46s, SD=3.335, 95% CI [1.48, 4.01]; $Z=.99$, $p=.363$).

5. Exclusion criteria

In Experiment 1, 8 children were replaced due to pointing to the left (or right) pictures throughout ($n=5$), refusal to point ($n=2$), or distraction ($n=1$). In Experiment 2, including both the initial study and the replication, 24 infants were replaced due to fussiness ($n=12$), parental interference ($n=2$), experimenter error ($n=1$), or distraction ($n=9$). In Experiment 3, 9 infants were replaced due to fussiness ($n=1$), parental interference ($n=1$), experimenter error ($n=2$), distraction ($n=4$) or hearing loss ($n=1$). In Experiment 4, 5 infants were replaced due to fussiness ($n=2$), parental interference ($n=2$), or not reaching into the boxes ($n=1$). In Experiment 5, 7 infants were replaced due to fussiness ($n=6$) or not returning the familiarization toys ($n=1$).

6. Procedure of the practice/training trials

6.1 Experiment 1

For the practice trial, the experimenter displayed two warm-up pictures (i.e., a beautiful beach and a dying flower) on the screen and said: "In this game I will show you two pictures. Sally will look at one of them and make a sound. We will need to guess which picture Sally is looking at. OK? I'll play the game first!" The experimenter played the positive sound, pointed to the picture of the beach, and said: "When Sally makes this sound, I think she is looking at the picture of a beautiful beach." Then she played the sad sound, pointed to the dying flower, and said: "When Sally makes this sound, I think she is looking at this dying flower." Then the experimenter said: "Now it's your turn to play the game!" and started the test trials.

6.2 Experiments 4 and 5

During the training phase, the experimenter took out a box and said: "Now let's play a game! This is my box. Let me see what's inside." The experimenter looked into the box through the top hole, saying: "I see something!" She reached through the felt opening and retrieved a ball. She looked at the child and said: "Look! It's a ball!" She squeaked the ball three times and

put it in the tray. Then she said: “Let me see what else is inside!” She searched the box for approximately three seconds and then opened her hand to show there was nothing on it, saying: “Nothing else!” Then the experimenter put the ball back in the box and affixed the box to the table with the felt opening facing the child. The experimenter told the child: “Now it’s your turn to play! Take the ball out and put it in the tray!” After the children retrieved the ball, they were encouraged to place it in the tray. Once they did, the experimenter took the ball away. A similar procedure was repeated with the other training box. When the experimenter searched the box the first time, she retrieved a red ball. When she searched the box again, she retrieved a toy car (Experiment 4) or a toy banana (Experiment 5). Then she searched the box again and found nothing.

7. Coding

7.1. Experiments 2 and 3

In both Experiments 2 and 3, a coder, blind to both the eliciting causes and the vocal expressions, coded the children’s looking offline from videotape during each 7-second interval (the 4 seconds when the vocalization was played and the 3 second pause that followed it). A second coder coded 33% of the clips. Inter-coder agreement was 90% for Experiment 2 and 96% for Experiment 3. Accuracy was calculated as children’s looking time to the image corresponding to the vocalization over their total looking time to both images during the 7-second intervals.

7.2. Experiments 4 and 5

Four behaviors were coded as “searching”: reaching into the box through the front opening, trying to reach into the box through the top hole, trying to look through the front opening, and trying to look through the top hole. All other behaviors (e.g., pushing the box,

playing with the felt, and resting a hand on the top hole) were coded as “not searching”. The coding criteria were the same for Experiments 4 and 5. Coders were trained together and then coded the video clips separately. All coders were blind to the experimental conditions. In Experiment 4, a primary coder coded all video clips, and a second coder coded 90% of the video clips. Inter-coder agreement was 92%. In Experiment 5, a primary coder coded all video clips and a second coder coded 60% of the video clips. Inter-coder agreement was 95%.

Table S1 Model selection in Experiment 1

Model		df	AIC	BIC	Log likely- hood	Test	Likeli- hood ratio	<i>p</i> -value
1 best fit	Age+Category+(1 Subject)	8	-156.5	-128.7	86.3	-	-	-
2	Age+Category+Age:Category +(1 Subject)	12	-155.8	-114.0	89.9	1 vs 2	7.3	.123

Table S2 Summary of the best-fit model in Experiment 1

Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	0.24	0.09	188	2.71	.008
Age	0.14	0.02	46	5.72	<.001
Category-Exciting	-0.01	0.03	188	-0.34	0.731
Category-Adorable	-0.01	0.03	188	-0.30	0.761
Category-Sympathetic	0.02	0.03	188	0.55	0.584
Category-Delicious	0.04	0.03	188	1.27	0.205
Random effects	Std. Deviation				
(Intercept)	.13				
Residual	.15				
Number of observations: 240; number of participants: 48					

Table S3 Model selection in Experiment 2

Model		df	AIC	BIC	Log likely- hood	Test	Likeli- hood ratio	<i>p</i> -value
1	Age+Time+Category+(1 Subject)	14	-470.4	-401.3	249.2	-	-	-
2	Age+Time+Category+(1 Subject) [corAR(1)]	15	-488.5	-414.4	259.2	1 vs 2	20.0	<.001
3 best fit	Age+Time+Category+Age:Time +(1 Subject) [corAR(1)]	21	-493.7	-390.0	267.8	2 vs 3	17.2	.009
4	Age+Time+Category+Age:Time +Age:Category+(1 Subject) [corAR(1)]	25	-487.9	-364.6	269.0	3 vs 4	2.3	.686
5	Age+Time+Category+Age:Time +Time:Category+(1 Subject) [corAR(1)]	45	-465.0	-242.9	277.5	3 vs 5	19.3	.735
6	Age+Time+Category+Age:Time +Age:Time:Category+(1 Subject) [corAR(1)]	49	-458.4	-216.6	278.2	3 vs 6	20.7	.836

Note: corAR(1) stands for autocorrelation structure of order 1.

Table S4 Summary of the best-fit model in Experiment 2

Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	.56	.10	980	5.64	<.001
Age	-.00	.01	30	-.73	.469
Second 2	-.03	.11	980	-.24	.809
Second 3	-.22	.12	980	-1.88	.061
Second 4	-.27	.12	980	-2.27	.023
Second 5	-.12	.12	980	-1.03	.304
Second 6	.06	.12	980	.53	.598
Second 7	-.12	.11	980	-1.09	.276
Category-Exciting	-.03	.02	980	-1.55	.122
Category-Adorable	-.02	.02	980	-.78	.438
Category-Sympathetic	.01	.02	980	.35	.728
Category-Delicious	-.03	.02	980	-1.40	.161
Age:Second 2	.00	.01	980	.56	.573
Age:Second 3	.02	.01	980	2.68	.007
Age:Second 4	.02	.01	980	2.91	.004
Age:Second 5	.01	.01	980	1.58	.114
Age:Second 6	-.00	.01	980	-.10	.920
Age:Second 7	.01	.01	980	1.49	.136
Random effects	Std. Deviation				
(Intercept)	.05				
Residual	.19				
Number of observations: 1028; number of participants: 32					

Table S5 Median split post-hoc analysis of age based on the best-fit model in Experiment 2

12-17-month-olds					
Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	.50	.03	486	14.87	<.001
Second 2	.02	.03	486	.58	.559
Second 3	.02	.03	486	.54	.589
Second 4	.00	.03	486	.09	.932
Second 5	.01	.03	486	.21	.837
Second 6	.04	.03	486	1.28	.200
Second 7	-.02	.03	486	-.55	.586
Category-Exciting	-.02	.03	486	-.80	.425
Category-Adorable	.01	.03	486	.28	.783
Category-Sympathetic	.03	.03	486	.93	.354
Category-Delicious	-.02	.03	486	-.70	.483
Random effects	Std. Deviation				
(Intercept)	.06				
Residual	.19				
Number of observations: 512; number of participants: 16					
18-23-month-olds					
Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	.48	.03	490	16.85	<.001
Second 2	.05	.03	490	1.87	.062
Second 3	.16	.03	490	5.38	<.001
Second 4	.14	.03	490	4.48	<.001
Second 5	.12	.03	490	3.89	<.001
Second 6	.06	.03	490	2.06	.040
Second 7	.10	.03	490	3.44	<.001
Category-Exciting	-.04	.03	490	-1.21	.227
Category-Adorable	-.03	.03	490	-1.23	.220
Category-Sympathetic	-.01	.03	490	-.38	.705
Category-Delicious	-.03	.03	490	-1.24	.217
Random effects	Std. Deviation				
(Intercept)	0.02				
Residual	0.19				
Number of observations: 516; number of participants: 16					

Table S6 Results replicating 18-23-month-olds in Experiment 2

Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	0.45	0.03	527	16.70	<.001
Second 2	0.06	0.02	527	2.47	0.014
Second 3	0.08	0.03	527	3.15	0.002
Second 4	0.13	0.03	527	5.09	<.001
Second 5	0.05	0.03	527	1.94	0.053
Second 6	0.11	0.03	527	4.10	<.001
Second 7	0.11	0.02	527	4.86	<.001
Category-Exciting	-0.01	0.03	527	-0.41	0.685
Category-Adorable	-0.01	0.03	527	-0.23	0.821
Category-Sympathetic	0.05	0.03	527	1.88	0.060
Category-Delicious	0.04	0.03	527	1.33	0.184
Random effects	Std. Deviation				
(Intercept)	0.03				
Residual	0.17				
Number of observations: 553; number of participants: 16					

Table S7 Model selection in Experiment 3

Model		df	AIC	BIC	Log likely-hood	Test	Likeli-hood ratio	<i>p</i> -value
1	Age+Time+Category+(1 Subject)	14	-177.4	-107.5	102.7	-	-	-
2 best fit	Age+Time+Category+(1 Subject) [corAR(1)]	15	-263.1	-188.2	146.6	1 vs 2	87.7	<.001
3	Age+Time+Category+Age:Time +(1 Subject) [corAR(1)]	21	-255.8	-150.9	148.9	2 vs 3	4.7	.583
4	Age+Time+Category+ Age:Category+(1 Subject) [corAR(1)]	19	-256.1	-161.1	147.0	2 vs 4	.925	.921
5	Age+Time+Category+ +Time:Category+(1 Subject) [corAR(1)]	39	-229.0	-34.1	153.5	2 vs 5	13.9	.949
6	Age+Time+Category+ +Age:Time:Category+(1 Subject) [corAR(1)]	Singularity in backsolve at level 0, block 1						

Table S8 Summary of the best-fit model in Experiment 3

Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	.52	.09	1052	5.91	<.001
Age	.00	.01	30	.18	.857
Second 2	-.04	.02	1052	-1.76	.078
Second 3	-.02	.02	1052	-.68	.494
Second 4	-.01	.02	1052	-.42	.674
Second 5	.04	.02	1052	1.64	.100
Second 6	-.01	.02	1052	-.61	.545
Second 7	.02	.02	1052	.71	.480
Category-Exciting	.00	.03	1052	.12	.901
Category-Adorable	-.01	.03	1052	-.45	.654
Category-Sympathetic	.02	.03	1052	.87	.385
Category-Delicious	.01	.03	1052	.37	.710
Random effects	Std. Deviation				
(Intercept)	.00				
Residual	.22				
Number of observations: 1094; number of participants: 32					

Table S9 Model selection in Experiment 5

Model		df	AIC	BIC	Log likely-hood	Test	Likeli-hood ratio	<i>p</i> -value
1 best fit	Age+Congruency+Category+(1 Subject)	6	756.4	773.7	-372.2	-	-	-
2	Age+Congruency+Category+Age:Congruency+(1 Subject)	7	757.1	777.3	-371.6	1 vs 2	1.32	.251
3	Age+Congruency+Category+Congruency:Category+(1 Subject)	7	758.2	778.3	-372.1	1 vs 3	.26	.613
4	Age+Congruency+Category+Age:Category+(1 Subject)	7	756.4	776.6	-371.2	1 vs 4	2.01	.157
5	Age+Congruency+Category+Age:Congruency:Category+(1 Subject)	Singularity in backsolve at level 0, block 1						

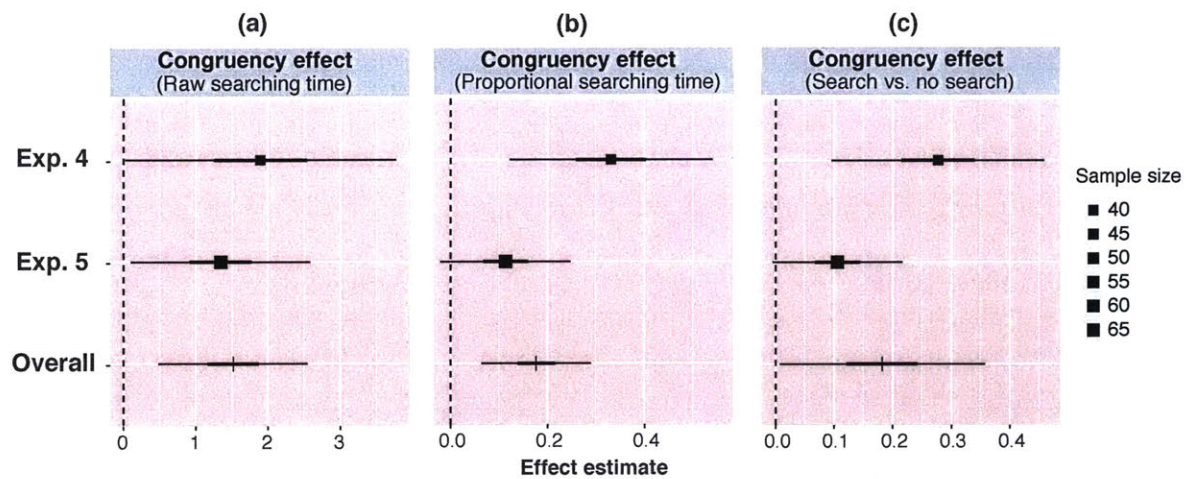
Table S10 Summary of the best-fit model in Experiment 5

Fixed effects	Estimate	Std. Error	df	<i>t</i> -value	<i>p</i> -value
(Intercept)	-6.01	4.09	65	-1.47	.146
Age	.55	.27	63	2.06	.044
Congruency-Incongruent	1.34	.63	65	2.12	.038
Category-Adorable	.36	.82	63	.45	.656
Random effects	Std. Deviation				
(Intercept)	2.07				
Residual	3.64				
Number of observations: 132; number of participants: 66					

Fig. S1 Representative eliciting cause stimuli from each category. See the main text for details.



Fig. S2 A meta-analysis looking at the effect of Congruency across Experiments 4 and 5. Three types of data were used: (a) participants' raw searching time in the Congruent and Incongruent conditions, (b) their proportional searching time in the Congruent and Incongruent conditions, (c) the number of participants searched (in contrast to did not search at all) in the Congruent and Incongruent conditions.



Appendix II. Study 2.2 Supporting Information

Text 1 Scenarios

1.1 The plane-crash scenario for Experiments 1 and 2

Participants in Experiments 1 and 2 read the following scenario:

“Grace and John are co-workers in Cleveland with an office in Boston. There are different versions of the story. In some versions of the story, Grace and John are close friends. In some versions, John has blackmailed Grace with information that would destroy her family and career, and Grace finds herself contemplating ways that she could kill John.

One day Grace is watching the news and learns that there has been an airplane crash on a Cleveland/Boston connection. There are only two direct flights between these cities and it happens that John was traveling that day. Grace doesn’t know which flight John was on. Grace quickly checks John’s itinerary. After Grace sees John’s itinerary, she might have the correct information about which flight John was on, or she might have the wrong information because John could have changed his flight.”

To make sure that the participants understood the scenario, we asked three catch questions based on its content. Participants who answered any question incorrectly were not included in the final analysis.

“1. How many direct flights are there between Cleveland and Boston?” Correct answer: Two.

“2. What does Grace do after she learns the news about the airplane crash?” Correct answer: She checks John’s itinerary.

“3. In what way might Grace be wrong about John’s flight after she sees John’s itinerary?” Correct answer: She might be wrong because John could have changed his flight.

1.2 The chemical-factory scenario for Experiments 3 and 4

Participants in Experiments 3 and 4 read the following scenario:

“Grace and John are co-workers. There are different versions of the story. In some versions of the story, Grace and John are close friends. In some versions, John has blackmailed Grace with information that would destroy her family and career, and Grace finds herself contemplating ways that she could kill John.

One day they are taking a tour of a chemical factory. When Grace goes over to the coffee machine to pour some coffee, John asks her to bring him a cup of coffee as well. There is a container with white powder by the coffee. Since they are in a chemical factory, the container might contain regular sugar, or it might contain a toxic substance left behind by a scientist that is deadly when ingested. Grace doesn’t know what is in the container, so she asks one of the scientists she sees passing by in the hall. After Grace gets the scientist’s answer, she might have the correct information about what’s in the container, or she might have the wrong information because the scientist could have been mistaken about which container she was referring to.”

Again, we asked three catch questions based on the content of the scenario, and participants who answered any question incorrectly were dropped from our study.

“1. Where are they taking a tour?” Correct answer: A chemical factory.

“2. What does Grace do to find out what's in the container?” Correct answer: She asks one of the scientists she sees passing by in the hall.

“3. In what way might Grace be wrong about what's in the container after she gets the scientist's answer?” Correct answer: She might be wrong because the scientist could have been mistaken about which container she was referring to.

Text 2 Stimuli

2.1 The photograph stimuli used in Experiments 1, 2a, 3 and 4

2.1.1 *Reaction*₁

We created one facial expression per condition based primarily on the assumptions that the fulfillment (or thwarting) of desire predicts positive (or negative) valence and the confirmation (or disconfirmation) of belief predicts the absence (or presence) of surprise. The original facial expressions used to generate these stimuli are presented in Fig. S1. They were found online as demonstrations of the six basic emotions proposed by Ekman (1992).

Aside from the two main assumptions, we did not make strong predictions about the individual facial expressions but adjusted the faces until they conformed to our intuitions that the face was plausible given the scenario. The components of each facial expression are shown in Table S1: Components (%). Specifically, when Grace's desire was thwarted, we used the expression of anger if the thwarted desire was a morally bad one (i.e., wanting John to die; Conditions 2 and 6), and used the expression of sadness if the desire was a morally good one (i.e., wanting John to be alive; Conditions 3 and 7). Even when a desire is fulfilled, a person with the morally bad desire to murder may look more aggressive than someone with a morally good desire. So in Condition 1, we added some anger in the eye region of Grace's face although overall the face expressed happiness. In Conditions 4 and 6, Grace believed that John would be alive and later she confirmed that he was alive. Presumably Grace's emotional reaction to this event was more neutral than in the conditions in which John died (e.g., Condition 7) or was unexpectedly alive. There was no "neutral" expression in the original stimulus set (see Fig. S1) so we neutralized the happy expression in Condition 4 and the angry expression in Condition 6 with emotional expressions of the opposite valence (i.e., some sadness in Condition 4 and some happiness in Condition 6). The fact that mixing the dominant expression with some of the

opposite valence generated a more neutral face was confirmed by people's perceptual judgments, as shown in Table S1: Perception results (0-100). For example, although there is 40% happiness in the expression for Condition 6, participants' ratings on the Happy Scale for this face were lower than those for all other negative expressions.

Note that these were all minor adjustments we made to create a plausible expression for each condition. To make sure that any results from this set of stimuli were not due to its arbitrary features, we used unmorphed facial expressions in Experiment 2b and movie stimuli in Experiment 3 Supplementary to replicate those results.

2.1.2 *Reaction₀*

For simplicity, we used the same face image for each pair of conditions sharing the same mental state, corresponding to the face used to indicate *Reaction₁* in the condition where the expected and actual outcomes match (e.g., *Reaction₀* in Conditions 1 and 2 uses *Reaction₁* from Condition 1; *Reaction₀* in Conditions 5 and 6 uses *Reaction₁* from Condition 6; see Fig. 2(a)).

2.2 The movie stimuli used in Experiment 3 Supplementary

The movie stimuli were generated by an actor with experience in university theater performances. He knew nothing about this study and was only told that we would like him to act out a short script. We gave him the chemical-factory scenario but replaced the protagonist's name "Grace" with "you". We told him that there were eight versions of this story, corresponding to whether he wanted his colleague John to die or live, whether he believed the powder was poison or sugar, and whether John died or lived. We asked him to respond to the final outcome of each story and we filmed his emotional reaction to the outcome. The lengths of the original video clips ranged from 8-13 seconds. We clipped them into 6-second video clips, containing most of the information in his reactions.

Text 3 Tasks

After reading either the plane-crash or the chemical-factory scenario, each participant was assigned to one of four tasks: (1) the prior task (and the action likelihood task for the chemical-factory scenario only), (2) the *Reaction*₀ likelihood task, (3) the *Reaction*₁ likelihood task, or (4) the test task: the mental state inference task (see Fig. S2). For a given task, the measures were within-subjects. All of the responses were elicited on a 0-100 scale, where 0 indicated “completely implausible” and 100 indicated “completely plausible,” except that for the action likelihood task, 0 indicated “definitely would not put the powder in the coffee” and 100 indicated “definitely would put the powder in the coffee.” The tasks were largely the same for both scenarios, but differed slightly depending on the details of the scenario. Both versions are described below.

3.1 The prior task (and the action likelihood task for the chemical-factory scenario only)

Given either the plane-crash or the chemical-factory scenario, participants were asked to judge the prior plausibility of the four possible combinations of Grace’s desire and belief: “*Based on the information you have now, how plausible do you think you would rate a version of the story described below?*”

For the plane-crash scenario: “*Grace wants John to [die/live], and after she sees John’s itinerary, she believes John was [on/not on] the crashed flight.*”

For the chemical-factory scenario: “*Grace wants John to [die/live], and after she gets the scientist’s answer, she believes the powder is [poison/sugar].*” Following the chemical-factory scenario only, participants were then asked to judge the action likelihood given each of the four mental states by responding to the prompt: “*Given that Grace wants John to [die/live] and*

believes the powder is [poison/sugar], how likely is it that she will put the powder in John's coffee?"

3.2 The *Reaction₀* likelihood task

In this task, participants were given different values of Grace's desire and belief and judged the plausibility of all four facial expressions when Grace expected but had not observed the outcome.

Specifically, following the plane-crash scenario, participants read: "*Grace wants John to [die/live], and after she sees John's itinerary, she believes John was [on/not on] the crashed flight. How plausible is each of Grace's responses?"*

Following the chemical-factory scenario, participants read: "*Grace wants John to [die/live], and after she gets the scientist's answer, she believes the powder is [poison/sugar]. Grace **puts** the powder in John's coffee and gives the coffee to John. Then Grace turns away and shows a facial expression. How plausible is each of these facial expressions?"*

3.3 *Reaction₁* likelihood task

In this task, participants were told Grace's desire and belief, and the final outcome. They then assessed the plausibility of all eight emotional expressions responding to the final outcome.

Following the plane-crash scenario, participants read: "*Grace wants John to [die/live], and after she sees John's itinerary, she believes John was [on/not on] the crashed flight. Later, Grace confirms from the newspaper that John [was on the crashed flight and has died/was not on the crashed flight and is still alive]. How plausible is each of Grace's responses?"*

Following the chemical-factory scenario, participants read: "*Grace wants John to [die/live], and after she gets the scientist's answer, she believes the powder is [poison/sugar]. Grace **puts** the powder in John's coffee and gives the coffee to John. Soon after the tour, Grace*

*gets to know that John [has **died** and it turns out that the powder is **poison/is ok and it turns out that the powder is sugar**]. How plausible is each of Grace's responses?"*

3.4 Mental state inference task

Given the plane-crash scenario (in which Grace's beliefs and desires were not specified), participants in Experiment 1 were told the final outcome and Grace's reaction to that outcome, i.e. "*Later, Grace confirms from the newspaper that John [was **on the crashed flight** and has **died/was not on the crashed flight and is still alive**]. Grace's response is this: [Reaction₁].*" Participants were then asked to infer Grace's desire and belief, i.e. "*Consider in this version whether Grace wants John to live or to die, and whether Grace believed John was on the crashed flight or not, before she gets final confirmation from the newspaper. Rate the plausibility of each of the four possible combinations of her desire and belief.*"

Participants in Experiment 2 performed the same judgments as in Experiment 1 with one exception. They were given an additional facial expression after Grace checks John's itinerary but before she gets the final outcome from the newspaper, i.e. "*After Grace sees John's itinerary, her response is this: [Reaction₀].*"

Given the chemical-factory scenario (in which Grace's belief and desire are not specified), participants in Experiment 3 were given Grace's action, the outcome of her action, and her emotional reaction to that outcome, i.e. "*Grace **puts** the powder in John's coffee and gives the coffee to John. Soon after the tour, Grace gets to know that John [has **died** and it turns out that the powder is **poison/is ok and it turns out that the powder is sugar**]. Grace's response is this: [Reaction₁].*" Participants were then asked to infer Grace's desire and belief, i.e. "*Consider in this version whether Grace wants John to live or to die, and whether Grace*

believed the powder is sugar or poison, before she gets to know the outcome. Rate the plausibility of each of the four possible combinations of her desire and belief.”

Participants in Experiment 4 performed the same judgments as in Experiment 3 but were given an additional facial reaction after Grace acted and before she knew the outcome, i.e.

“Grace turns away and shows a facial expression: [Reaction₀].”

Text 4 Analyses of the judgments used to calibrate the model

All judgments used to calibrate the model were analyzed with both mixed effects models and One Sample t-tests (just as the test responses were). The detailed analyses reflect the results reported in the main text.

4.1 Prior

The ratings for the prior probability of each mental state combination was comparable across the two scenarios (the plane-crash scenario used in Experiments 1-2 and the chemical-factory scenario used in Experiments 3-4). Given the plane-crash scenario, the main effect of Mental State trended towards significance ($F(3, 168) = 2.17, p = .093$). Further analyses showed that no pairwise comparison was significant (all $|z|s < 2.20$, all $ps > .168$; p values were corrected with Bonferroni method throughout). One Sample t-tests revealed that Die&Safe was rated significantly below 50 ($t(56) = -2.84, p = .025$); no other mental state differed significantly from 50 (all $|t|s < 0.16$, all $ps = 1.000$). See Fig. 2(b)(i).

Given the chemical-factory scenario, the main effect of Mental State was significant ($F(3, 164) = 6.62, p < .001$). Post-hoc analyses showed that the mental state Live&Sugar was rated significantly higher than the mental states Die&Sugar ($z = 4.11, p < .001$) and Live&Poison ($z = 3.55, p = .002$); no other pair-wise comparison was significant (all $|z|s < 2.32$, all $ps > .122$). One

Sample t-tests showed that Live&Sugar was rated significantly above 50 ($t(54) = 3.04, p = .015$); no other mental state differed significantly from 50 (all $|t|s < 2.30$, all $ps > .103$). See Fig. 2(c)(i).

4.2 Action likelihood (for the chemical-factory scenario only)

Fig. 2(c)(ii) presents participants' ratings of the likelihood of Grace's action (i.e., putting the powder in John's coffee). The main effect of Mental State was significant ($F(3, 168) = 36.03, p < .001$). Post-hoc analyses showed that participants rated the action as significantly more likely given the two mental states in which the action would fulfill Grace's desire given her belief (Die&Poison and Live&Sugar) than given the two mental states where the action would not fulfill her desire given her belief (Die&Sugar and Live&Poison; all $zs > 2.63$, all $ps < .052$). The action likelihood given the two plausible mental states did not differ from each other ($z = 1.23, p = 1.000$), however, participants rated the action as less plausible given Live&Poison than Die&Sugar ($z = 5.69, p < .001$). Similar results were found using One Sample t-tests. The action was rated significantly above 50 for the two mental states (Live&Sugar and Die&Poison) where the action would fulfill her desire given her belief (both $ts > 3.04$, both $ps < .014$). The action likelihood given the mental state Die&Sugar was rated non-significantly different from 50 ($t(56) = -0.33, p = .741$), and the action likelihood given the mental state Live&Poison was rated significantly below 50 ($t(56) = -11.941, p < .001$).

4.3 *Reaction*₀ likelihood

The *Reaction*₀ likelihood elicited by both the plane-crash and chemical-factory scenarios was consistent with participants judging that Grace would feel positive if she believed that her desire would be fulfilled given her belief, and that she would feel negative if she believed her desire would not be fulfilled. See Fig. 2(b)(ii), Fig. 2(c)(iii), and Fig. 5(b)(i).

In Experiment 2a (see Fig. 2(b)(ii)), the judgments were elicited by the plane-crash scenario. There was no main effect of Condition ($F(3, 735) = 1.25, p = .290$) but a significant main effect of Mental State ($F(3, 735) = 7.21, p < .001$) and a significant interaction between Mental State and Condition ($F(3, 735) = 147.38, p < .001$). Further analyses showed that the main effect of Mental State was significant in each condition (all $F_s > 74.06$, all $p_s < .001$). The positive expression used in Conditions 1 and 2 was rated more plausible given the two mental states where Grace expected her desire to be fulfilled (Die&Crash, Live&Safe) than the two mental states where she expected it would not (Die&Safe, Live&Crash; all $|z|s > 15.12$, all $p_s < .001$). Participants did not further distinguish the two plausible mental states ($z = .15, p = .882$). The same pattern was found for the positive expression used in Conditions 3 and 4. The negative expression used in Conditions 5 and 6 was rated more plausible given the two mental states where Grace did not expect her desire to be fulfilled (Die&Safe, Live&Crash) than the two mental states where she expected it would be (Die&Crash, Live&Safe; all $z_s > 8.34$, all $p_s < .001$). Conditions 7 and 8 showed the same pattern (all $z_s > 8.89$, all $p_s < .001$). However, for both negative expressions, participants further differentiated the two plausible mental states: the facial expression (i.e., anger) used in Conditions 5 and 6 was rated more plausible given the mental state Die&Safe ($z = 3.11, p = .030$) than Live&Crash, and the one (i.e., sadness) used in Conditions 7 and 8 was rated more plausible given Live&Crash than Die&Safe ($z = 4.19, p < .001$). One Sample t-tests showed similar patterns. The two positive facial expressions were rated significantly above 50 given the two mental states where Grace expected her desire to be fulfilled (all $t_s > 7.27$, all $p_s < .001$) but significantly below 50 given the two mental states where her desire would not (all $t_s < -11.46$, all $p_s < .001$). The angry expression used in Conditions 5 and 6 was uniquely rated significantly above 50 given the mental state Die&Safe ($t(49) = 8.85, p$

< .001), and the sad expression used in Conditions 7 and 8 was uniquely rated significantly above 50 given the mental state Live&Crash ($t(49) = 8.85, p < .001$). Participants' tendency to make relatively fine-grained discrimination between the two negative expressions is consistent with our intuition in designing the stimuli: when a morally bad desire is thwarted, a person is more likely to feel angry than sad; if a morally good desire is thwarted, vice versa.

Experiment 2b also used the plane crash scenario but the facial expressions were de-morphed. These generated roughly the same results (see Fig. 5(b)(i)) as in Experiment 2a. There was a non-significant trend towards a main effect of Mental State ($F(3, 627) = 2.42, p = .065$) but a significant main effect of Condition ($F(2, 627) = 33.83, p < .001$) and a significant interaction between Condition and Mental State ($F(6, 627) = 279.64, p < .001$). Further analyses showed that the main effect of Mental State was significant in all conditions (all F s > 74.06, all p s < .001). Participants judged the positive face used in Conditions 1-4 more plausible given the two mental states where Grace expected to fulfill her desire than the two mental states where she expected she would not (all z s > 23.23, all p s < .001). Participants did not further distinguish the two plausible mental states ($z = .71, p = 1.000$). Again, participants had more fine-grained judgments of the two negative faces; they judged the angry expression more plausible given Die&Safe than the other mental states, and the sad expression more plausible given Live&Crash than the other mental states (all z s > 10.35, all p s < .001). One Sample t-tests showed similar patterns. The two positive facial expressions were rated significantly above 50 given the two mental states where Grace expected to fulfill her desire (both t s > 14.07, all p s < .001) but significantly below 50 given the two mental states where she expected she would not (both t s < -19.98, all p s < .001). The negative expression used in Conditions 5 and 6 and the one used in

Conditions 7 and 8 were uniquely rated significantly above 50 given Die&Safe and Live&Crash, respectively ($t_{5,6}(57) = 7.74, p_{5,6} < .001$; $t_{7,8}(57) = 12.99, p_{7,8} < .001$).

These results were replicated in Experiment 4 with the chemical-factory scenario (see Fig. 2(c)(iii)). There were significant main effects of Mental State ($F(3, 842) = 24.69, p < .001$) and Condition ($F(3, 842) = 31.42, p < .001$). The interaction between Mental State and Condition was also significant ($F(9, 842) = 90.21, p < .001$). Further analyses showed that the main effect of Mental State was significant in all conditions (all $F_s > 33.74$, all $p_s < .001$). Pairwise comparisons revealed that the positive facial expression used in Conditions 1 and 2 was judged more plausible given the two mental states where Grace expected to fulfill her desire than the two mental states where she expected she would not (all $|z|s > 9.37$, all $p_s < .001$). Participants did not further differentiate between the two plausible mental states ($z = 2.99, p = .067$). See Fig. 2(b)(iii). The same pattern was found for the positive facial expression used in Conditions 3 and 4. The angry expression used in Conditions 5 and 6 was rated more plausible given the mental state Die&Sugar than all other mental states (all $z_s > 4.72$, all $p_s < .001$), and the sad facial expression used in Conditions 7 and 8 was rated more plausible given the mental state Live&Poison than all other mental states (all $z_s > 5.85$, all $p_s < .001$). One Sample t-tests showed roughly the same results. The two positive facial expressions were rated significantly above 50 given the two mental states in which Grace expected to fulfill her desire (all $t_s > 5.08$, all $p_s < .001$) but significantly below 50 given the two mental states in which she expected she would not (all $t_s < -3.55$, all $p_s < .013$). The angry expression used in Conditions 5 and 6 was uniquely rated significantly above 50 given the mental state Die&Sugar ($t(54) = 3.25, p = .032$). The sad expression used in Conditions 7 and 8 was rated non-significantly different from 50 when Grace

expected she would not fulfill her desire (both $|t|s < 2.94$, both $ps > .133$), but significantly below 50 given the other two mental states (both $ts < -14.90$, both $ps < .001$).

4.4 *Reaction*₁ likelihood

As noted in the main text, with the exception of the unmorphed surprised face in Experiment 2, participants found the facial expressions most plausible given the desires used to generate the expressions but did not distinguish the plausibility of the expressions based on the target belief. See Fig. 2(b)(iii), Fig. 2(c)(iv) and Fig. 5(b)(ii). For example, as shown in Fig. 2(b)(iii): Condition 1, participants found the facial expression plausible given both the mental states Die&Crash and Die&Safe, although the stipulated mental state was Die&Crash.

Specifically, the *Reaction*₁ likelihood in Experiment 1 (see Fig. 2(b)(iii)) was elicited by the plane-crash scenario. There was no main effect of Mental State ($F(3, 1356) = .96, p = .410$) but a significant main effect of Condition ($F(7, 1356) = 3.72, p < .001$) and a significant interaction between Mental State and Condition ($F(21, 1356) = 91.72, p < .001$). Further analyses showed that the main effect of Mental State was significant in all conditions (all $Fs > 48.16$, all $ps < .001$). Participants rated the facial expression used in each condition more plausible given the two mental states with the correct desire than the two mental states with the incorrect desire (all $zs > 7.02$, all $ps < .001$) but did not distinguish the two beliefs (all $|z|s < 1.83$, all $ps = 1.000$). Similarly, One Sample t-tests showed that given the two mental states with the correct desire, the facial expressions were rated significantly above 50 (all $ts > 3.34$, all $ps < .054$) with a non-significant trend in the same direction in Condition 2 (both $|t|s < .190$, both $ps = 1.000$); the facial reaction was rated significantly below 50 given the two mental states with the incorrect desire: both $ts < -9.57$, all $ps < .001$).

Similar results were found using the chemical-factory scenario in Experiment 3. There was no main effect of Mental State ($F(3, 1647) = .96, p = .412$) but a significant main effect of Condition ($F(7, 1647) = 20.14, p < .001$) and a significant interaction between Mental State and Condition ($F(21, 1647) = 65.07, p < .001$). Further analyses showed that the main effect of Mental State was significant in all conditions (all F s > 31.24 , all p s $< .001$). Only in Condition 2 did participants rate the facial expression more plausible given the target mental states than the other three mental states (all z s > 3.38 , all p s $< .038$). In the remaining conditions, they rated each facial reaction more plausible given the two mental states with the correct desire than the two mental states with the incorrect desire (all z s > 4.79 , all p s $< .001$) but did not distinguish between the two beliefs (all $|z|$ s < 2.95 , all p s $> .150$). Similarly, One Sample t -tests showed that the facial reactions were rated significantly above 50 given the two mental states with the correct desire in Conditions 1, 3, 4, 6, 7 (all t s > 3.44 , all p s $< .036$) with a similar trend in Conditions 5, 6, 8 (i.e., either significantly above 50 (t s > 4.35 , p s $< .001$) or equal to 50 ($|t|$ s < 3.14 , p s $> .088$)); the mental states with the incorrect desire were rated significantly below 50 (all t s < -7.06 , all p s $< .001$). In Condition 2, the facial reaction was rated non-significantly different from 50 given the target mental state ($t(51) = .83, p = 1.000$), but significantly below 50 given all other mental states (all t s < -6.35 , all p s $< .001$).

In Experiment 2b⁴, all the facial expressions were unmorphed. There were both significant main effects of Mental State ($F(3, 1311) = 3.83, p = .010$) and Condition ($F(5, 1311) = 2.99, p = .011$) and a significant interaction between Mental State and Condition ($F(15, 1311)$

⁴ Because some of the facial expressions were identical across conditions (e.g., *Reaction*₀ in Conditions 1,2 and 3,4; *Reaction*₁ to the “live” outcome in Condition 2b and 8b; and *Reaction*₁ to the “die” outcome in Condition 1 and 5a), we only measured likelihood judgments for each pair of identical expressions once. However, for completeness in the table (see Fig. 5(a)), we report the likelihood data for all conditions, repeating the identical ratings as needed.

= 82.37, $p < .001$). Further analyses showed that the main effect of Mental State was significant in each condition (all F s > 3.38, all p s < .020).

For all the valenced expressions (i.e., used in Conditions 1, 2a, 3a, 4, 5a, 6, 7, 8a), the results replicated those found in Experiments 1 and 3. Participants rated these expressions more plausible given the two mental states with the correct desire than the two mental states with the incorrect desire (all z s > 8.47, all p s < .001) but did not further discriminate the two beliefs (all $|z$ s < 2.27, all p s > .839). One Sample t -tests showed converging results. All the valenced expressions were rated significantly above 50 given the two mental states with the correct desire (all t s > 3.84, all p s < .011).

Given the pure surprised faces, participants were better able to distinguish Grace's beliefs. In Conditions 2b and 8b, participants rated the surprised expression as more plausible given the two mental states with the correct belief than the two mental states with the incorrect belief (all z s > 5.51, all p s < .001). For the surprised expression used in Conditions 3b and 5b, however, participants' ratings did not significantly differ from each other given any pair of the four mental states (all $|z$ s < 3.09, all p s > .072). As noted in the main text, this may be because surprise is considered a normal response to someone's death even when the death is anticipated. Similarly, One-Sample t -tests showed that there was a non-significant trend for the surprised reaction in Conditions 2b and 8b to be rated above 50 given the two mental states with the correct belief (both $|t$ s < 1.13, both p s = 1.000); the ratings given the two mental states with the incorrect belief were significantly below 5 (both t s < -6.60, both p s < .001). The surprised expression in Conditions 3b and 5b did not differ significantly from 50 given any of the four mental states (all $|t$ s < 2.80, all p s > .224).

Text 5 Equations

5.1 Experiment 1:

$$P(\text{Belief}, \text{Desire} | \text{Outcome}, \text{Reaction}_1) \propto$$

$$P(\text{Reaction}_1 | \text{Belief}, \text{Desire}, \text{Outcome}) \times$$

$$P(\text{Belief}, \text{Desire})$$

5.2 Experiment 2:

$$P(\text{Belief}, \text{Desire} | \text{Outcome}, \text{Reaction}_0, \text{Reaction}_1) \propto$$

$$P(\text{Reaction}_1 | \text{Belief}, \text{Desire}, \text{Outcome}) \times$$

$$P(\text{Reaction}_0 | \text{Belief}, \text{Desire}) \times$$

$$P(\text{Belief}, \text{Desire})$$

5.3 Experiment 3:

$$P(\text{Belief}, \text{Desire} | \text{Action}, \text{Outcome}, \text{Reaction}_1) \propto$$

$$P(\text{Reaction}_1 | \text{Belief}, \text{Desire}, \text{Action}, \text{Outcome}) \times$$

$$P(\text{Action} | \text{Belief}, \text{Desire}) \times P(\text{Belief}, \text{Desire})$$

5.4 Experiment 4:

$$P(\text{Belief}, \text{Desire} | \text{Action}, \text{Outcome}, \text{Reaction}_0, \text{Reaction}_1) \propto$$

$$P(\text{Reaction}_1 | \text{Belief}, \text{Desire}, \text{Action}, \text{Outcome}) \times$$

$$P(\text{Reaction}_0 | \text{Belief}, \text{Desire}, \text{Action}) \times$$

$$P(\text{Action} | \text{Belief}, \text{Desire}) \times P(\text{Belief}, \text{Desire})$$

Text 6 Experiment 2b Supplementary

A group of participants ($n = 61$) were asked to rate on a Likert scale of 0 (neutral) to 100 (extremely intense) to what degree Grace's facial reactions (the photograph stimuli) conveyed each of the following emotions: happy, surprise, sad, anger, fear, disgust, and an additional emotion unhappy counting all the negative valences. See Table S1: Perception results (0-100).

We found that the sub-components of the facial reactions significantly predicted participants' perceptions of the corresponding emotions (surprise: $\beta = .573$, $t(486) = 8.80$, $p < .001$; happy: $\beta = .768$, $t(486) = 25.97$, $p < .001$; sad: $\beta = .500$, $t(486) = 11.98$, $p < .001$; anger: $\beta = .254$, $t(486) = 5.40$, $p < .001$; unhappy: $\beta = .542$, $t(486) = 13.09$, $p < .001$). The sub-components of the facial reactions also explained a significant proportion of variance in participants' ratings on the corresponding emotions (surprise: $R^2 = .138$, $F(1, 486) = 77.49$, $p < .001$; happy: $R^2 = .581$, $F(1, 486) = 674.27$, $p < .001$; sad: $R^2 = .228$, $F(1, 486) = 143.58$, $p < .001$; anger: $R^2 = .238$, $F(1, 486) = 29.11$, $p < .001$; unhappy: $R^2 = 0.261$, $F(1, 486) = 171.32$, $p < .001$).

These judgments were also used as perceptual features of the stimuli when we evaluated the Event-Features Model in the main text. For the same purpose, a separate group of participants ($n = 58$) made the same judgments of the prototypical facial expressions used in Experiment 2b. These judgments are reported in Table S2.

Text 7 Experiment 3 Supplementary

7.1 *Reaction₁* likelihood (movie stimuli)

Fig. S3(a) reports participants' likelihood judgments of the movie stimuli given Grace's belief, desire, and the outcome specified in each condition. The pattern of results was similar to those elicited in other experiments (e.g., see Fig. 2(b)(iii), Fig. 2(c)(iv)). There was a non-significant trend towards a main effect of Mental State ($F(3, 1550) = 2.41$, $p = .065$) but a significant main effect of Condition ($F(7, 1550) = 2.60$, $p = .012$) and a significant interaction between Mental State and Condition ($F(21, 1550) = 33.63$, $p < .001$). Further analyses showed that the main effect of Mental State was significant in all conditions (all F s > 2.94 , all p s $< .035$). As with the photograph stimuli, participants rated each facial expression more plausible given the two mental states with the correct desire than the two mental states with the incorrect desire

(all z s > 4.10 , all p s $< .001$) but did not distinguish between the two beliefs (all $|z|$ s < 2.97 , all p s $> .143$). The exception was Condition 4, in which participants did not distinguish either the desires or beliefs and no mental state rating differed significantly from any other (all $|z|$ s < 2.78 , all p s $> .258$). Similarly, One Sample t -tests showed that the facial expressions were rated significantly above 50 given the two mental states with the correct desire in Conditions 1, 3, 7, 8 (all t s > 3.42 , all p s $< .040$) with a non-significant trend in the same direction in Conditions 2, 5, 6 (i.e., either above 50 (t s > 4.67 , p s $< .001$) or equal to 50 ($|t|$ s < 3.33 , p s $> .052$); the ratings given the mental states with the incorrect desire were significantly below 50: all t s < -4.637 , all p s $< .001$). In Condition 4, the facial expression did not differ significantly from 50 given any of the four mental states (all $|t|$ s < 1.89 , all p s = 1.000).

7.2 Mental state inferences (movie stimuli)

A separate group of participants rated the plausibility of the different combinations of beliefs and desires based on Grace's action, the outcome, and the emotional reaction specified in each condition. See Fig. S3(b). The results converged with those using the photograph stimuli in Experiment 3 (see Fig. 3(c)). There was no main effect of Condition ($F(7) = 1.02$, $p = .417$) but a significant main effect of Mental State ($F(3) = 199.23$, $p < .001$) and a significant interaction between Condition and Mental State ($F(21) = 29.00$, $p < .001$). The main effect of Mental State was significant in all conditions (all F s > 11.72 , all p s $< .001$). We further looked at whether participants rated the target mental state significantly higher than the other three mental states in each condition. Consistent with the results from the photograph stimuli, participants rated the target mental state significantly higher than all other mental states (all z s > 6.65 , all p s $< .001$) in the congruent conditions (Conditions 1-4). Also consistent with the results from the photograph stimuli, in the incongruent conditions (Conditions 5-8) participants rated the mental state with

the correct desire and the belief congruent with the desire given the action significantly higher than the target mental state (all z s > 3.61 , all p s $< .007$). Similarly, One Sample t-tests showed that participants uniquely rated the target mental state significantly above 50 in Conditions 1-3 ($t_1(51) = 18.39, p_1 < .001$; $t_2(51) = 7.83, p_2 < .001$; $t_3(51) = 3.93, p_3 = .008$) with a non-significant trend in the same direction in Condition 4 ($t(51) = 2.98, p = .140$); the other three mental states were rated significantly below 50 (all t s < -5.57 , all p s $< .001$). In the incongruent conditions, participants rated the mental state with the correct desire and the belief congruent with the desire given the action significantly above 50 in Condition 5 ($t(51) = 5.96, p < .001$) and show a non-significant trend in the same direction in Conditions 6, 7, and 8 ($t_6(51) = 1.62, p_6 = 1.000$; $t_7(51) = 2.09, p_7 = 1.000$; $t_8(51) = 1.958, p_8 = 1.000$); all other mental states were rated equal to or significantly below 50 (all t s < -2.42 , all p s $< .621$)).

Model predictions for people's mental state inferences were generated according to the Equation in SI Text 5.3 using the *Reaction*₁ likelihood from the movie stimuli as well as data from the prior and action likelihood tasks. Similar to the photograph stimuli, the model for the movie stimuli assigned the highest posterior probability to the mental state with the desire favored by the *Reaction*₁ likelihood function and the belief congruent with the desire due to the action likelihood function. For instance, in Condition 1, the observed facial reaction favors the mental states Die&Poison and Die&Sugar, while the observed action favors the mental states Die&Poison and Live&Sugar, resulting in Die&Poison receiving the highest posterior probability. People's judgments correlated highly with the model predictions ($r=0.908$).

Fig. S1 Prototypical facial expressions of the six basic emotions. (As mentioned in the main text, because we were unable to track down the copyright permissions for the actual photographs used, figures in this paper show hand drawn pencil sketches from those photographs.)

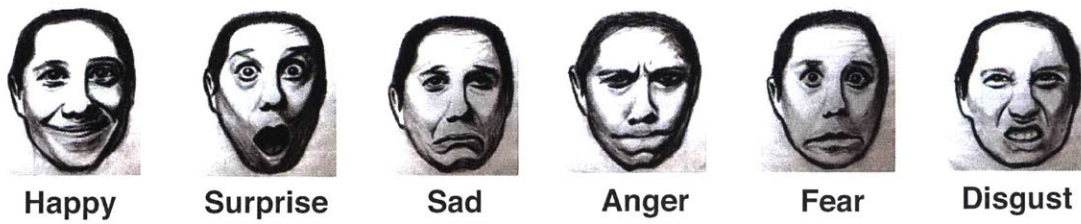


Fig. S2 The structure of the tasks

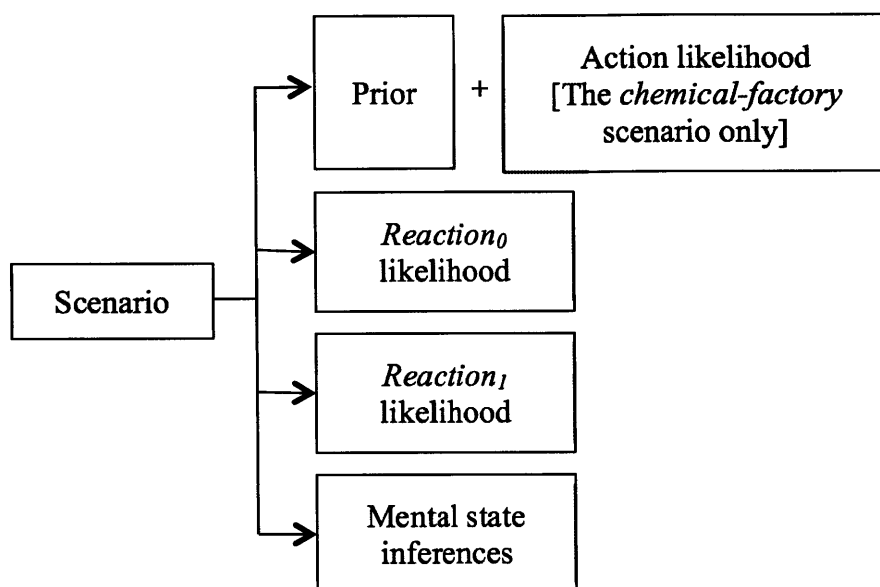
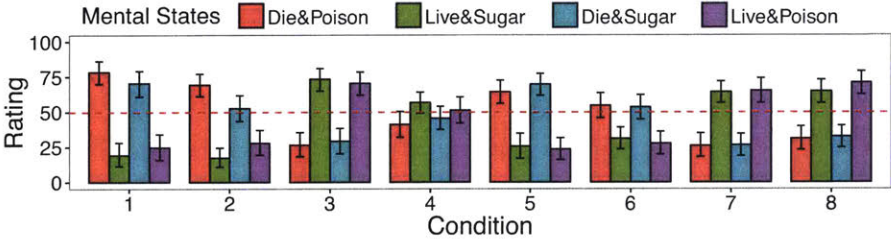


Fig. S3 The likelihood of the movie stimuli, people’s mental state inferences and model predictions in Experiment 3 Supplementary. Error bars indicate 95% confidence intervals.

Experiment 3 Supplementary

(a) Likelihood of Reaction1 (motive stimuli): $\hat{P}(R_1 \mid B, D, A = \text{True}, O)$



(b) Mental state inferences

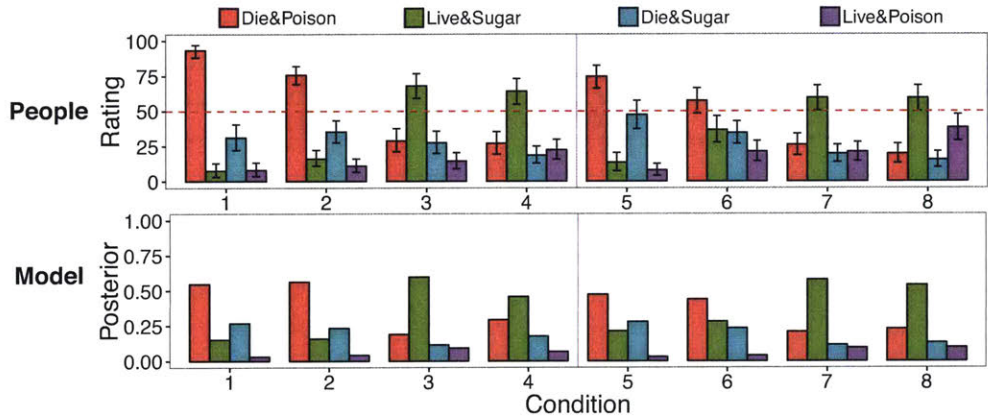


Table S1 The creation and assessment of the photograph stimuli. (See SI Text 2.1.1 for detailed explanation of the components of these faces.)













Desire&Belief		Die&Crash		Live&Safe		Die&Safe		Live&Crash	
Outcome		Die	Live	Die	Live	Die	Live	Die	Live
Reaction ₁									
Condition		1	2	3	4	5	6	7	8
Components (%)		Mouth: <u>Happy</u> 100 Eye: <u>Happy</u> 60 <u>Anger</u> 40	<u>Anger</u> 50 <u>Surprise</u> 50	<u>Sad</u> 50 <u>Surprise</u> 50	<u>Happy</u> 77 <u>Sad</u> 23	<u>Happy</u> 80 <u>Surprise</u> 20	<u>Anger</u> 60 <u>Happy</u> 40	<u>Sad</u> 100	<u>Happy</u> 60 <u>Surprise</u> 40
Perception results (0-100)	Happy	80	11	8	75	74	18	8	47
	Surprise	23	42	41	22	37	19	10	59
	Sad	11	35	39	10	10	48	79	18
	Anger	12	25	20	8	10	30	13	17
	Fear	13	57	63	11	15	25	25	23
	Disgust	14	29	31	11	12	37	23	16
	Unhappy	12	40	43	9	13	45	67	19

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

Prototypical facial expressions					
		Happy	Surprise	Sad	Anger
Perception results (0-100)	Happy	83	17	5	5
	Surprise	11	88	9	11
	Sad	5	7	90	31
	Anger	5	8	13	68
	Fear	5	35	22	14
	Disgust	4	9	15	36
	Unhappy	5	13	78	59

References

- Adam, C., Herzig, A., & Longin, D. (2009). A logical formalization of the OCC theory of emotions. *Synthese*, 168(2), 201-248.
- Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal Processing of Social Information from Faces in Autism. *Journal of Cognitive Neuroscience*, 13(2), 232-240.
- Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! Responsibility judgments when cooperation breaks down. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84-89).
- Amminger, G. P., Schafer, M. R., Papageorgiou, K., Klier, C. M., Schlogelhofer, M., Mossaheb, N., ... & McGorry, P. D. (2011). Emotion Recognition in Individuals at Clinical High-Risk for Schizophrenia. *Schizophrenia Bulletin*, 38(5), 1030-1039.
- 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.
- Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.
- Astington, J. W., & Gopnik, A. (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology*, 9(1), 7-31.
- Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33(6), 717-746.
- Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111), 1225-1229.
- 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., and 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.
- Banerjee, M. (1997). Hidden emotions: Preschoolers' knowledge of appearance-reality and emotion display rules. *Social Cognition*, 15(2), 107-132.
- Banerjee, R. (2002). Children's understanding of self-presentational behavior: Links with mental-state reasoning and the attribution of embarrassment. *Merrill-Palmer Quarterly*, 48(4), 378-404.
- Banerjee, R., & Yuill, N. (1999a). Children's explanations for self-presentational behaviour. *European Journal of Social Psychology*, 29(1), 105-111.
- Banerjee, R., & Yuill, N. (1999b). Children's understanding of self-presentational display rules: Associations with mental-state understanding. *British Journal of Developmental Psychology*, 17(1), 111-124.
- 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., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child psychology and Psychiatry*, 38(7), 813-822.

- Baron-Cohen, S. (1991). Do people with autism understand what causes emotion?. *Child Development*, 62(2), 385-395.
- Barrett, L. F. (2006a). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10, 20-46.
- Barrett, L. F. (2006b). Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40(1), 35-55.
- 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*. Guilford Publications.
- Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, 20(5), 286-290.
- Barrett, L. F., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2014). The conceptual act theory: a road map. In L. F. Barrett & J. A. Russell (eds.), *The Psychological Construction of Emotion* (pp. 83-110). New York: Guilford Press.
- Bartsch, K., & Estes, D. (1996). Individual differences in children's developing theory of mind and implications for metacognition. *Learning and Individual Differences*, 8, 281-304.
- 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.
- Beatty, A. (2013). Current emotion research in anthropology: Reporting the field. *Emotion Review*, 5(4), 414-422.

- 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.
- Bergelson, E. & Swingle, D. (2012) At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253-3258.
- Berman, J. M., Chambers, C. G., & Graham, S. A. (2010). Preschoolers' appreciation of speaker vocal affect as a cue to referential intent. *Journal of Experimental Child Psychology*, 107(2), 87-99.
- 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.
- Broomfield, K. A., Robinson, E. J., & Robinson, W. P. (2002). Children's understanding about white lies. *British Journal of Developmental Psychology*, 20(1), 47-65.
- Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.
- Caron, R. F., Caron, A. J., & Myers, R. S. (1985). Do infants see emotional expressions in static faces?. *Child Development*, 56(6), 1552-1560.
- 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.
- Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science*, 323(5918), 1222-1226.

- 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.
- Cole, P. M. (1986). Children's spontaneous control of facial expression. *Child Development*, 57(6), 1309-1321.
- Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, 12(3), 426-437.
- Crivelli, C., Russell, J. A., Jarillo, S., & Fernandez-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)
- Davis, T. L. (1995). Gender differences in masking negative emotions: Ability or motivation?. *Developmental Psychology*, 31(4), 660.
- De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3), 242-249.
- De Gelder, B., de Borst, A. W., & Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2), 149-158.

- De Rose, S. J. (2005, April 21). The Compass DeRose Guide to Emotion Words. Retrieved from <http://www.deroose.net/steve/resources/emotionwords/ewords.html>
- De Rosnay, M., Fink, E., Begeer, S., Slaughter, V., & Peterson, C. (2014). Talking theory of mind talk: young school-aged children's everyday conversation and understanding of mind and emotion. *Journal of Child Language*, 41(5), 1179-1193.
- 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.
- de Wied, M., Gispen-de Wied, C., & van Boxtel, A. (2010). Empathy dysfunction in children and adolescents with disruptive behavior disorders. *European Journal of Pharmacology*, 626(1), 97-103.
- DeLoache, J. S., & LoBue, V. (2009). The narrow fellow in the grass: Human infants associate snakes and fear. *Developmental Science*, 12(1), 201-207.
- Denham, S. A., Blair, K. A., DeMulder, E., Levitas, J., Sawyer, K., Auerbach-Major, S., & Queenan, P. (2003). Preschool emotional competence: pathway to social competence? *Child Development*, 74(1), 238-256.
- 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.
- Eibl-Eibesfeldt, I. (1989). *Human ethology*. New York, NY: Aldine de Gruyter.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169-200.

- Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, 3(4), 364-370.
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124-129.
- Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1, 56-75.
- Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, 30(1), 527-554.
- El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). Flame-fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-agent systems*, 3(3), 219-257.
- Elfenbein, H. A., Beaupre, M., Levesque, M., & Hess, U. (2007). Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion*, 7(1), 131-146.
- Ellsworth, P. C. & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, H. H. Goldsmith (eds), *Handbook of Affective Sciences* (pp. 572-595). Oxford University Press.
- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429-433.
- Eugenides, J. (2002). *Middlesex*. New York: Farrar, Straus, and Giroux.
- 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.

- Feigenson, L. & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science*, 6(5), 568-584.
- Feinman, S., Roberts, D., Hsieh, K. F., Sawyer, D., & Swanson, D. (1992). A critical review of social referencing in infancy. In *Social referencing and the social construction of reality in infancy* (pp. 15-54). Springer US.
- Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expression by neonates. *Science*, 218(4568), 179-181.
- Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, 43(1), 238-252.
- 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), 998-998.
- 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.

- Frederickson, N., Petrides, K. V., & Simmonds, E. (2012). Trait emotional intelligence as a predictor of socioemotional outcomes in early adolescence. *Personality and Individual Differences*, 52(3), 323-328.
- Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.
- Gebhard, P. (2005, July). ALMA: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems* (pp. 29-36). ACM.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287-292.
- Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2263-2268).
- Gnepp, J. (1983). Children's social sensitivity: Inferring emotions from conflicting cues. *Developmental Psychology*, 19(6), 805-814.
- Gnepp, J., & Hess, D. L. (1986). Children's understanding of verbal and facial display rules. *Developmental Psychology*, 22(1), 103.
- 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.
- Gold, R., Butler, P., Revheim, N., Leitman, D. I., Hansen, J. A., Gur, R. C.,... Javitt, D. C. (2012). Auditory Emotion Recognition Impairments in Schizophrenia: Relationship to Acoustic Features and Cognition. *American Journal of Psychiatry*, 169(4), 424-432.

- Goodman, N. D., & Stuhlmuller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1), 173-184.
- Gopnik, A. & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6), 1085-1108.
- Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4), 269-306.
- Gross, A. L., & Ballif, B. (1991). Children's understanding of emotion from facial expressions and situations: A review. *Developmental Review*, 11(4), 368-398.
- Gross, D., & Harris, P. L. (1988). False beliefs about emotion: Children's understanding of misleading emotional displays. *International Journal of Behavioral Development*, 11(4), 475-488.
- Grueneisen, S., Wyman, E., & Tomasello, M. (2015). "I Know You Don't Know I Know..." Children Use Second-Order False-Belief Reasoning for Peer Coordination. *Child development*, 86(1), 287-293.
- 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. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Direction in Psychological Science*, 22(3), 186-193.
- 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.

- Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. *Neuropsychology Review*, 20(3), 290-322.
- 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. (2008). Children's understanding of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3 ed., pp. 320-331). New York: Guildford Press.
- Harris, P. L. (2016). Emotion, imagination and the world's furniture. *European Journal of Developmental Psychology*, 1-12.
- Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development*, 57(4), 895-909.
- 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.
- Harwood, M. D., & Farrar, M. J. (2006). Conflicting emotions: The connection between affective perspective taking and theory of mind. *British Journal of Developmental Psychology*, 24(2), 401-418.

- Haviland, J. M., & Lelwica, M. (1987). The induced affect response: 10-week-old infants' responses to three emotion expressions. *Developmental Psychology*, 23(1), 97-104.
- Hedger, J. A., & Fabricius, W. V. (2011). True belief belies false belief: Recent findings of competence in infants and limitations in 5-year-olds, and implications for theory of mind development. *Review of Philosophy and Psychology*, 2(3), 429-447.
- Hobson, R. P. (1986). The Autistic Child's Appraisal of Expressions of Emotion. *Journal of Child Psychology and Psychiatry*, 27(3), 321-342.
- Hoehl, S., & Striano, T. (2008). Neural Processing of Eye Gaze and Threat? Related Emotional Facial Expressions in Infancy. *Child Development*, 79(6), 1752-1760.
- Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 57(3), 567-582.
- Hornik, R., & Gunnar, M. R. (1988). A descriptive analysis of infant social referencing. *Child Development*, 59(3), 626-634.
- Hughes, C., & Dunn, J. (1998). Understanding mind and emotion: longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, 34(5), 1026-1037.
- Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts.
- Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115(2), 288-299.
- James, W. (1890). *The principles of psychology*. New York: Dover.
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8), 589-604.

- Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best preschoolers use past reliability over age when learning new words. *Psychological Science*, 17(9), 757-758.
- Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, 21(10), 1541-1547.
- Jones, D. C., Abbey, B. B., & Cumberland, A. (1998). The development of display rule knowledge: Linkages with family expressiveness and social competence. *Child Development*, 69(4), 1209-1222.
- Josephs, I. E. (1994). Display rule behavior and understanding in preschool children. *Journal of Nonverbal Behavior*, 18(4), 301-326.
- 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* (4 ed., pp. 467-482). New York, NY: Guilford.
- Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in Neuroscience*, 15(4), 241-252.
- King, S. (2010). *Danse Macabre*. Simon and Schuster.
- Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123-1128).
- Koenig, M. A., Clement, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, 15(10), 694-698.

- 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.
- Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244-247.
- Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4), 917-940.
- LaBounty, J., Wellman, H. M., Olson, S., Lagattuta, K., & Liu, D. (2008). Mothers' and fathers' use of internal state talk with their young children. *Social Development*, 17(4), 757-775.
- 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. (1966). *Psychological stress and the coping process*. New York: McGraw-Hill.
- Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist*, 46(8), 819-834.

- 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* (4 ed., pp. 495-509). New York, NY: Guilford.
- Leventhal, H. (1980). Toward a comprehensive theory of emotion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 13, pp. 139-197). New York: Academic Press.
- Leventhal, H. (1984). A perceptual-motor theory of emotion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 17, pp. 117-182). New York: Academic Press.
- Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (2014). The child as econometrician: A rational model of preference understanding in children. *Plos One*, 9, e92160.
- Lutz, C. (1982). The domain of emotion words on Ifaluk. *American Ethnologist*, 9(1), 113-128.
- Lyons, W. (1980). *Emotion*. Cambridge, MA: Cambridge University Press.
- MacLaren, R., & Olson, D. (1993). Trick or treat: Children's understanding of surprise. *Cognitive development*, 8(1), 27-46.
- Marinier III, R. P., Laird, J. E., & Lewis, R. L. (2009). A computational unification of cognitive behavior and emotion. *Cognitive Systems Research*, 10(1), 48-69.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.
- Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K.R. Scherer, T. Bänziger, & E. Roesch (Eds.), *A Blueprint for an Affectively Competent Agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford: Oxford University Press, in press

- Martinez, L., Falvello, V. B., Aviezer, H., & Todorov, A. (2016). Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cognition and Emotion*, 30(5), 939-952.
- 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.
- McShane, B. B. & Bockenholt, U. (2017). Single-Paper Meta-Analysis: Benefits for Study Summary, Theory Testing, and Replicability. *Journal of Consumer Research*, 43(6), 1048-1063.
- 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.
- Minsky, M. (2007). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. New York: Simon and Schuster.
- Mintz, T. H. & Gleitman, L. R. (2002). Adjectives really do modify nouns: The incremental and restricted nature of early adjective acquisition. *Cognition*, 84(3), 267-293.
- Misailidi, P. (2006). Young children's display rule knowledge: Understanding the distinction between apparent and real emotions and the motives underlying the use of display rules. *Social Behavior and Personality: an international journal*, 34(10), 1285-1296.
- Moll, H., Kane, S., & McGowan, L. (2016). Three-year-olds express suspense when an agent approaches a scene with a false belief. *Developmental Science*, 19(2), 208-220.
- Montague, D. P., & Walker-Andrews, A. S. (2001). Peekaboo: a new look at infants' perception of emotion expressions. *Developmental Psychology*, 37(6), 826-838.

- 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.
- Moses, L. J., Baldwin, D. A., Rosicky, J. G., & Tidball, G. (2001). Evidence for referential understanding in the emotions domain at twelve and eighteen months. *Child Development*, 72(3), 718-735.
- Mumme, D. L., & Fernald, A. (2003). The infant as onlooker: Learning from emotional reactions observed in a television scenario. *Child Development*, 74(1), 221-237.
- Naito, M., & Seki, Y. (2009). The relationship between second-order false belief and display rules reasoning: the integration of cognitive and affective social understanding. *Developmental Science*, 12(1), 150-164.
- Nunner-Winkler, G., & Sodian, B. (1988). Children's understanding of moral emotions. *Child Development*, 59(5), 1323-1338.
- Nussbaum, M. (1990). *Love's knowledge*. Oxford, UK: Oxford University Press.
- Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion*, 1, 29-50.
- Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, 143, 141-162.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge, UK: Cambridge University Press.
- Panksepp, J. (1992). A critical role for “affective neuroscience” in resolving what is basic about basic emotions. *Psychological Review*, 99(3), 554-560.

- Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3(9), 337-344.
- Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3), 437-471.
- Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2), 125-137.
- 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.
- Posamentier, M. T., & Abdi, H. (2003). Processing faces and facial expressions. *Neuropsychology Review*, 13(3), 113-143.
- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 715-734.
- Powell, L. J. & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, 110(41), E3965-E3972.
- 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.

- Roseman, I. J., Antoniou, A. A., & Jose, P. E. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3), 241-277.
- Ruffman, T., & Keenan, T. R. (1996). The belief-based emotion of surprise: The case for a lag in understanding relative to false belief. *Developmental Psychology*, 32(1), 40-49.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.
- Russell, J. A. (1990). The Preschooler's Understanding of the Causes and Consequences of Emotion. *Child Development*, 61(6), 1872-1881.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.
- Russell, J. A. & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805-819.
- Russell, J. A., & Bullock, M. (1986a). Fuzzy Concepts and the Perception of Emotion in Facial Expressions. *Social Cognition*, 4(3), 309-341.
- Russell, J. A., & Bullock, M. (1986b). On the dimensions preschoolers use to interpret facial expressions of emotion. *Developmental Psychology*, 22(1), 97-102.
- Saarni, C. (1984). An observational study of children's attempts to monitor their expressive behavior. *Child Development*, 55(4), 1504-1513.
- Saarni, C., & Harris, P. L. (1991). *Children's Understanding of Emotion*. CUP Archive.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1), 27-52.

- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6), 2408-2412.
- Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87-124.
- Schachter, S. (1964). The interaction of cognitive and physiological determinants of emotional state. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 49-80). New York: Academic Press.
- Scherer, K. R. (1984). On the nature and function of emotions: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293-317). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Scherer, K. R. (1994). Affect bursts. In S. H. M. Van Goodzen, N. E. Van de Poll, & J. A. Sergeant (eds) *Emotions: Essays on Emotion Theory* (pp. 161-193). Lawrence Erlbaum Associates.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227-256.
- 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., 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., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press.

- Scheutz, M. (2004). Useful roles of emotions in artificial agents: A case study from artificial life. In D. McGuinness & G. Ferguson (eds.), *Proceedings of the National Conference on Artificial Intelligence* (pp. 42-47). San Jose, CA: MIT press.
- Schmidt, M. F. & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PLoS One*, 6(10), e23223.
- Scott, R. M. (2017). Surprise! 20-month-old infants understand the emotional consequences of false beliefs. *Cognition*, 159, 33-47.
- 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.
- Shariff, A. F., & Tracy, J. L. (2011). What are emotion expressions for?. *Current Directions in Psychological Science*, 20(6), 395-399.
- 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. (2015). *Abstract representations of attributed emotion: evidence from neuroscience and development*. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.
- 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., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48, 813-838.
- Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and the emotions. *Cognition & Emotion*, 7(3-4), 233-269.
- Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558-1568.
- Soderstrom, M., Reimchen, M., Sauter, D., & Morgan, J. L. (2017). Do infants discriminate non-linguistic vocal expressions of positive emotions?. *Cognition and Emotion*, 31(2), 298-311.
- Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, 5, 39-43.
- Soken, N. H. & Pick, A. D. (1999). Infants' perception of dynamic affective expressions: do infants distinguish specific expressions?. *Child Development*, 70(6), 1275-1282.
- Solomon, R. C. (1976). *The passions: Emotions and the meaning of life*. New York: Doubleday.
- Sorce, J. F., Emde, R. N., Campos, J. J., & Klinnert, M. D. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology*, 21(1), 195-200.
- Spelke, E. S. (2017). Core Knowledge, Language, and Number. *Language Learning and Development*, 13(2), 147-170.
- Stein, N. L. & Levine, L. J. (1989). The causal organisation of emotional knowledge: A developmental study. *Cognition and Emotion*, 3(4), 343-378.

- Stein, N. L. & Trabasso, T. (1992). The organisation of emotional experience: Creating links among emotion, thinking, language, and intentional action. *Cognition and Emotion*, 6(3-4), 225-244.
- Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2012). A formal model of emotion triggers: an approach for BDI agents. *Synthese*, 185(1), 83-129.
- Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30(3), 395.
- Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development*, 26(5), 436-444.
- Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, 43(3), 804.
- Talwar, V., Murphy, S. M., & Lee, K. (2007). White lie-telling in children for politeness purposes. *International Journal of Behavioral Development*, 31(1), 1-11.
- 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.
- Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332(6033), 1054-1059.
- 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.
- Thompson, R. A., & Lagattuta, K. H. (2006). Feeling and understanding: Early emotional development. In K. McCartney, & D. Phillips (Ed.), *The Blackwell Handbook of Early Childhood Development* (pp. 317-337). Oxford, England: Blackwell.
- Tomkins, S. (1962). *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer publishing company.
- 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.
- Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental Psychology*, 45(2), 534-543.
- Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychological Bulletin*, 121(3), 437-456.
- Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions: Contributions of auditory and visual information. *Infant Behavior and Development*, 14(2), 131-142.
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, 76(5), 820-838.

- 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*. Oxford University Press.
- Wellman, H. M. & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3), 245-275.
- 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., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75(2), 523-541.
- 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.
- Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emotion Review*, 5(1), 72-77.
- 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* (4 ed., pp. 307-318). New York, NY: Guilford Press.

- Widen, S. C., Pochedly, J. T., & Russell, J. A. (2015). The development of emotion concepts: A story superiority effect in older children and adolescents. *Journal of Experimental Child Psychology, 131*, 186-192.
- Widen, S. C. & Russell, J. A. (2004). The relative power of an emotion's facial expression, label, and behavioral consequence to evoke preschoolers' knowledge of its cause. *Cognitive Development, 19*(1), 111-125.
- Widen, S. C. & Russell, J. A. (2010). Children's scripts for social emotions: Causes and consequences are more central than are facial expressions. *British Journal of Developmental Psychology, 28*(3), 565-581.
- Widen, S. C., & Russell, J. A. (2008a). Children acquire emotion categories gradually. *Cognitive Development, 23*(2), 291-312.
- Widen, S. C., & Russell, J. A. (2008b). Young children's understanding of other's emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (3 ed., pp. 348-363). New York, NY: Guilford.
- 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., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2014). Joint inferences of belief and desire from facial expressions. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1796-1801).

- Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3), 223-250.
- Xu, F., Bao, X., Fu, G., Talwar, V., & Lee, K. (2010). Lying and truth-telling in children: From concept to action. *Child Development*, 81(2), 581-596.
- Xu, F., Cote, M., & Baker, A. (2005). Labeling guides object individuation in 12-month-old infants. *Psychological Science*, 16(5), 372-377.
- 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.